Application of High Resolution Melting analysis for haplotype detection in phylogeographic research and case studies of *Arenaria ciliata*, *A. norvegica* and *Minuartia recurva* (Caryophyllaceae)

Xiaodong Dang

In fulfilment of the requirement for the degree of

Doctor of Philosophy

National University of Ireland, Maynooth

Department of Biology

September 2012

Head of department: Prof. Paul Moynagh

Supervisor: Dr. Conor Meade

Table of contents

Declaration

This thesis has not been submitted in whole, or in part, to this, or any other University for any other degree and is, except where otherwise stated, the original work of the author.

Signed:_____

Xiaodong Dang

Acknowledgements

First of all I would like to thank my supervisor, Dr. Conor Meade for giving me the opportunity to pursue my PhD study at NUI Maynooth, and for the advice he provided throughout my works, both in bench works and in methodology.

During my study, my co-supervisor, Dr. Colin Kelleher and other researchers from the National Botanic Gardens of Ireland (Glasnevin) also provided advice and support for both laboratory and field works. Emma Howard-Williams has helped me a lot, especially in carrying me in the car looking for the plants everywhere in Ireland and Scotland. Undergraduate students Brian Collopy, Dale Oflaherty, Kevin Logan, Matthew Brocklebank, Ronan Morris and Ross Campbell, who each worked in the lab for months, also have contributed part of the DNA data. I owe my big thanks to them for their support.

Here I also thank other collaborators including Adreas Tribsch (Salzburg), Kevin Walker (BSBI), Bogdan Jovanovic (Heidelberg), Pablo Vargas (Madrid), Patrick Kuss (Bern) and Zachary Dwight (Salt Lake City), who helped in software use, locating, collecting and identifying samples, and

giving advice. Their helps are essential for the works in the present study.

I owe my thanks to department and college staffs, as well as researchers from other labs in the department of biology, who helped me in laboratory teaching, using and maintaining the instruments, delivering reagents and devices, and in those paper works for administrative use. Their helps have made my life easier in Ireland.

I am also grateful to the university who granted the John & Pat Hume Scholarship to support my study for the past four years. As part of the bigger project of a biogeographic research on Irish arctic/alpine plants, the study was funded by Science Foundation Ireland (Grant No.: SFI/08/RFP/EOB1545). The Roche LightCycler® 480 System real-time PCR instrument used in the study was also funded by Science Foundation Ireland (Grant No.: SFI/07/RFP/GEN/F571/ECO7).

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Tables

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Abstract

The present study includes two aspects in the field of phylogeography. First, the technology of High Resolution Melting (HRM) analysis based on real-time PCR is introduced as a tool for haplotype detection in phylogeographic research. Second, phylogeographic study of three arctic-alpine or alpine plant species in the family Caryophyllaceae is carried out based on the haplotype data obtained through established protocol of HRM analysis.

In Chapter 2, experimental protocols of HRM analysis combined with posterior DNA sequencing as a complementary and confirmatory method are established for haplotype detection in the case study of three species, *Arenaria ciliata*, *A. norvegica* and *Minuartia recurva*. Non-coding chloroplast DNA loci, rps16 intron (c.750bp for *Arenaria* and c.690bp for *Minuartia*) and trnT-trnL (c.640bp for *Arenaria* and c.540bp for *Minuartia*) are used in HRM analysis, where they are split into smaller (<400bp) amplicons for each real-time PCR reaction. The protocol is able to reveal 19 out of 20 haplotypes of rps16 and all of the 24 haplotypes of trnT-trnL in the case of *Arenaria* species, and to reveal all of the four haplotypes of rps16 and three out of eight (or five if variation in SSRs is

not considered) haplotypes of trnT-trnL in the case of *M. recurva*. Posterior DNA sequencing reveals only one more haplotype with rps16 in the case of *Arenaria* species, which indicates a high sensitivity of HRM analysis in both cases.

In Chapter 3 and 4, phylogeographic studies are carried out for the *Arenaria* species and *M. recurva*. Based on the haplotype identities and their distribution within and among the sampled populations, a complete phylogeographic study becomes possible. Maximum-likelihood phylogenetic trees and statistical parsimony networks are constructed among the haplotypes, and the genealogical relationship among the haplotypes is combined with their geographic distribution to understand the migratory history of the populations, population genetic analysis is made to understand the genetic diversity and structure of the populations and mismatch analysis is performed to understand demographic history of the populations. In Chapter 3, deeply diverged clades are revealed which are not in accordance of either subspecies or geographic localities. The data indicate a much older establishment of the populations of *A. ciliata* on Ben Bulben in northwest Ireland than thought before, possibly as early as 150-250 thousand years ago. Thus these Irish populations may have survived the last ice age *in situ*, rather than having immigrated after the

end of Pleistocene (c. 12, 000 years ago). However, the result shows that the Irish populations are more closely related to the Iberian populations than they are related to the Alps populations. In Chapter 4, much lower level of genetic polymorphisms is revealed in the *M. recurva* populations, although they cover a European range comparable to that in the *Arenaria* case. The Balkans region is suggested as the refugium for the species, while little variation is found across the species distribution from the Alps, the Pyrenees, the north side of Spain and Ireland, which indicates a recent dispersal of the species in west Europe. Also a close relationship is suggested between the Irish populations and the Iberian populations.

In Chapter 5, a further theoretical assessment of possibility of missed detection in HRM analysis is carried out via *in silico* HRM simulation. Based on an amplicon of rps16I in *A. ciliata*, random mutations are made and HRM sensitivity is evaluated to different classes of single nucleotide substitutions. Class I and II class I (A/G or C/T) and II (A/C or G/T) substitutions are demonstrated easier to be detected than class III (C/G) and IV (A/T) substitutions. Further analyses suggest that between 50 and 650bp, amplicons of greater sizes are more likely to yield multiple melting peaks, which is favourable for higher sensitivity in HRM analysis. Between 100 and 550bp when all the amplicons render double melting

peaks, amplicons of greater sizes tend to provide lower sensitivity in HRM analysis. Amplicons smaller than 350bp with double melting peaks are considered to generate an acceptable rate of missed detection (<10-20%). In addition, *in silico* HRM analysis is tested with available DNA sequences of mitochondrial and chloroplast loci from published phylogeographic studies, and is demonstrated to help distinguish most of the extant haplotypes, although with a proportion of haplotypes missed (typically 10-20%). The results provide information for possible improvements of HRM analysis to be widely applied for haplotype detection in phylogeographic research.

Chapter 1

General introduction

1.1 Biogeographic questions on arctic/alpine carnation species in Ireland

1.1.1 Arctic/alpine carnation plants in Ireland

The term arctic plants refers to plants found in arctic and/or subarctic areas but are scarce or absent in temperate areas, and the term alpine plants refers to those occurring at high altitudes above the tree line but rarely found in lowland areas (Webb 1983). Arctic-alpine plants can be defined as those that have both the above distributions, and are considered to be highly vulnerable to climate change because they are limited to habitats with low temperatures and susceptible to invasive species once it becomes warmer in the local area (Pauchard *et al.* 2009). Alpine plants are usually found at or above c.2,000 meters in mountainous areas in the temperate region, e.g. on the Alps and the Pyrenees in south Europe, while in northern areas such as in Ireland they are found on mountaintops above c.500 meters. There are some native arctic-alpine plants in Ireland, including *Silene acaulis*, and *Arenaria ciliata*, and alpine plants including *Minuartia recurva* in the family Caryophyllaceae (Angiosperm Phylogeny Group III 2009), which are listed in *Flora (Protection) Order, 1999*, as species to be protected in

Ireland. As their distributions are limited to the mountaintops at only a few sites in Ireland, they are regarded as susceptible to the adverse effects of global warming (Webb 1983).

In the entire region of Europe, there has been much interest in how the plants and animals survived the climatic oscillations that occurred throughout the Quaternary when arctic ice sheets expanded southward and receded northward in cycles of approximately 100,000 years. This interest is especially strong regarding the latest Pleistocene glaciation from around 130,000 years ago to 12,000 years ago, and in particular how organisms migrated during the postglacial period when the ice sheets were retreating northward. It is believed that many organisms had survived in three major potential refugia in southern Europe during the last glacial maximum (c. 25,000 to 18,000 years ago), which include the Iberian Peninsula, Italy, and the Balkans, where mountain ridges of the Pyrenees and the Alps served as barriers between temperate habitats in the south and extensive ice sheets and tundra in the north (Taberlet *et al.* 1998; Hewitt 2000). The limited and fragmented distribution of biota in these areas could lead to loss of diversity in separated populations of the involved species, allowing intraspecific divergence to arise as new mutations accumulated within each population while inter-population

gene flow was limited by spatial isolation. Later during postglacial migration, when different lineages met in central Europe and formed admixed populations, a mosaic pattern of genetic identities formed, which may possibly show higher genetic diversity than any of the individual refuge populations (Petit *et al.* 2003).

Three major patterns of the migration routes from the refugia to the current distributions of the plants and animals were proposed by Hewitt (2000) considering the effects of common geological and geomorphic factors on different taxa in Europe, which were represented by three species, ‘grasshopper’, ‘hedgehog’ and ‘bear’.

In the first patter as shown by common meadow grasshoppers (*Chorthippus parallelus*), populations across northern Europe show close genetic relationship to the Balkan populations with little variation among themselves, while isolated populations in southern Europe including Italy and Spain carry local haplotypes that are more deeply divergent from one another. This pattern represents recent postglacial migration from the Balkans north-westward to central Europe.

In the second patter shown by two sister species of hedgehog, *Erinaceus*

europeus and *E. concolor*, deeply divergent clades were revealed to exist within each species. The clades are estimated to have diverged several million years ago and been isolated to different south/north ‘strips’ from western Europe to the east areas including Turkey and Israel. The northern populations have been established by northward colonization from different refugia separately.

In the third pattern shown by brown bears (*Ursus arctos*), the populations in central Europe are revealed to have migrated both from the Iberia in the west and from Caucasian/Carpathian refugia in the east. The two different lineages may have met ~9 thousand years ago in central Sweden. Compared to these, the lineages in Italy and the Balkans are restricted to their current locations and are not found to disperse to northern areas.

A number of species, including both animals and plants, have been reported to fall into each of the above three patterns in Europe, as reviewed by Hewitt (2000). However, it is unknown if other patterns also exist and extensive investigations are still needed to understand the biogeographic history of more species of interest.

Within the European context the Irish flora was traditionally considered

to be a sampling of north-west European flora, although there are some subspecies endemic to Ireland (Webb 1983). The smaller size of the Irish flora compared to the British flora used to be attributed to the late-glacial and/or post-glacial migration (around 12,000 years ago) of plants from Europe to Ireland via Britain as the transit station (Webb 1983; Wingfield 1995), which implies that the Irish flora derived from the British flora and thus constitutes a subset of the latter.

However, the question arose regarding 18 plant species found only in Ireland but not in Britain, including *A. ciliata* and *M. recurva*. One possibility is that their British populations have become extinct after the species migrated to Ireland, although this is unlikely considering the abundance of suitable habitats for them in Britain (Webb 1983). Based on studies in biogeography and palaeogeography, two alternative hypotheses have been proposed to answer the question. One hypothesis is that some Irish species may have come via a post-glacial land bridge along the Coast of Celtic Sea from the Iberian Peninsula (Wingfield 1995). Studies of Irish trees including oaks have shown support for this hypothesis (Mitchell & Ryan 1997; Kelleher *et al.* 2004; Mitchell 2006). Another hypothesis is that some Irish species may have survived throughout the last ice age on the ice-free surfaces of some inland mountains. It is

supported by the discovery of *A. ciliata* and *S. acaulis* specimens dated to about 30,000 years ago at Derryvree, Co. Fermanagh (Colhoun *et al.* 1972). Synge and Wright (1969) also found that the western surface of Ben Bulben may have been free from glaciation throughout the Pleistocene, constituting a refuge for cold-resistant plants.

These three hypotheses need to be tested to give an assessment of the uniqueness of the Irish populations of arctic and alpine plants from a biogeographic perspective. It will be helpful to investigate when and how they came into Ireland and how they have responded to the geographic and climate changes in the past, which will provide some advice on how to protect the Irish plant diversity in future.

The present study is focused on three arctic-alpine or alpine species occurring in Ireland and a range of places across Europe. The three plant species are in the family Caryophyllaceae, including *Arenaria ciliata* as an arctic-alpine species, its arctic-distributed sister species *A. norvegica* and *Minuartia recurva*, an alpine species. As described above, *A. ciliata* and *M. recurva* are not found to occur in Britain, which makes them interesting in terms of biogeography. The taxonomic and biogeographic details of the two *Arenaria* species and *M. recurva* are provided in

Chapter 3 and 4 respectively. Photographs of the three species are provided in Figure 1.1 to show their morphology, and the locations of the sampled populations are shown on the map in Figure 1.2.

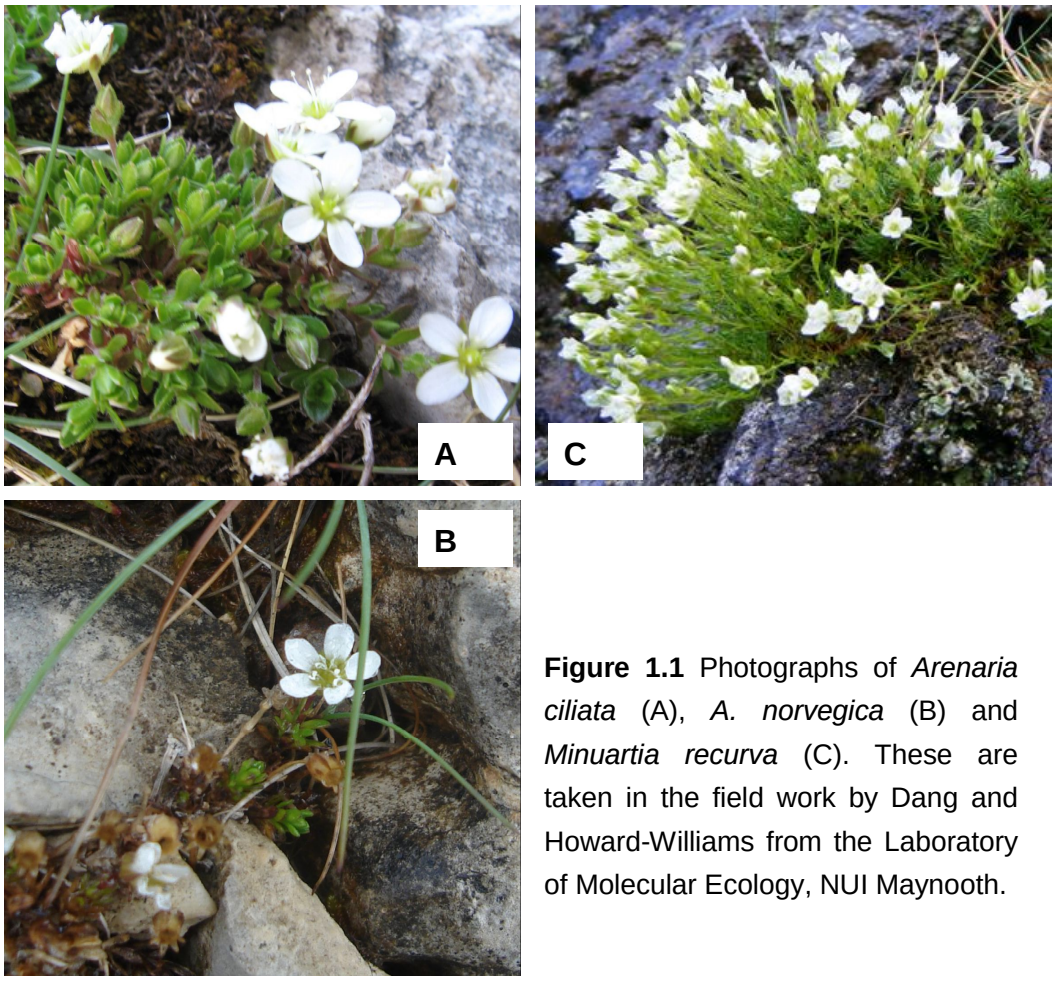


Figure 1.1 Photographs of *Arenaria ciliata* (A), *A. norvegica* (B) and *Minuartia recurva* (C). These are taken in the field work by Dang and Howard-Williams from the Laboratory of Molecular Ecology, NUI Maynooth.

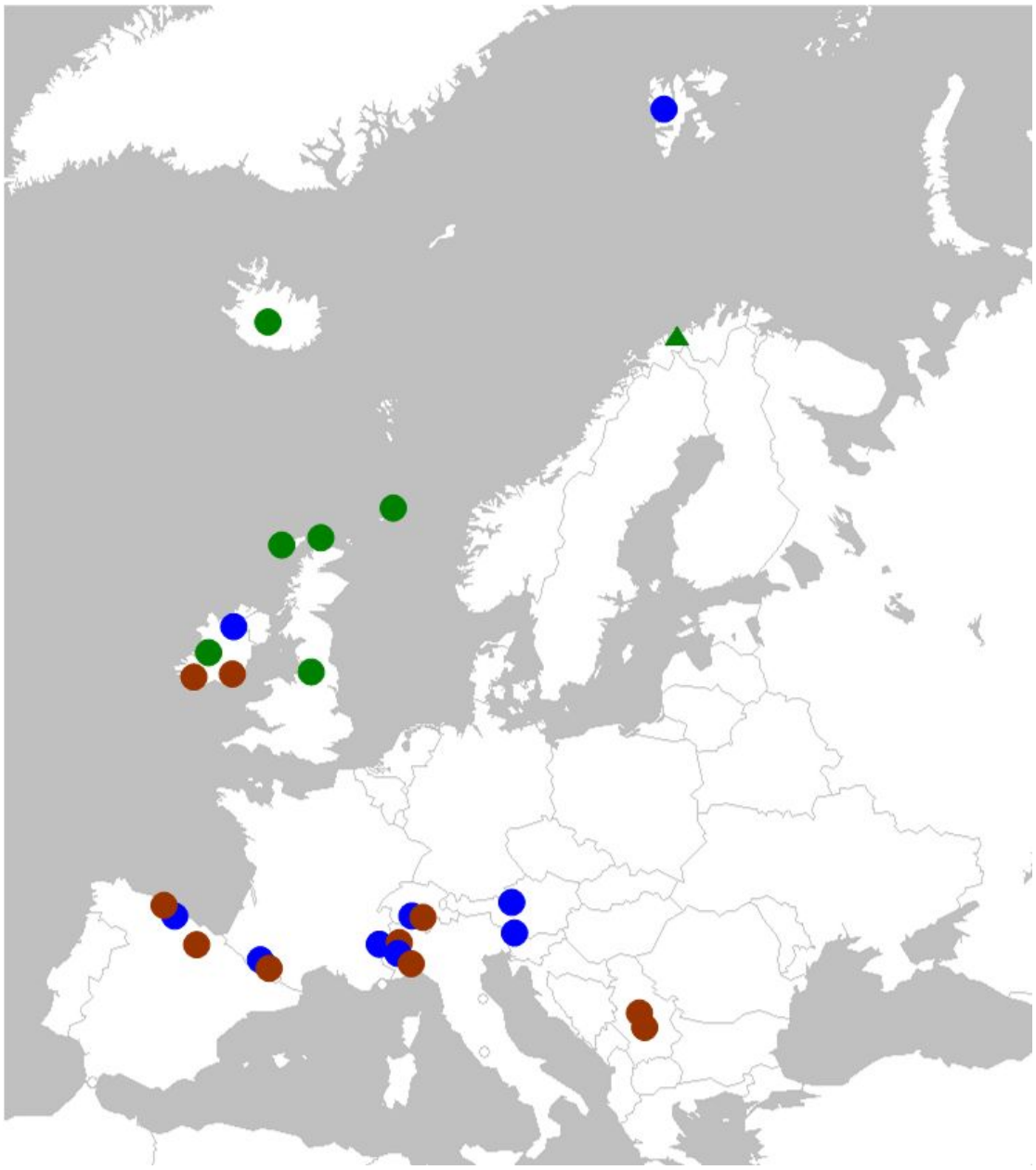


Figure 1.2 Distribution of the sampled population of *Arenaria ciliata* (blue), *A. norvegica* (green) and *Minuartia recurva* (brown) in Europe. The green triangle indicates the location of an accession of *A. norvegica* from Genbank (Accession No. HM772117; Westergaard *et al.* 2011).

1.1.2 Principles and methods of phylogeographic research

As a research field emerged since 1980s, phylogeography generally

investigates the evolutionary history of contemporary biodiversity by focusing on the genealogical relationship among individuals of one or more closely related species across populations from different geographic localities (Avise *et al.* 1987). It generally attempts to answer where the current populations of particular taxa came from, when the migration occurred and how the patterns of genetic diversity have been shaped by historical and geological events (Reviewed by Avise 2000). Based on the coalescent theory depending on haploid DNA data (Kingman 1982; Hudson 1983; Tajima 1983), the evolutionary relationship between individuals can be inferred from their DNA sequences of selected loci, making it possible to trace the common history of existing conspecific organisms.

On some occasions, different species may have been affected by the same geographic and historical factors, sharing the same migration routes and thus showing the same contemporary patterns of population genetic distribution. Comparative phylogeography has been proposed to test whether such patterns exist among a group of species co-occurring within a community or in the same region (Taberlet *et al.* 1998; Bermingham & Moritz 1998; Vamosi *et al.* 2009).

Due to both theoretical and technical limitations, haploid DNA data obtained from mitochondrial or chloroplast genomes rather than diploid DNA data are normally used for phylogeographic studies of animals and plants respectively, while diploid data from nuclear genome, mostly acquired through coarse-grained techniques, e.g. SSRs (simple sequence repeats, or microsatellites), RFLPs and AFLPs and occasionally through cloning-sequencing of ITS loci and other gene regions, are usually used as a plus to give additional support of both maternal and paternal genetic information to the results (e.g. Valcárcel *et al.* 2006; Bettin *et al.* 2007; Flanders *et al.* 2009; Teacher *et al.* 2009; Wei *et al.* 2010). One advantage of haploid data is that mitochondrial and chloroplast genomes are uniparentally (usually maternally) inherited in most organisms. In phylogeographic studies of angiosperm plants, DNA sequence data of chloroplast genomes are predominantly used to track the maternal line of the populations, which exactly represents the migration events via seed dispersal. Non-coding DNA loci are usually chosen for analysis because they are expected to show higher levels of variation than coding regions, and to be less impacted by selection pressure (Shaw *et al.* 2005, 2007).

In phylogeographic studies, usually several to tens of wild populations are sampled from distantly located places to cover the focused

distribution of the studied taxa while less than ten individuals on average from each population are subject to DNA sequencing analysis (e.g. Pauls *et al.* 2006; Bettin *et al.* 2007; Flanders *et al.* 2009; Teacher *et al.* 2009). One issue of concern is that the polymorphism level of each population may be under-estimated by insufficient sampling, as potentially there could be rare haplotypes unrepresented in the sampled individual. Biased or even incorrect conclusions could be made due to such insufficiency of data. Although in plant studies as many as thirty individuals can be collected from each population, the high cost of DNA sequencing poses a limitation that prevents researchers from exhausting all the existent alleles of a selected DNA locus when there are hundreds or even thousands of individual plants to be assayed. A common technique currently used in many phylogeographic studies is that of PCR-RFLP. This technique has been successfully used across a range of taxa (Taberlet *et al.* 1998), but significant workload of initial screening is needed to obtain suitably informative markers and it is not possible to detect point mutations that are not covered by restriction enzyme cut sites. As a result a lower-cost and more efficient method is needed to reveal the variation among as many haplotypes as possible from the field.

1.2 High Resolution Melting analysis as a potential method of haplotype detection in phylogeographic research

DNA melting analysis is based on real-time PCR (polymerase chain reaction) techniques, which incorporate duplex DNA-binding fluorescent dyes, e.g. LC Green and SYBR Green I, into traditional PCR reactions to monitor the progress of DNA amplification (Wittwer *et al.* 1997). The melting process is initiated after the completion of amplification, making the programmed increase of temperature to dissociate the amplified double-strand DNA fragment, leading to a decrease in the strength of detected fluorescent signals. A melting curve is thus obtained by plotting the fluorescence strength against the temperature increase. As the shape of the melting curve and the precise temperature of maximum dissociation (the melting peak, Tm) are determined by the size and the sequence of the involved DNA amplicon, melting curve analysis has become a valuable tool for genotype (or haplotype for haploid genomes) identification and detection (Ririe *et al.* 1997).

With recent advances in real-time PCR technologies and new-generation saturating dyes, such as the ResoLight Dye patented by the Roche company, introduced into DNA melting analysis, High Resolution

Melting (HRM) analysis has thus been developed and widely used for genotyping and mutation scanning, principally in the realm of human genetic and medical research (Reviewed by Wittwer 2009). Multiple combinations of different dyes and reaction buffer systems have been tested on various platforms (Monis *et al.* 2005; Herrmann *et al.* 2006), for example the LightCycler systems provided by Roche have been validated as an efficient tool of genotyping and mutation scanning (Lyon & Wittwer 2009; Tindall *et al.* 2009).

In the past few years, HRM analysis has also been introduced into areas of agricultural research (e.g. Dong *et al.* 2009; Wu *et al.* 2009) and population biology (e.g. Smith *et al.* 2010; Mader *et al.* 2010). In these studies, HRM was used as a genotype validation method, i.e. the expected identities and exact sequences of the studied alleles had been predefined before HRM analyses were employed.

On the other hand, the mutation scanning function of HRM analysis has also been utilized in population studies, e.g. lineage screening in a perennial ryegrass breeding scenario (Studer *et al.* 2009), where HRM anslysis revealed unknown mutant alleles of the selected loci without the need for *a-priori* sequencing. The success of this sort of studies indicates

that the mutation scanning utility of HRM analysis can also be applied to phylogeographic research, where potentially a plural of unknown alleles of a particular DNA locus are supposed to exist in the wild populations, neither the number nor the sequence identities of which are known before the populations are put into high-throughput assays of allele detection and counting.

One convenience for HRM analysis to be used in phylogeographic research is that heterozygous genotypes are not usually seen from the majority data sources from haploid chloroplast and mitochondrial genomes, as the scanning of heterozygous genotypes by HRM analysis is still being developed and is not sufficiently sensitive to be applied for blind genetic diversity screening at present. For detection of haploid genotypes (haplotypes), in theory HRM analysis can show which individual samples share the same haplotype and which samples differ in haplotype, allowing a relatively small number of representive samples to be then sequenced to get full information of haplotype identites and their distribution among the sampled populations.

1.3 Limitations of HRM analysis in haplotype detection and the scope for improvement

Despite the potential advantages of HRM analysis to be used for haplotype detection in phylogeographic studies, a key consideration however are the technical limits, as the size and specific nucleotide content of the amplicon influences the discriminating ability of HRM analysis, due to the complex physical-chemical mechanisms of DNA duplex denaturation (Ririe *et al.* 1997; Wittwer *et al.* 2003). In general, the sensitivity of HRM analysis is influenced by the amplicon size, the melting domain and the classes of nucleotide substitutions. Furthermore, HRM analysis in theory is a cursory assay with inherent lack of capability of exhausting all the possible mutant identities of an amplicon. The limitation of HRM analysis has to be quantitatively assessed with both realistic DNA loci, i.e. the chloroplast and mitochondrial loci widely used in phylogeographic research, and computer-generated random DNA sequences. The quantitative assessment is powered by *in silico* simulation of HRM analysis with the software uMelt^{SM} (Dwight *et al.* 2011), which helps generate possible melting curves of given DNA sequences and tell if two alleles can be distinguished by providing their Tm results. The possibility of missed detection is thus quantitatively evaluated and

potential solutions are provided to enhance the success rate.

1.3.1 Inherent limitation of HRM analysis

There is an inherent possibility of missed detection of variation between different haplotypes by HRM analysis when Tm values are used for comparison between templates.

For instance, a target DNA locus with the amplicon size of 200bp can have 4^200 (~2.6 E120) possible sequence identities when only nucleotide substitutions are considered. At the same time, considering the minimal threshold of Tm difference between different amplicons being 0.1°C within the normal melting range of temperature between 65°C and 95°C, there are at most 300 well distinguished Tm values with a single melting peak. Where two Tm peaks are evident for an amplicon this still only equates to 300*299 = 89,700 discrete Tm combinations, which is far from being capable to distinguish all the possible sequence compositions of the amplicon. This inherent insufficiency is also the reason why HRM analysis can never be used in place of sequencing for exact information of DNA identities.

However the haplotypes are in fact homologous DNA regions with high similarity in their nucleotide sequence. Thus the variation between two haplotypes of a 200bp target locus is much smaller than the variation between any two of the randomly generated 200bp amplicons. Usually in phylogeographic studies, less than 5-10%, typically 1-2% of the nucleotides within the amplified DNA fragment are polymorphic between haplotypes found from different populations (e.g. Flanders *et al.* 2009; Teacher *et al.* 2009). As a result, the likely number of observed haplotypes within a given 200bp DNA loci is much smaller than the figure described above. Nevertheless, the likelihood and sensitivity limits associated with patterns of missed detection need to be quantitatively examined.

1.3.2 HRM sensitivity and amplicon size

In general, for a single nucleotide polymorphism (SNP) between two haplotypes, the divergence in Tm values becomes greater when less flanking nucleotides are included in the targeted amplicon, thus small fragments are preferred for many HRM applications because genotype divergence is clearer (Liew *et al.* 2004). For precision diagnostic purposes, e.g. SNP mutations, amplicon size is typically c. 50-100bp,

although amplicons in the 150-250bp size range are less commonly used for amplicon variant identification (Vossen *et al.* 2009; Wittwer 2009). Reed & Wittwer (2004) reported in an experimantal evaluation of HRM that across a range of amplicon size classes, identification of one SNP between otherwise homologous amplicons was achievable in 100% of cases for product <300bp in length, but with lower success rate for products between 300 and 1000bp. Although new-generation saturating dyes are developed and recipes of PCR buffers, e.g. LightCycler 480 HRM Master Mix (Roche), are formulated to improve the sensitivity, the manufacturer's guidelines suggest it is still difficult to distinguish amplicons greater than 400bp with a SNP between them.

In phylogeographic analysis, target DNA loci between 200 and 1400bp in length (typically 700-1000bp) are amplified and sequenced to generate informative data (Shaw *et al.* 2005, 2007). Such large amplicon sizes are generally unsuitable for HRM analysis, which requires smaller sizes of the target DNA amplicon to achieve high discrimination sensitivity.

However, within larger amplicons, nucleotide substitutions and indels (insertions/deletions) between different haplotypes are more likely to occur at multiple sites, which may potentially contribute to greater

variation between their corresponding melting curves and thus may help offset the lower discriminating capability with the larger amplicons. This multiple nucleotide inter-haplotype variation is in fact one of the requirements to provide sufficient polymorphism information for phylogeographic analysis, and the principle reason why larger amplicons are preferred in phylogeographic studies.

Furthermore, larger amplicons have the potential to give rise to more than one melting domains, generating multiple Tm peaks in the result of HRM analysis which are informative by indicating variation at multiple nucleotide sites: in amplicons of 50 to 150bp in length only one Tm value is evident, however fragments longer than 200bp can have two or more melting peaks (Vossen *et al.* 2009; Wittwer 2009; and see Chapter 2 and 3). The potential occurrence of multiple melting domains provides an additional advantage of using larger amplicons with HRM analysis, although it is not guaranteed to occur in every case, depending on the exact nucleotide composition and sequence of the amplicon (Vossen *et al.* 2009; Wittwer 2009).

Considering these various factors the appropriate length for HRM amplicons at each target locus needs to be optimized, so as to cover as

much polymorphic nucleotide sites as possible while not losing discriminating sensitivity between haplotypes. One part of this PhD research focuses on this problem, and with the aim of elucidating how amplicon design effects sensitivity, a quantitative investigation is made into the effect of amplicon size and melting domains on HRM sensitivity. This will provide some needed guidance on how amplicon optimization can be achieved.

1.3.3 HRM sensitivity to SNPs: Discriminating between haplotypes with the smallest possible template variation

HRM analysis is known to have different sensitivity limits for the four classes of nucleotide substitutions, which are class I (A/G and C/T) , class II (A/C and G/T), class III (C/G) and class IV (A/T) substitutions. In general class IV substitutions are regarded as the most difficult to be identified by HRM analysis among the four classes, and class III substitutions can also be problematic, while class I and II mutations are usually well distinguished with least difficulty (von Ahsen *et al.* 2001; Liew *et al.* 2004; Reed & Wittwer 2004). However, the possibility of missed detection for the four classes need to be quantitatively evaluated.

1.3.4 Evaluating HRM analysis with *in silico* simulation

With accumulated theoretical studies on the melting dynamics of duplex DNA molecules (Poland 1974; Yeramian *et al.* 1990; Blake *et al.* 1999; Tøstesen *et al.* 2003; Huguet *et al.* 2010), the profile of the DNA melting process can be simulated *in silico* using the web-based software, uMelt^{SM} developed by Dwight *et al.* (2011), where any possible DNA template of 50-3600bp in size can be filled into the online input box (http://www.dna.utah.edu/umelt/um.php) so that the profile of its melting curve is simulated and illustrated, while the position (in temperature) of each melting peak is recorded as a Tm value (see Figure 1.3). The theoretical variation between different mutants of the same DNA locus in their melting profiles (including the Tm differences) can thus be predicted on computer. With help from *in silico* HRM analysis, better decision of amplified locus and the exact priming position can be made via testing how different combinations of amplicon and primers will impact on the DNA melting profile and potentially on the sensitivity.

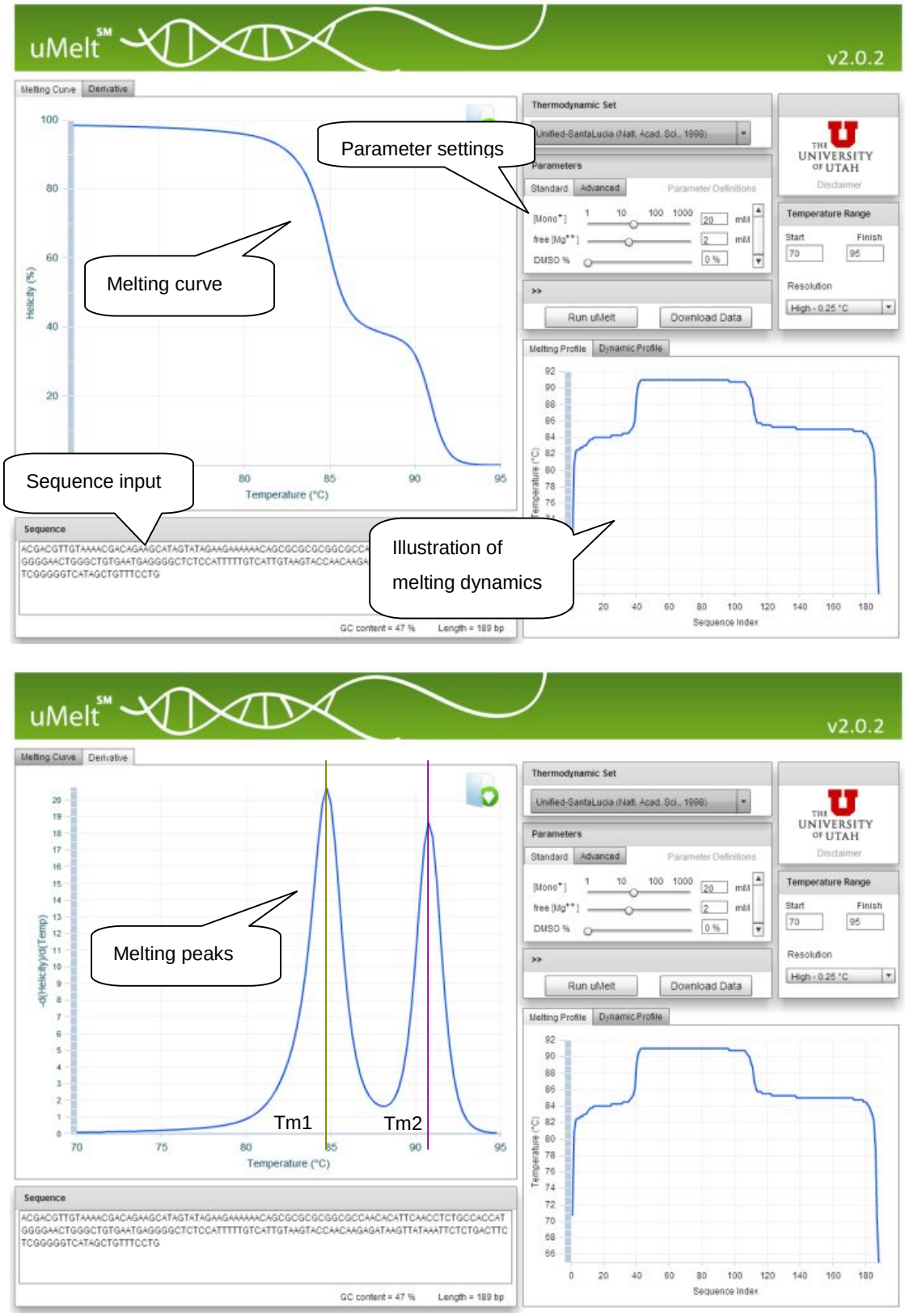


Figure 1.3 The interface of uMelt[SM] (retrieved from http://www.dna.utah.edu/umelt/um.php, developed by Dwight *et al.* 2011). The upper panel shows the melting curve and the lower panel shows the melting peaks as the derivative curve of the above.

As this is new software, it is an ideal time to evaluate the reliability of the *in silico* HRM estimates. At each locus, the precise correlation between actual *in-vitro* HRM data and modelled *in-silico* HRM can be evaluated. With known haplotypes revealed by *in vitro* HRM analysis and verified by sequencing, their sequences can be uploaded to the uMelt^{SM} software to obtain the simulated theoretical melting profile and peak Tm values. Then the *in silico* Tm variation between different haplotypes could be thus calculated, to be compared with the corresponding *in vitro* Tm variation. Knowledge is thus acquired of how consistent the *in silico* test is with the *in vitro* test based on known DNA sequences and how reliable and informative the *in silico* test is.

With *in silico* simulation, it is also possible to evaluate the rate of missed detection with a specific DNA locus by arbitrarily creating possible mutants with different types and numbers of mutations within the locus region and looking into how their *in silico* melting profiles vary between each other. In such cases the possibility that two alleles can be distinguished by HRM can be calculated.

The broad efficacy of HRM analysis in phylogeographic research can also be evaluated by using *in silico* investigation of haplotype sequence

DNA data from published work in the literature, by examining how many of the recorded haplotypes within specific studies could have been distinguished by HRM analysis (based on optimised amplicon design for each locus) had it been used in place of exhaustive sequencing.

1.4 Objectives

In general, the present study is aimed at demonstrating and assessing the efficacy of HRM analysis in phylogeographic research through two case studies of arctic/alpine carnation plants in Ireland and continental Europe, and also through theoretical studies with the aid of *in silico* simulation.

In Chapter 2, the method of HRM analysis is applied to haplotype detection from sample populations of *Arenaria ciliata*, *A. norvegica* and *Minuartia recurva*. Experimental protocols are established, the sensitivity and reliability of HRM analysis are assessed and the consistency between *in vitro* and *in silico* HRM analysis is examined. The haplotypes detected by HRM analysis in this way and confirmed by DNA sequencing are then collated for use in further analyses.

In Chapter 3, population genetic and phylogeographic analyses are

performed with haplotype data for the two *Arenaria* species, to understand how their current diversity and biogeographic distribution in Europe have been formed by their population history.

In Chapter 4, population genetic and phylogeographic analyses are performed with haplotype data for *Minuartia recurva*. The phylogeographic history of the species in Europe is examined and compared with that of the *Arenaria* species.

In Chapter 5, a theoretical evaluation of the efficacy of HRM analysis is performed with both computer-generated DNA sequences and available DNA data from published phylogeographic studies with the aid of *in silico* simulation. The probability of missed detection by HRM when used in phylogeographic studies is quantitatively evaluated regarding the considerations in section 1.4.

In Chapter 6, a further discussion is provided on two aspects. First, the prospect of HRM analysis is discussed as a method of genotype detection to be widely used in phylogeographic research. Secondly, the phylogeographic patterns found with the three species are discussed jointly regarding the biogeographic history of Ireland.

Chapter 2

High Resolution Melting analysis: case studies in carnation species

2.1 Introduction

In the previous chapter High Resolution Melting (HRM) analysis was introduced as a potential method for haplotype detection in phylogeographic research where the number and identities of all the haplotypes are unknown within a studied population. However, it is also evident that there is the possibility of missed detection in HRM analysis, arising from sensitivity limitations regarding amplicon size and the varying detectability of different mutation classes. Thus it is needed to evaluate the efficacy and limitation of HRM analysis in realistic studies, and to find possible ways of improving the performance of HRM analysis in haplotype detection.

This chapter deals with the application of HRM analysis in the study of three species in the family of Caryophyllaceae, *Arenaria ciliata*, *A. norvegica* and *Minuartia recurva*. Two non-coding regions of chloroplast genome, the rps16 intron and the trnT-trnL intergenic spacer, were analyzed with the three species. There were four principal objectives in this work:

(i) To establish a protocol of *in vitro* HRM analysis for detection of

chloroplast haplotypes within sampled populations of the three species;

(ii) To conduct and optimize a posterior amplicon sequencing strategy that validates the indentification of haplotypes via HRM analysis; and based on which,

iii) To assess the sensitivity of HRM analysis in the detection of mutant varieties at two non-coding chloroplast loci, trnT-L and rps16 based on differences in the characteristic melt curve profiles associated with each putative haplotype;

(iv) To evaluate the utility of *in silico* HRM simulation as a support for *in vitro* HRM analysis.

2.2 Materials

2.2.1 *Arenaria ciliata* and *A. norvegica*

As the major material studied in the present work, the species *Arenaria ciliata* L. is an arctic-alpine calcicole herb occurring in high-altitude mountainous areas in southern Europe and is also recorded from a few

sites in northern regions including Scandinavia and northwest Ireland. *Arenaria. norvegica* Gunnerus generally occurs in subarctic areas in Iceland, Scandinavia, Britain and Ireland (Jalas & Suominen 1983; Tutin *et al.* 1993; Walker *et al.* in press). The two plants are closely related sister species in taxonomy, falling within the *A. ciliata* L. Complex which also includes *A. moehringioides* Murr (= *A. multicaulis* L.) and *A. gothica* Fries (Wyse Jackson & Parnell 1987). The two sister species have similar habitats as both are most commonly found from shallow poorly-formed soils on exposed limestone, however they are not known to co-occur at any single location, leaving open the question if they are geographic subspecies of the same taxon.

Leaf tissues of 480 individual samples were collected from 17 populations of *A. ciliata* and 6 populations of *A. norvegica* across Europe, with a median population sample size of 20. The coding of populations, their locations and sampling sizes are listed in Table 2.1. This table is modified from supplementary Table 1 from Dang *et al.* (2012) with a few corrections.

Table 2.1 Location and size of sample populations of *Arenaria ciliata* (*Ac*) and *A. norvegica* (*An*) used in High Resolution Melt Analysis.

Population code	*Species and Location*	*Latitude/ Longitude*	*Sample size*
Ir1	Ac, King's Mountain, Co. Sligo, Ireland	N54^{o} 20.672' W08^{o} 27.373'	30
Ir2	Ac, Gowlaun Valley, Co. Sligo, Ireland	N54^{o} 21.353' W08^{o} 27.290'	30
Ir3	Ac, Glencarbury Mine, Co. Sligo, Ireland	N54^{o} 21.549' W08^{o} 24.044'	30
Ir4	Ac, Glendarragh Valley, Co. Sligo, Ireland	N54^{o} 21.879' W08^{o} 25.705'	30
Au1	Ac, Niedere Tauren, Steiermark, Austria	N47^{o} 16.270' E14^{o} 21.210'	27
Au2	Ac, Karawanken, Karten, Austria	N46^{o} 30.200' E14^{o} 29.120'	25
It1	Ac, Rifugio Mongioie, Piemonte Italy	N44^{o} 09.864' E07^{o} 47.201'	26
It2	Ac, Lago Visaisia, Piemonte, Italy	N44^{o} 27.227' E06^{o} 55.313'	1
It3	Ac, Colle dell'Agnello, Piemonte, Italy	N44^{o} 40.709' E06^{o} 59.484'	4
Fr1	Ac, Col D'Agnel, Provence- Alpes, France	N44^{o} 46.880' E06^{o} 40.638'	5
Pi1	Ac, Minas de Altaiz, Picos de Europa, Cantabria, Spain	N43^{o} 09.533' W04^{o} 49.302'	19
Pi2	Ac, Cabana Veronica, Picos de Europa, Cantabria, Spain	N43^{o} 10.644' W04^{o} 49.967'	15
Pi3	Ac, Corarrobres, Picos de Europa, Cantabria, Spain	N43^{o} 09.374' W04^{o} 48.213'	19
Py1	Ac, Valle de Benasque, Aragón, Spain	N42^{o} 40.957' E00^{o} 36.177'	20
Py2	Ac, Hospital de Benasque, Aragón, Spain	N42^{o} 40.957' E00^{o} 36.177'	16
Sw1	Ac, Site1, Gemmipass, Leukerbad, Switzerland	N46^{o} 23.883' E07^{o} 34.596'	8
Sw2	Ac, Site2, Gemmipass, Leukerbad, Switzerland	N46^{o} 25.165' E07^{o} 37.478'	29
Sv1	Ac subsp. pseudofrigida, Midtre Lovénbreen, Svalbard	N78^{o} 54.48' E12^{o} 04.70'	2
Sv2	Ac subsp. pseudofrigida, Bohemanflya, Svalbard	N78^{o} 23.42' E14^{o} 44.23'	5
NB	An, Black Head, Co. Clare, Ireland	N53^{o} 08.243' W09^{o} 16.048'	30
NE	An, Yorkshire, England	N54^{o} 17.000' E02^{o} 33.010'	19
NIc	An, Eldgja gorge, Herobreio, Iceland	N64^{o} 24.680' W18^{o} 42.252'	2
NIn	An, Inchnadamph, Highlands, Scotland	N58^{o} 07.493' W04^{o} 55.374'	30
NR	An, Rum, Western Isles, Scotland	N56^{o} 59.647' W06^{o} 18.863'	29
NS	An, Shetland Islands, Scotland	N60^{o} 30.835' W01^{o} 21.674'	29
Mean sample size			19
Median sample size			20

2.2.2 *Minuartia recurva*

The species *Minuartia recurva* Schinz & Thell. is a tufted perennial herb with woody basal stems growing in non-calcareous rocks, occurring in mountainous areas across continental Europe and at two separate sites in Ireland (Jalas & Suominen 1983; Tutin *et al.* 1993). Leaf tissues of 250 individual samples were collected from 13 different populations of *M. recurva* across Europe, from Kosovo to Ireland, with a median population sample size of 20. The coding of populations, their locations and sampling sizes are listed in Table 2.2 (on next page).

Table 2.2 Location and size of sample populations of *Minuartia recurva* used in High Resolution Melt Analysis. It is noted that MR8 and MR10 are collected from the same site but on different dates. One population recorded as MR9, collected from Cantabria, Spain (near the site of MR7), has been confirmed to be the other species, *M. verna*, and thus is not included here.

Population code	*Location*	*Latitude/ Longitude*	*Sample size*
MR1	Comeragh Mountains, Co. Waterford, Ireland	N52º 14.163' W07º 31.202'	20
MR2	West summit, Caha Mountains, Co. Kerry, Ireland	N51º 44.368' W09º 43.161'	30
MR3	East summit, Caha Mountains, Co. Kerry, Ireland	N51º 44.592' W09º 42.314'	30
MR4	Ligurische Alpen, Piemonte, Italy	N44º 09.734' E07º 47.082'	10
MR5	Cottische Alpen, Piemonte, Italy	N44º 33.086' E07º 07.135'	3
MR6	Mountain Kopaonik, Serbia	N43º 16.375' E20º 49.132'	26
MR7	Pico El Vigia, Cantabria, Spain	N43º 03.341' W04º 44.651'	35
MR8	Huesca, valle de Benasque, Spain	N42º 32.936'	24
MR10		E00.33.159'	
MR11	Laguna Negra, Soria, Spain	N41º 59.763' W02º 50.915'	13
MR12	Kosovo		5
MR13	Site 1, South Gspon, Staldenried, Switzerland	N46º 13.906' E07º 55.300'	27
MR14	North Gspon, Staldenried, Switzerland		8
MR15	Site 2, South Gspon, Staldenried, Switzerland	N46º 13.381' E07º 54.945'	19
Total			250
Mean sample size			19.2
Median sample size			20

2.3 Methods

2.3.1 DNA extraction and initial assays

Total genomic DNA was extracted from silica-dried leaf tissue using a modified CTAB protocol (Doyle & Doyle 1987) and dissolved in de-ionized H_2O. The target locus for haplotype identification was selected based on comparative sequence analysis of selected accessions from different populations of each species, i.e. one individual from each population, at 5 non-coding chloroplast loci prompted by Shaw et al. (2005, 2007), including the rpl16, rps16 introns, the intergenic trnS-trnG, trnT-trnL and rpl32-trnL spacers. Primer sequences for all the tested loci were as per Shaw et al. (2005; 2007) and primers were synthesized by Applied Biosystems BV. PCR was performed on a PTC-200 Thermal Cycler (MJ Research) with GoTaq® Flexi DNA polymerase/ buffer system (Promega). Reaction volume was 25µl containing 1×buffer, 2mM $MgCl_2$, 0.2mM each of the four dNTPs, 0.4µM each of the forward and reverse primers, 1µl of DNA template (ranging from 20 to 200ng DNA), 0.5U Taq polymerase and de-ionized water. The uniform thermal cycling programme for rpl16, rpl32-trnL, trnS-trnG, and trnT-trnL was pre-denaturing at 80°C for 5 minutes, followed by 35 cycles of denaturing at 95°C for one minute, annealing at 52°C for one minute, a slow ramp at

0.3°C/s up to 65°C and the extension at 65°C for four minutes, with the final extension at 65°C for ten minutes. The thermal cycling programme for rps16 was pre-denaturing at 80°C for 5 minutes, followed by 35 cycles of denaturing at 95°C for 30 seconds, annealing at 52°C for 30 seconds and extension at 72°C for one minute, with the final extension at 72°C for ten minutes. PCR products were screened on 1.5% agarose gels using SYBR Safe DNA gel stain (Invitrogen). Confirmed PCR products were purified using the mi-PCR Purification Kit (Metabion) and sequenced by Eurofins MWG Operon. Sequencing primer sets were the same as the ones used for initial PCR. Completed sequences were checked and aligned within BioEdit 7.0.9 (Hall 1999) and compared between samples from different populations. Selection of optimal loci for HRM analysis was based on evident inter-population variation and outline size of primed PCR amplicons. The sequences of the used primers and their amplicon sizes are listed in Table 2.3.

2.3.2 Locus selection and internal primer design for HRM analysis

The five tested loci varied in size between 640 and 1350bp in length (as in Table 2.3), thus internal primers were needed to divide each of the loci into c.3-400bp amplicons more suitable for HRM analysis.

Table 2.3 Information of the DNA loci and primers used for initial PCR (from Shaw et al. 2005, 2007)

Amplicon name	Forward primer	Reverse primer	Aligned size (bp)
rpl16	5'-GCTATGCTTAGTGTGTGACTCGTTG-3'	5'-CCCTTCATTCTTCCTCTATGTTG-3'	~850
rps16	5'-AAACGATGTGGTARAAAGCAAC-3'	5'-AACATCWATTGCAASGATTCGATA-3'	~750
rpl32-trnL	5'-CAGTTCCAAAAAAACGTACTTC-3'	5'-CTGCTTCCTAAGAGCAGCGT-3'	~910
trnS-trnG	5'-AACTCGTACAACGGATTAGCAATC-3'	5'-GAATCGAACCCGCATCGTTAG-3'	~1350
trnT-trnL	5'-CATTACAAATGCGATGCTCT-3'	5'-TCTACCGATTTCGCCATATC-3'	~640

Table 2.4 Information of amplicons and primers used for HRM analysis with *Arenaria* species

Amplicon name	Forward primer	Reverse primer	Aligned size (bp)
rps16			~750
rps16 I	5'-ATGCTCTTGACTCGACATCTT-3'	5'-GGGTTTAGACATTACTTCGTTGATT-3'	~360
rps16 II	5'-AAGTAATGTCTAAACCcAATGA TTCAA-3'	5'-CGTATAGGAAGTTTTCTCCTCGTA-3'	~390
trnT-trnL			~640
trnTL I	5'-TCTTTAATAATAATATATATAAATT CAAATCAAATTCAAA-3'	5'-GCCATTTGTAATACTCAAAGGATC-3'	~330
trnTL II	5'-CGATCCTTTGAGTATTACAAATGGC-3'	5'-GTCCCAATTTTATGTTTTCCCTTCC-3'	~260

Table 2.5 Information of amplicons and primers used for HRM analysis with *Minuartia recurva*

Amplicon name	Forward primer	Reverse primer	Aligned size (bp)
rps16			~690
rps16 I	5’-GCTCTTGACTCGACATCT-3’	5’-AATGGCAGCAACATACCT-3’	~350
rps16 II	5’-AGGTATGTTGCTGCCATT-3’	5’-TGACCAATCCAATAAGTCCATA-3’	~350
trnT-trnL			~540
trnTL I	5’-TTTTCGTCTAGAGCCATTT-3’	5’-TCGTCTTAGTCTCTGAATGA-3’	~310
trnTL II	5’-TTCATTCAGAGACTAAGACGAA-3’	5’-GGATTAATATACCGAACAGTGTT-3’	~240

The design of internal primers rested on two criteria, (i) the internal primers must bind to conserved regions of the DNA locus to ensure broad intra-specific sensitivity and (ii) as few internal primers as possible were sought for each locus to maximize the efficiency of HRM analysis, as long as allowing that the amplicon sizes did not exceed 400bp.

Internal primers were designed with AlleleID 7 (Premier Biosoft International) using its SYBR® Green Design function. Cross Species Design was conducted with aligned locus sequences from different populations of *A. ciliata* and *A. norvegica*, and the same task was also conducted with *Minuartia recurva*. For primer searching, the target amplicon length was set between 250 and 400bp, the length of primers was set between 18 and 30bp and the primer Ta (annealing temperature) within the range of 55.0±5.0°C. Considering the requirements for appropriate amplicon lengths, optimal coverage of polymorphic sites and conservation of sequence at primer sites, rps16 and trnT-trnL was finally selected for HRM analysis of the three species, and meanwhile the internal primers were designed for *Arenaria* and *Minuartia* separately.

The sequences of the designed internal primers are listed in Table 2.4 and Table 2.5 for *Arenaria* and *Minuartia* separately. For rps16 with *Arenaria*, the two internal amplicons are adjacent with 17bp overlap, covering the

c.750bp length, and for trnT-trnL with Arenaria, the two internal amplicons are adjacent with 25bp overlap, covering the c.640bp overall length. Figure 2.1 shows the relative positions of the primers and the approximate sizes of their amplicons with *Arenaria* species. The same priming pattern was also used with *Minuartia recurva*, with slightly different amplicon sizes as per Table 2.4.

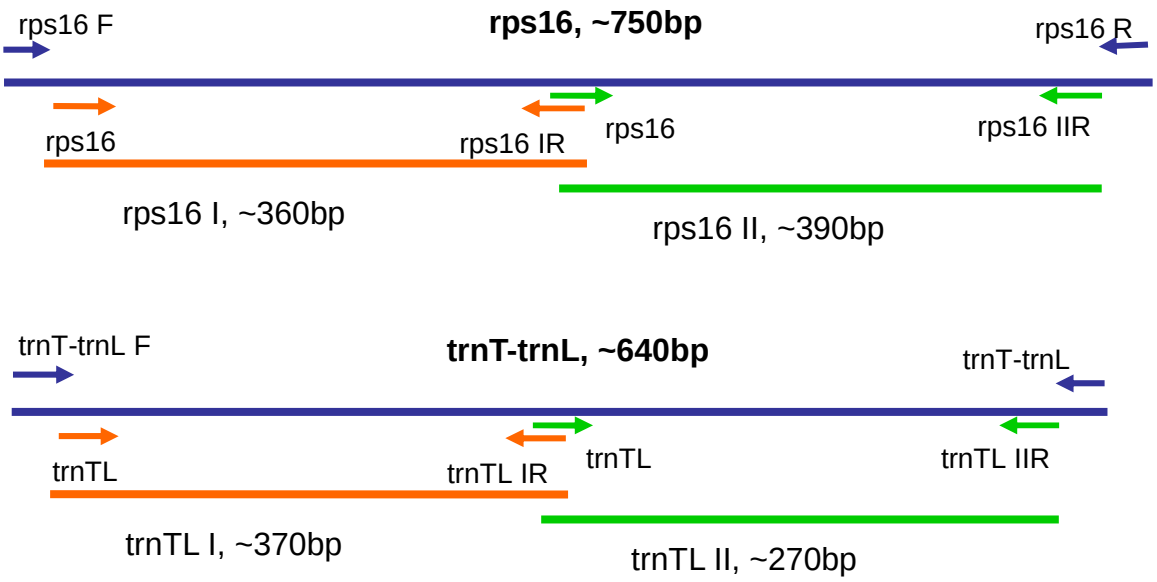


Figure 2.1 Diagram of primers used with the two chloroplast DNA regions for *Arenaria* species. The primers rps16 F, rps16 R, trnT-trnL F and trnT-trnL R were used for initial PCR and sequencing, while rps16 IF, rps16 IR, rps16 IIF, rps16 IIR, trnT-L IF, trnT-L IR, trnT-L IIF, trnT-L IIR were used for HRM analysis based on real-time PCR.

2.3.3 High-Resolution Melting analysis: *in vitro* protocols

Real-time PCR assays were conducted in Roche LightCycler® 480

(LC480) Multiwell plates (96-well white) on the LC480 instrument (Hoffman-La Roche, Basel, Switzerland) using the LC480 HRM Master Mix reagent kit (Roche). The template DNA concentration for each individual sample was measured using a Nanodrop 1000 spectrophotometer (Thermo Scientific), and then diluted to 20–30ng/µl. At the beginning, the reaction system was determined and the real-time PCR programme was optimized to generate the target amplicons correctly, before high-throughput HRM analysis was carried out to screen the populations in batches. Each real-time PCR reaction was conducted in a 15µl final volume containing 7.5µl of the Taq enzyme mix (No. 1 reagent from the Roche HRM Master kit containing Taq polymerase, dNTPs, buffer system and saturating dsDNA binding dye), 1.5µl of 25mM $MgCl_2$ solution (No. 2 reagent from the Roche HRM Master kit, final Mg concentration 2.5mM), 1.5µl of primer mix (final concentration 1µM each), 3.5µl of de-ionized water (No. 3 reagent from the Roche HRM Master kit) and 1µl of DNA template (20 – 30ng). The real-time PCR programmes for different amplicons are listed in Table 2.6 and 2.7 for *Arenaria* and *Minuartia* separately. Fluorescence values for each sample in each running were recorded through the SYBR Green (483-533nm) channel using the default LC480 data acquisition settings.

Table 2.6 Real-time PCR programmes for HRM assays of *Arenaria* species. All reactions completed on Roche LightCycler® 480 system using the SYBR Green I/HRM channel, with 15µl reaction volume.

Amplicon	Stage of HRM Programme	Target Temperature (°C)	Duration (hh:mm:ss)	Ramp rate (°C/s)	Number of Cycles	Acquisitions per °C	Analysis mode
rps16*I* and rps16 *II*	Pre-incubation	95	00:10:00	4.4	1	-	-
	Amplification	95	00:00:15	4.4	30	-	Quantification
		60	00:00:15	2.2		-	
		72	00:00:25	1		Single	
	Melting	95	00:00:05	4.4	1	-	Melting curve
		65	00:01:00	2.2		-	
		97	-	0.01		50	
	Cooling	40	30	1.5	1	-	-
trnT-trnL I	Pre-incubation	95	00:10:00	4.4	1	-	-
	Amplification	95	00:00:10	4.4	40	-	Quantification
		50	00:00:15	2.2		-	
		60	00:00:30	1		single	
	Melting	95	00:00:05	4.4	1	-	Melting curve
		55	00:01:00	2.2		-	
		97	-	0.01		50	
	Cooling	40	30	1.5	1	-	-
trnT-trnL *II*	Pre-incubation	95	00:10:00	4.4	1	-	-
	Amplification	95	00:00:15	4.4	25	-	Quantification
		60	00:00:15	2.2		-	
		72	00:00:25	1		single	
	Melting	95	00:00:05	4.4	1	-	Melting curve
		60	00:01:00	2.2		-	
		97	-	0.01		50	
	Cooling	40	30	1.5	1	-	-

Table 2.7 Real-time PCR programmes for HRM assays of *Minuartia recurva*. All reactions completed on Roche LightCycler® 480 system using the SYBR Green I/HRM channel, with 15µl reaction volume.

Amplicon	Stage of HRM Programme	Target Temperature (°C)	Duration (hh:mm:ss)	Ramp rate (°C/s)	Number of Cycles	Acquisitions per °C	Analysis mode
rps16*I* **and** **trnT-trnL*II***	Pre-incubation	95	00:10:00	4.4	1	-	-
	Amplification	95	00:00:15	4.4	35	-	Quantification
		60	00:00:15	2.2		-	
		72	00:00:20	1		Single	
	Melting	95	00:00:05	4.4	1	-	Melting curve
		70	00:01:00	2.2		-	
		97	-	0.01		50	
	Cooling	40	30	1.5	1	-	-
rps16 *II*	Pre-incubation	95	00:10:00	4.4	1	-	-
	Amplification	95	00:00:15	4.4	35	-	Quantification
		55	00:00:15	2.2		-	
		60	00:00:20	1		single	
	Melting	95	00:00:05	4.4	1	-	Melting curve
		70	00:01:00	2.2		-	
		97	-	0.01		50	
	Cooling	40	30	1.5	1	-	-
trnT-trnL *I*	Pre-incubation	95	00:10:00	4.4	1	-	-
	Amplification	95	00:00:10	4.4	35	-	Quantification
		52	00:00:15	2.2		-	
		60	00:00:25	1		single	
	Melting	95	00:00:05	4.4	1	-	Melting curve
		60	00:01:00	2.2		-	
		97	-	0.01		50	
	Cooling	40	30	1.5	1	-	-

Batch HRM assays were conducted mostly with a single population in each running. Only small populations, e.g. the populations IT2, IT3 and FR1 of *A. ciliata* were assayed with other populations because of their small sizes. The aim of doing it this way is to manage inter-individual and inter-plate errors arising from variation of template concentration and quality across samples from different population extractions. In addition, one or two individual accessions from the previous runs were always included in a new run with the same amplicon, with the aim of giving an idea of inter-batch error, and also serving as positive controls. Each individual accession from each population was run in duplicate (in cases where the batch size is around 30) or triplicate (in cases where the batch size is around 20) for HRM analysis depending on the batch size in the run.

2.3.4 Interpretation and analysis of HRM curve profiles

Peak Tm Calling Analysis was performed after each real-time PCR reaction run within the LC480 Software (release 1.5.0 SP3, Version 1.5.0.39), based on which clear melting peaks and shoulders were revealed for the amplicons. The temperatures at the melting peaks are recorded as Tm values. Figure 2.2 shows the examples of Tm readings based on the *Arenaria* results, where two melting peaks recorded as Tm1

and Tm2 are revealed with amplicon rps16*I* (Figure 2.2A and B), while one prominent clear peak, named Tm3, and a less prominent lower temperature shoulder (TmX) were revealed for amplicon rps16*II* (Figure 2.2C and D). Tm values for the obtained melting peaks were given initially by automatic Tm calculation under 'SYBR Green I Format' where the number of 'Maximal Peaks' was set as '2 or less' by default, and these values were manually adjusted to mark the temperatures where the melting peaks were at their maximum heights.

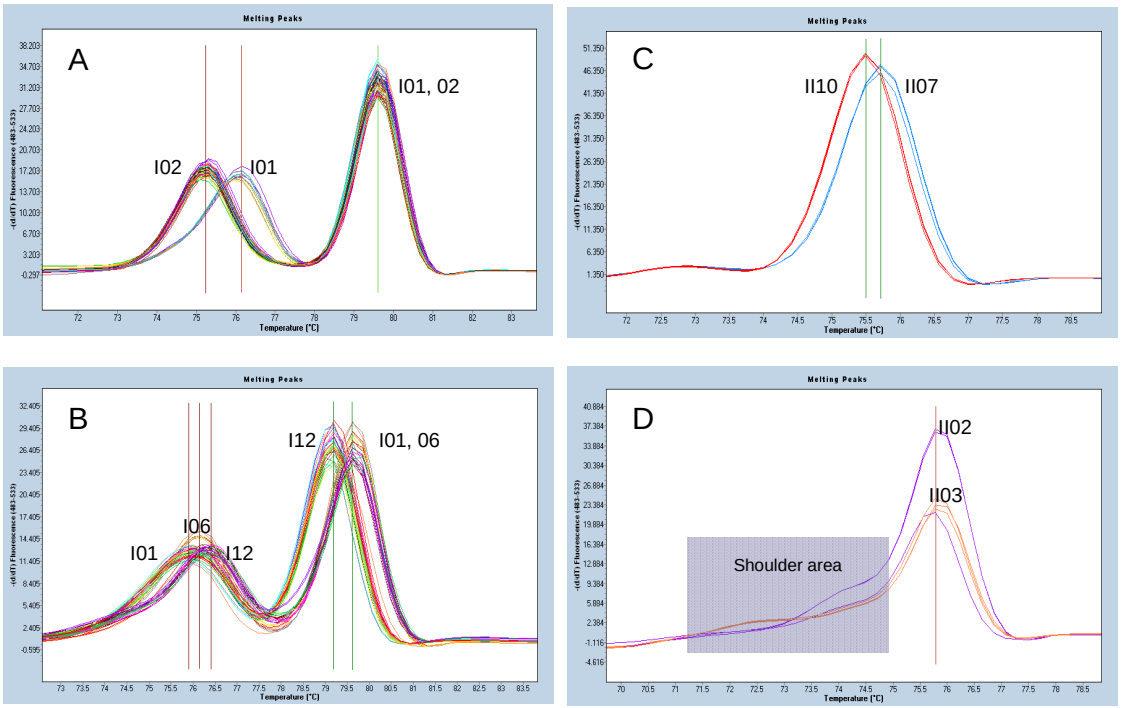


Figure 2.2 Sensitivity of *in vitro* HRM to haplotype variation. **A.** The melting profiles of two amplotypes of rps16*I* found in Ir1. **B.** The melting profiles of three amplotypes of rps16*I* found in Au2. **C.** The melting profiles of two amplotypes of rps16*II* found in It1 and It3, which differ by a single G to C mutation. **D.** The melting profiles of two amplotypes of rps16*II* found in Ir1 and Ir2, which differ by a large indel. The melting curves of rps16*I* provide double melting peaks, both of which can show variation between amplotypes. The melting curves of rps16*II* provide one melting peak with a less prominent shoulder, which can also help distinguish between amplotypes.

Analysis runs were carried out on the central 60 wells of the 96-well plate to minimize the possibility of edge-effect variation. As described in 2.2.3, two or three replicates were used to validate the consistency of the melting profiles for each analyzed accession, based on the majority law. When two replicates were used and they show divergent melting profiles, the accession was re-analyzed within another assay until the consistent melting profile was obtained for that individual. Final Tm values were manually validated for each accession based on the standardized curves of the replicates for each amplicon, which were then used to determine the identities of the amplicons.

Two additional analyses were carried out to validate the consistency of the method; (i) single-plate analysis of multiple accessions of one amplotype drawn from several different populations, to evaluate the relative variation in HRM profiles associated with population-specific factors and (ii), full-plate replicates comprising 60 individual HRM analyses of a single sample accession, to evaluate well-to-well variation across the plate. These consistency validating assays were conducted with the *Arenaria* I01 amplotype of rps16*I* .

2.3.5 Determination of haplotype identities based on melting profiles

The confirmation of the amplicon identity (hereafter termed '*amplotype*') for each sample was carried out using a standard protocol. Within each run, samples were first grouped into different hypothesized amplotypes based on their validated Tm values. After initial analysis runs, variation of ≥ 0.2°C in Tm was considered the likely level of divergence which would be evident between variant amplotypes, with the understanding that ongoing analysis might alter this assumption (no guidance was available from the literature on haplotype Tm divergence in amplicons larger than 150bp). Sample profiles that differed by ≥ 0.2°C at any of the 3 Tm peaks or the shoulder region were regarded as having potentially distinct amplotypes, the Tm value difference of each melting peak between putative amplotypes termed 'ΔTm'.

For each discrete Tm group within each population the amplotype sequence identity was obtained. Where within-group Tm variance was evident but at less than 0.1°C, two individuals were randomly selected from this group to be sequenced over the whole region of the DNA locus (e.g. the whole length of rps16 was sequenced once variation was suspected within rps16*I* or *II*). When within-group variance was between 0.1°C and 0.2°C, the individuals with the lowest and the highest Tm

values were chosen to be sequenced for confirmation. If different amplotypes were detected within a pre-classified group in the latter manner, one or two additional samples from each new subgroup were chosen for sequencing to confirm the revised amplotype groupings. Each putative amplotype recorded was assigned an identity, e.g. *I*01, *I*02 for rps16*I* and *II*01, *II*02 for rps16*II* etc, before the haplotype of the locus was finally determined by combining the amplotype identities, e.g. the rps16 haplotype identity was obtained by concatenating the amplotype identities of rps16*I* and *II*.

To control inter-population errors, the sampling protocol required that in each population a minimum of two replicates of each putative amplotype were sequenced, even where the same melting curve profile had been encountered in another population. This strategy was judged to be necessary as a second level of assurance in addition of the evaluation of inter-batch errors by inclusion of reference individuals between batches (section 2.2.3). The inter-population deviation is perceivable based on the repeatability evaluation (described in 2.2.4 with results shown in Table 2.9) despite the putative Tm values for each amplotype can be managed through manual adjustment based on the use of inter-batch reference samples (described in 2.2.3 and the adjusted Tm values are seen in Table 2.10). On the other hand, there is a theoretical possibility that different

amplotypes may share similar or even identical melting profiles, which need to be assessed at both inter- and intra-population levels. The above strategy has thus been adopted in order to exclude inter-population errors and to minimize possible missed detection.

2.3.6 *In silico* simulation of HRM analysis and correlation tests

Using the web based software uMelt^{SM} (Dwight *et al.* 2011), the sequence identities of all the verified amplotypes were input to *in silico* HRM analysis. The *in silico* Tm values were obtained for each melting peak and were compared to the corresponding *in vitro* Tm values. The thermodynamic set of Huguet *et al.* (2010) was chosen as this is the latest published compared to those of other authors. The free Mg^{2+} concentration was set at 2.5mM as used with the in vitro analysis, and the DMSO concentration was set at 10% to give similar Tm values as seen in the *in vitro* assays. The correlation analysis between the *in silico* and *in vitro* Tm values were conducted via Mantel test (Mantel 1967) using the R software (R Development Core Team 2011).

The correlation between Tm difference and genetic distance was also analyzed via Mantel test (Mantel 1967) by adding up the Tm1, Tm2 and Tm3 differences between each pair of haplotypes of rps16 to be tested

against their Tajima-Nei's *D* (Tajima & Nei 1984) as the genetic distance.

2.4 Results

2.4.1 Haplotype detection for *Arenaria* species

2.4.1.1 Uniformity of the assay within a plate

In order to evaluate the possible across-well variation on a single plate, 60 replicates of a single PCR reaction with rps16*I* of the individual Ir1.25, an *A. ciliata* sample, were made in wells from B2 to G11 (as illustrated in Figure 2.3) on a 96-well white plate supplied by Roche. The side wells were not used to remove possibility of any edge effects. The readings of Tm1 and Tm2 are listed in Table 2.8. Visualized cross-well variation is shown in Figure 2.4 and 2.5. Fifty eight out of the sixty replicates have their Tm1 values between 75.31°C and 75.37°C (Mean 75.32°C, Median 75.32°C, s.d. 0.013°C) with two outliers at 75.21°C (in the well G5) and 75.24°C (in the well G11). The maximum deviation for Tm1 is 0.06°C excluding the outliers and 0.16°C including the outliers. A similar pattern is also shown with Tm2 values, where the majority of the replicates have Tm2 between 79.59°C and 79.63°C (Mean 79.61°C, Median 79.61°C, s.d. 0.007°C), with three outliers at 79.56°C (B5),

79.51°C (G6) and 79.56°C (G11). The maximum deviation for Tm2 is 0.04°C/0.12°C (excluding/including outliers). From the result it is seen that the systems error between replicates in a single run on a plate is lower than the 0.2°C threshold for discriminating between discrete haplotypes.

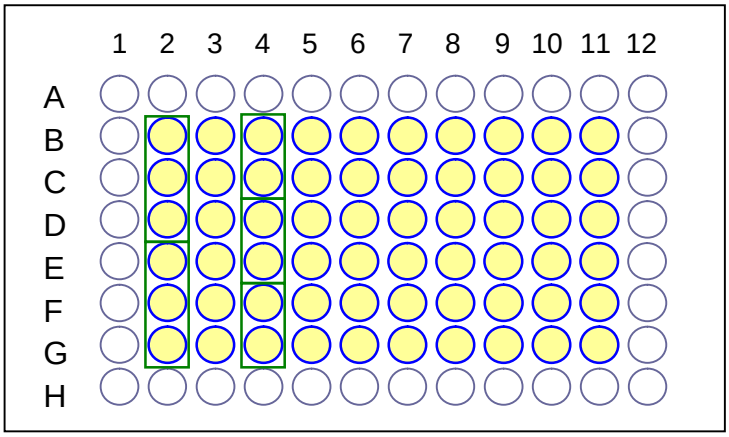


Figure 2.3 The schematic diagram of a 96-well plate. The wells with blue borders and yellow fillings are used for real-time PCR reactions. In batch assays each individual sample is run in 2 or 3 replicates, as shown in green squares.

Table 2.8 Tm1 readings from 60 wells on the plate (in °C)

Tm1										
	2	3	4	5	6	7	8	9	10	11
B	75.37	75.31	75.32	75.31	75.32	75.32	75.31	75.32	75.32	75.34
C	75.33	75.32	75.32	75.31	75.32	75.32	75.32	75.31	75.32	75.32
D	75.33	75.33	75.33	75.32	75.32	75.32	75.32	75.32	75.31	75.32
E	75.35	75.37	75.35	75.31	75.32	75.31	75.32	75.32	75.32	75.32
F	75.33	75.32	75.32	75.32	75.32	75.31	75.32	75.32	75.32	75.32
G	75.35	75.32	75.32	75.32	75.21	75.32	75.31	75.31	75.31	75.24

Tm2										
	2	3	4	5	6	7	8	9	10	11
B	79.61	79.61	79.59	79.56	79.61	79.61	79.60	79.61	79.62	79.62
C	79.61	79.61	79.60	79.61	79.61	79.61	79.61	79.61	79.61	79.61
D	79.63	79.61	79.61	79.61	79.61	79.61	79.61	79.61	79.61	79.62
E	79.61	79.62	79.61	79.61	79.61	79.61	79.61	79.61	79.63	79.62
F	79.61	79.61	79.61	79.61	79.61	79.61	79.61	79.61	79.61	79.61
G	79.63	79.61	79.61	79.61	79.51	79.61	79.61	79.61	79.61	79.56

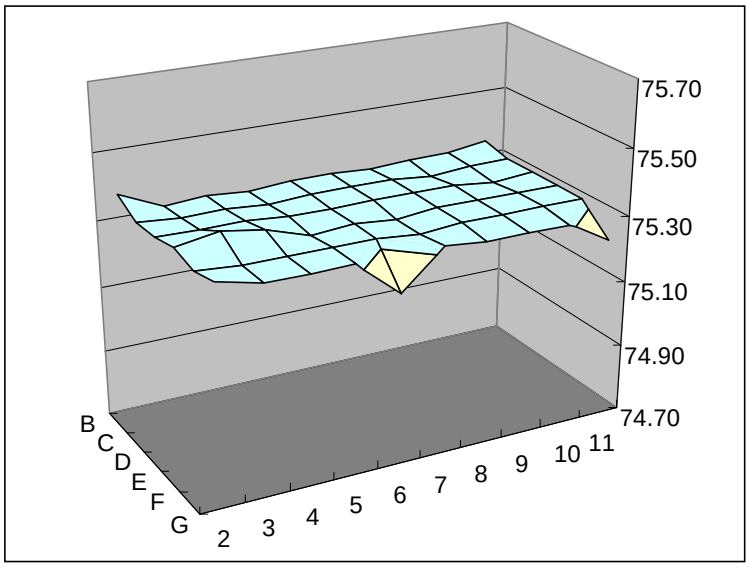


Figure 2.4 The thermal variation among 60 replicates on a 96-well plate in LightCycler 480 based on the Tm1 readings with rps16*I* amplified from the individual sample of *A. ciliata*, IR1.25.

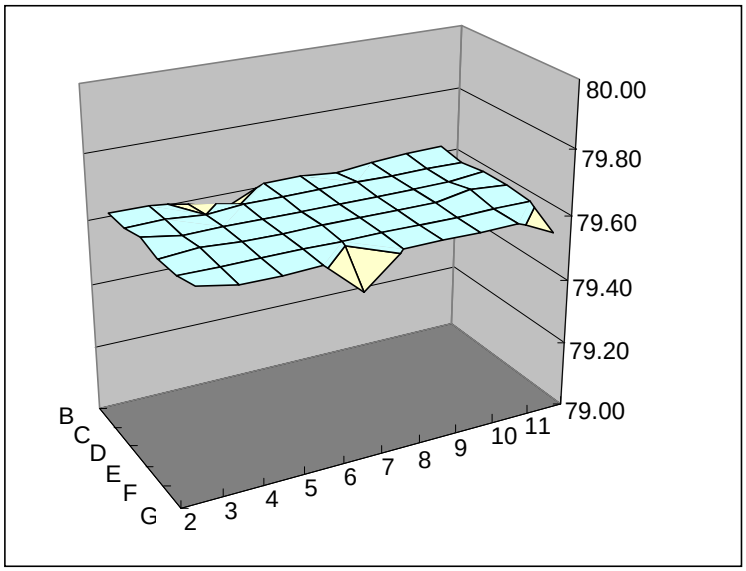


Figure 2.5 The thermal variation among 60 replicates on a 96-well plate in LightCycler 480 based on the Tm2 readings with rps16*I* amplified from the individual sample of *A. ciliata*, IR1.25.

2.4.1.2 Inter-batch repeatability of HRM analysis

After amplotype grouping and haplotype determination via direct sequencing of representative accessions from each population, the individuals listed in Table 2.9 were confirmed to share the same haplotype of rps16, recorded as rpsC01. They were then put on the same plate in a single run for each of rps16*I* and *II*. Tm values were recorded for each replicate of each individual, with the aim of assessing the errors at both intra- and inter-individual levels.

It is seen from Table 2.9 that the inter-individual variation is generally greater than intra-individual variation across replicates, however the inter-individual variation in Tm2 and Tm3 is comparable with or even less than the inter-replicate variation in the plate uniformity test described in 2.3.1.1. The greater inter-individual variation is seen in Tm1 within this haplotype, which is as great as 0.23°C, mainly due to the low Tm1 values of the two individuals, Pi2.7 and Pi2.15 from the Picos Population. The variation between these two individuals, however, is as small as 0.02°C, which implies that they may have become outliers due to unknown factors specific to the Picos populations.

Table 2.9 Tm values for individual accessions with the same haplotype (rpsC01) for rps16 *I* and *II*

Sample	rps16*I*				rps16*II*	
	Tm1 (°C)		Tm2 (°C)		Tm3 (°C)	
Ir1.15	76.39	Mean	79.75	Mean	76.32	Mean
	76.37	76.37	79.73	79.74	76.32	76.32
	76.35	s.d. 0.02	79.75	s.d. 0.01	76.32	s.d. 0.00
Ir1.24	76.31	Mean	79.73	Mean	76.32	Mean
	76.31	76.31	79.73	79.74	76.32	76.32
	76.31	s.d. 0.00	79.75	s.d. 0.01	76.32	s.d. 0.00
Ir2.13	76.20	Mean	79.64	Mean	76.26	Mean
	76.20	76.20	79.64	79.64	76.26	76.26
	76.20	s.d. 0.00	79.64	s.d. 0.00	76.26	s.d. 0.00
Ir2.15	76.24	Mean	79.63	Mean	76.37	Mean
	76.24	76.24	79.63	79.63	76.39	76.38
	76.24	s.d. 0.00	79.63	s.d. 0.00	76.37	s.d. 0.01
Ir3.13	76.22	Mean	79.65	Mean	76.22	Mean
	76.22	76.21	79.65	79.64	76.22	76.22
	76.20	s.d. 0.01	79.63	s.d. 0.01	-	s.d. 0.00
Ir3.16	76.22	Mean	79.72	Mean	76.30	Mean
	76.22	76.23	79.72	79.72	76.30	76.30
	76.24	s.d. 0.01	79.72	s.d. 0.00	76.30	s.d. 0.00
Pi2.7	76.15	Mean	79.62	Mean	76.20	Mean
	76.15	76.16	79.63	79.63	76.20	76.20
	76.17	s.d. 0.01	79.63	s.d. 0.006	76.20	s.d. 0.00
Pi2.15	76.14	Mean	79.64	Mean	76.21	Mean
	76.14	76.14	79.64	79.64	76.21	76.21
	76.14	s.d. 0.00	79.64	s.d. 0.00	76.21	s.d. 0.00
Maximum deviation		0.23		0.12		0.17
Standard deviation		0.08		0.05		0.06

The impurity and thus chemical difference are suspected to exist in the extracted DNA solutions between the two and other individuals, which may potentially have affected the repeatability of HRM analysis for individuals from different populations. The populations are collected from various places at different times, and then stored for different periods before their DNA materials were extracted through the traditional CTAB method, which may not guarantee the chemical uniformity in the final DNA solutions.

2.4.1.3 Haplotype detection in the sampled populations with rps16

Within the 484 individual samples from 23 populations of *A. ciliata* and *A. norvegica*, 14 amplotypes of rps16I and 14 amplotypes of rps16II were detected, combined to yield a total of 20 haplotypes of the overall rps16 locus, 18 of them for *A. ciliata* and 2 for *A. norvegica* (Tables 2.10-12, cited from Dang *et al.* 2012). It should be noticed that the Tm variation may occur between individuals sharing the same amplotype due to random human errors, inter-population chemical variation of DNA templates and/or inter-batch system errors, thus the Tm values included in Table 2.10 are adjusted mean values rounded by 0.05°C for each amplotype, where inter-batch (and thus part of inter-population) errors are minimized with the aid of the reference samples used for inter-batch cross

check and random inter-individual errors are eliminated by averaging the values. Recorded haplotypes differed at multiple nucleotide sites by both nucleotide substitutions and insertion-deletion events (Table 2.11 and 12).

Table 2.10 Summary of all rps16 haplotypes identified in *A. ciliata* (rpsC-) and *A. norvegica* (rpsN-) using HRM *in vitro* and *in silico* analysis of rps16*I* and rps16*II* amplicons.

Composite rps16 haplotype	rps16*I* amplotype	*In vitro* Tm1 (°C)	*In silico* Tm1 (°C)	*In vitro* Tm2 (°C)	*In silico* Tm2 (°C)	rps16*II* amplotype	*In vitro* Tm3 (°C)	*In silico* Tm3 (°C)	*In silico* TmX (°C)
rpsC01	*I*01	76.25	77.7	79.65	80.6	*II*01	76.15	77.7	75.8
rpsC02	*I*02	75.35	77.3	79.65	80.6	*II*02	75.70	77.4	76.3
rpsC03	*I*02	75.35	77.3	79.65	80.6	*II*03	75.70	77.4	75.7
rpsC04	*I*03	75.45	77.5	79.65	80.6	*II*04	76.05	77.7	76.5
rpsC05	*I*04	76.35	77.7	79.65	80.7	*II*01	76.15	77.7	75.8
rpsC06	*I*01	76.25	77.7	79.65	80.6	*II*05	76.15	77.7	76.3
rpsC07	*I*01	76.25	77.7	79.65	80.6	*II*06	76.00	77.7	76.3
rpsC08	*I*05	75.95	77.5	79.85	80.8	*II*05	76.15	77.7	76.3
rpsC09	*I*06	76.45	78.1	79.65	80.6	*II*07	75.80	77.3	76.1
rpsC10	*I*07	76.40	78.1	79.95	80.9	*II*08	75.60	77.2	76.1
rpsC11	*I*08	76.50	78.0	79.10	80.3	*II*09	75.70	77.4	76.3
rpsC12	*I*09	75.95	77.5	79.65	80.6	*II*05	76.15	77.7	76.3
rpsC13	*I*06	76.45	78.1	79.65	80.6	*II*10	75.60	77.0	76.1
rpsC14	*I*01	76.25	77.7	79.65	80.6	*II*11	76.15	77.7	76.3
rpsC15	*I*10	75.60	77.5	79.00	80.0	*II*12	76.55	77.9	76.3
rpsC16	*I*11	76.35	77.8	79.65	80.6	*II*07	75.80	77.3	76.1
rpsC17	*I*01	76.25	77.7	79.65	80.6	*II*13	76.10	77.7	76.6
rpsC18	*I*12	75.85	77.9	79.10	80.3	*II*09	75.70	77.4	76.3
rpsN01	*I*13	76.10	77.4	79.65	80.6	*II*05	76.15	77.7	76.3
rpsN02	*I*14	75.05	77.0	79.65	80.6	*II*14	76.50	77.9	76.3

Table 2.11 Polymorphic nucleotide sites within the rps16*I* region aligned among the revealed haplotypes of *Arenaria* species. The shaded nucleotides show the sites where a SNP is the sole difference between two amplotypes.

Haplotype of rps16*I*	Nucleotide sites (363bp aligned)																	
	29	49	93	113	124-129	160	166	217	231-237	241	245	246	252	290-295	296	306	314	321-325
I01	G	T	C	A	AAAGAA	G	G	G	ATATATC	T	A	C	C	------	C	T	G	-----
I02	G	T	C	A	AAAGAA	G	G	G	-------	T	A	A	C	TTATAA	C	T	G	-----
I03	G	T	C	A	AAAGAA	G	G	G	ATATATC	T	A	C	C	------	C	T	G	TTTTT
I04	G	T	C	A	------	G	G	G	ATATATC	T	A	C	C	------	C	T	G	-----
I05	G	T	C	C	AAAGAA	G	T	G	-------	T	A	A	C	------	C	T	G	-----
I06	G	T	C	A	AAAGAA	G	G	G	ATATAGC	C	A	A	C	------	C	T	C	-----
I07	G	G	C	A	AAAGAA	G	G	G	ATATAGC	C	A	A	C	------	C	T	C	-----
I08	A	T	C	A	AAAGAA	G	G	G	ATATATC	C	C	A	G	------	C	T	C	-----
I09	G	T	C	A	AAAGAA	G	T	G	-------	T	A	A	C	------	C	T	G	-----
I10	G	T	T	A	AAAGAA	G	T	G	-------	T	A	A	C	------	C	T	G	-----
I11	G	T	C	A	AAAGAA	T	G	G	ATATAGC	C	A	A	C	------	C	T	C	-----
I12	A	T	C	A	AAAGAA	G	G	G	ATATATC	C	C	A	G	------	T	T	C	-----
I13	G	T	C	A	AAAGAA	G	T	T	-------	T	A	A	C	------	C	G	G	-----
I14	G	T	C	A	AAAGAA	G	G	G	-------	T	A	A	T	TTATAA	C	T	G	-----

Table 2.12 Polymorphic nucleotide sites within the rps16*II* region aligned among the revealed haplotypes of *Arenaria* species. The shaded nucleotides show the sites where a SNP is the sole difference between two amplotypes.

Haplotype of rps16*II*	Nucleotide sites (431bp aligned)																				
	38	79	94	111	119	129-155	179	183	186	194	200	213	221	223-227	281	291	303	322	336	338	354-359	372-384
II01	-	C	G	T	C	----------	T	C	A	G	T	T	T	-----	T	T	T	C	C	T	TAT	----------
																					ATA	
II02	-	C	G	T	G	----------	T	T	A	G	T	T	T	-----	T	T	T	C	C	T	------	----------
II03	-	C	G	T	G	----------	T	T	A	G	T	T	T	-----	T	T	T	C	C	T	------	ATATAGA
																						TATAAT
II04	-	C	G	T	G	----------	T	C	A	G	T	T	T	-----	T	G	T	C	C	T	------	----------
II05	-	C	G	T	G	----------	T	C	A	G	T	T	T	-----	T	T	T	C	C	T	------	----------
II06	-	C	G	T	G	----------	T	C	A	G	T	T	A	TTG	T	T	T	C	C	T	------	----------
														AT								
II07	T	T	G	T	G	TTAAATTGATTCTAA	T	C	A	T	G	-	T	-----	G	T	T	T	C	T	TAT	----------
						ATGAGACACAAC															AGA	
II08	T	T	G	T	G	----------	T	C	A	T	G	-	T	-----	G	T	T	T	C	T	TAT	----------
																					AGA	
II09	-	G	A	T	G	----------	T	C	T	G	T	T	T	-----	G	T	T	C	T	G	------	----------
II10	T	T	G	T	G	TTAAATTGATTTTAA	T	C	A	T	G	-	T	-----	G	T	T	T	C	T	TAT	----------
						ATGAGACACAAC															AGA	
II11	-	C	G	T	C	----------	T	C	A	G	T	T	T	-----	T	T	T	C	C	T	------	----------
II12	-	C	G	T	G	----------	C	C	A	G	T	T	T	-----	T	T	T	C	C	T	------	----------
II13	-	C	G	T	G	----------	T	C	A	G	T	T	T	-----	T	T	C	C	C	T	------	----------
II14	-	C	G	G	G	----------	T	C	A	G	T	T	T	-----	T	T	T	C	C	T	------	----------

Rps16*I* data

All except 2 pairs of amplotypes of rps16*I* were readily distinguishable from one another by the combination of Tm1 & Tm2 values, including 6 cases where single SNPs were the sole difference between amplotype templates. Amplotypes rps16*I*04 and *I*11 differed greatly in sequence composition (by 5 substitutions and one 6-bp indel) but displayed the same Tm1 and Tm2 values during the in vitro HRM analysis. These amplotypes are seen to be endemic to populations in Ireland (Ir3) and Spain (Pi2), respectively, each with only one representative individual sample discovered. If there were any other individual samples in each of the populations sharing the same Tm1 and Tm2 values as the two amplicons, at least two of the individuals from each population were required to be sequenced for haplotype confirmation. However it was unnecessary in this case because only one individual was found in each population showing the above combination of Tm1 and Tm2 values. In addition the composite rps16 sequence of both amplotypes differed in rps16*II* and the respective Tm3 values were clearly distinct, providing a valid HRM diagnostic. Amplotypes *I*06 and *I*11 were only distinguished by 0.1°C difference in Tm1, which is below the normal threshold. However they are confirmed not co-occurring in the same populations.

Rps16*II* data

Amplicon melting curves for rps16*II* showed lower resolution than for rps16*I*. Three different amplotypes, *II*01, *II*05 and *II*11 that differ by one G/C substitution and/or a 6bp indel have the same Tm3 at 76.15°C. *II*01 can be distinguished from both of the latter at TmX, however *II*05 and *II*11 (which co-occur in population Py2 as part of composite haplotypes C06 and C14 respectively, Figure 2.6) cannot be distinguished from one another based on the rps16*II* melting curve. Amplotype *II*09 differs in sequence content markedly (minimum 7 point mutations) from both *II*02 and *II*03, but has the same Tm3 value as both (75.70°C). The latter two amplotypes differ by a single indel of 13bp, which is associated with divergent TmX shoulder curves between the two (Figure 2.2D), however none of these rps16*II* amplotypes co-occur within a single population (Figure 2.6). The averaged Tm3 values of two geographically distant rps16*II* amplotypes *II*12 (*A. ciliata* Spain) and *II*14 (*A. norvegica* Scotland) varied by only 0.05°C, which is below the limit of HRM resolution in out study, however they are not found in the same populations as well (Figure 2.6).

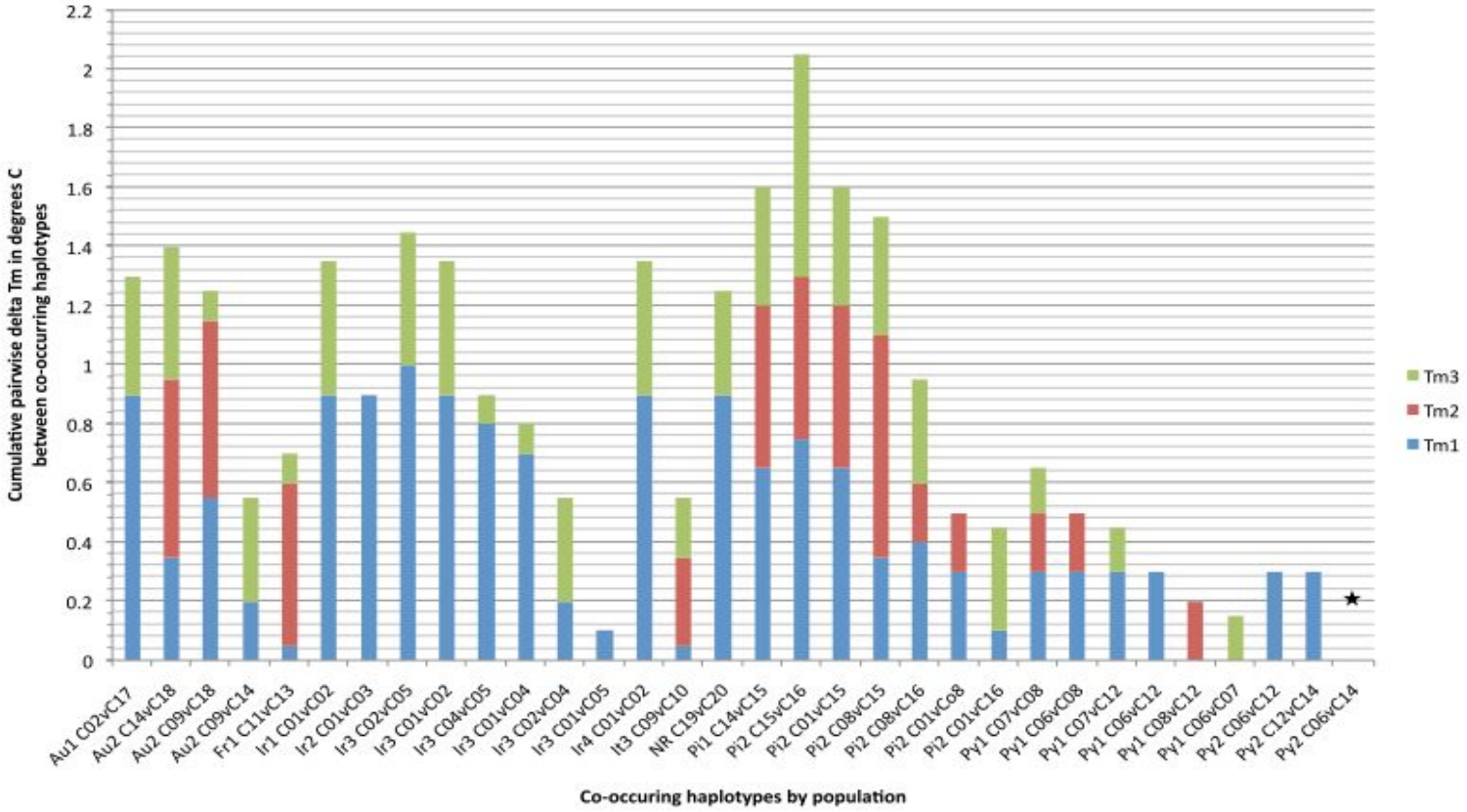


Figure 2.6 Observed total pairwise ΔTm between haplotypes that co-occur in the same sampled population based on combined differences in Tm1, Tm2 and Tm3 values. Only Py2 CO6 v C14 (corresponding to RTC05 and 09) failed to yield any discrete ΔTm value between the haplotypes, indicated by a star. Cited from Dang *et al.* 2012 with kind permission from collaborator authors.

In vitro sensitivity to SNPs

Among all possible pairwise comparisons between amplotypes, single nucleotide polymorphisms constituted the sole difference between templates in 12 cases (shown by shaded nucleotides in Table 2.11-12), 11 of which were identifiable by HRM (e.g. *II*07 and *II*10, Figure 2.2C). In 8 cases the pairwise ΔTm was >0.2^{o}C. In two cases (*II*04 vs. *II*05 and *II*05 vs. *II*13) the Tm3 discrimination was <0.2^{o}C but the detection was achieved using the TmX shoulder in rps16*II*, and in one case (*I*06 vs. *I*11) a shift of <0.2^{o}C was evident in Tm1 in rps16*I*. Only in one case (*II*05 vs. *II*11) a class III transversion SNP in rps16*II* (G to C) was not discernable via initial HRM analysis. Overall class I transition SNPs (G to T or A to C) were the most easily detected (5 of 6 ΔTm >0.2^{o}C), followed by class II transversion SNPs (A to C or G to T) (3 of 5 ΔTm >0.2^{o}C). No class IV transversion SNPs were recorded.

No differences were discernable in the melting curve profile between *II*05 and *II*11, and *II*05 was initially detected by sequencing at least 2 samples that showed the putative *II*11 amplotype curve for rps16*II* in population Py2. Sequence analysis of all 15 putative *II*11 amplotypes in Py2 confirmed just one individual of *II*05.

Sensitivity to differences between co-occurring haplotypes

Here one case with rps16*I* and three cases with rps16*II* were seen where different amplotypes were concealed by similar or even identical DNA melting profiles, i.e. the combination of Tm values. However, only in one of the cases two of the undistinguishable amplotypes were found to co-occur within the same population, but were revealed by their linkage to different rps16*I* amplotypes. The strategy of sequencing the whole rps16 locus instead of only the rps16*I* or *II* amplicons provides additional insurance to avoid missing detection, based on the assumption of the linkage of mutations within the two adjacent amplicons.

Among co-occurring haplotypes within the sampled populations, only 3 of 32 inter-haplotype comparisons failed to yield at least one Δ Tm value that exceeded the nominal discriminating threshold of 0.2°C for Tm1, Tm2 or Tm3 (Figure 2.3). Two of these inter-haplotype comparisons were reliably distinct below the 0.2°C threshold; C06/ C07 in Py1 (ΔTm3=0.15°C) and C01/C05 in Ir3 (ΔTm1=0.1°C). Only one inter-haplotype comparison, C06/ C14 in Py2 (ΔTm1, 2, 3 = 0), as discussed above, was not distinguishable by HRM. The combined array of Tm1, Tm2, Tm3 and TmX values thus provided a unique identifier for 18 of 20 composite haplotypes identified in the analysis and validated by sequencing (Table 2.4, Figure 2.3). While the overall sequence composition for each of these

composite haplotypes was unique, seventeen of the twenty shared at least one rps16*I* or *II* sequence identity with another haplotype, only three composite haplotypes included entirely unique sequences in both regions (C04, C10 and C15).

2.4.1.4 Haplotype detection in the sampled populations with trnT-trnL

Within the same set of samples of *A. ciliata* and *A. norvegica*, 23 amplotypes of trnT-trnL*I* and 8 amplotypes of trnT-trnL*II* were detected, combined to yield a total of 24 haplotypes of the overall rps16 locus, 20 of them for *A. ciliata* and 4 for *A. norvegica* (Table 2.13). The Tm values included in Table 2.13 are adjusted mean values rounded in 0.05°C for each amplotype, where the reference samples were used for inter-batch cross-checking. Recorded haplotypes differed at multiple nucleotide sites by both nucleotide substitutions and insertion-deletion events (Table 2.14-15).

In the amplified DNA regions of trnT-trnL, a few micro-satellite (or simple sequence repeats, SSR) sites were identified, including 220-241bp and 333-341bp within trnT-trnL*I* and 152-154bp within trnT-trnL*II*, which all happened to be single adenine repeats. They were included for

haplotype identification and in Tm analysis. However it turns out that the *in vitro* HRM analysis was not sensitive to variation on these SSR sites. One example is the simple adenine repeats at 152-154bp within trnT-trnL*II*, where the difference among *II*06, *II*07 and *II*08 cannot be distinguished by *in vitro* HRM analysis, although *in silico* HRM simulation predicts some Tm variation should occur between *II*06 and *II*07/08. In trnT-trnL*I*, *I*02, *I*11 and *I*23are only different in their adenine repeat numbers at 333-341bp, and are undistinguishable based on *in silico* HRM simulation. Their *in vitro* Tm differences are small (ΔTm=0.15°C between *I*02 and *I*11/23), which may be no more than an inter-batch or inter-population error. However, the three amplotypes (corresponding to the haplotypes TLC02, TLC12 and TLN04) are found in different populations from Ireland, Austria and Britain respectively. The individuals carrying the three different haplotypes were confirmed by posterior DNA sequencing.

With trnT-trnL*II*, the melting curves of some of the amplotypes showed a shoulder or a less prominent peak at 71.5-72.5°C. This was not predicted by the *in silico* simulation and not shown in Table 2.13. However, the shoulder helped to distinguish *II*01 from *II*06/07/08 which shared the same Tm3 value, as *II*01 carries a low melting peak at 71.7°C while *II*06/07/08 only show a shoulder around 72.4°C (Figure 2.7A). In the

other case, *II*01 and *II*04 share the same Tm3 value but have a second melting peak at 71.7°C and 72.1°C respectively. While the difference between 71.7°C and 72.1°C for the second peak is small, the difference in the overall shape of the melting curves also helps distinguish the two amplotypes (Figure 2.7B).

(See Figure 2.7 on next page.)

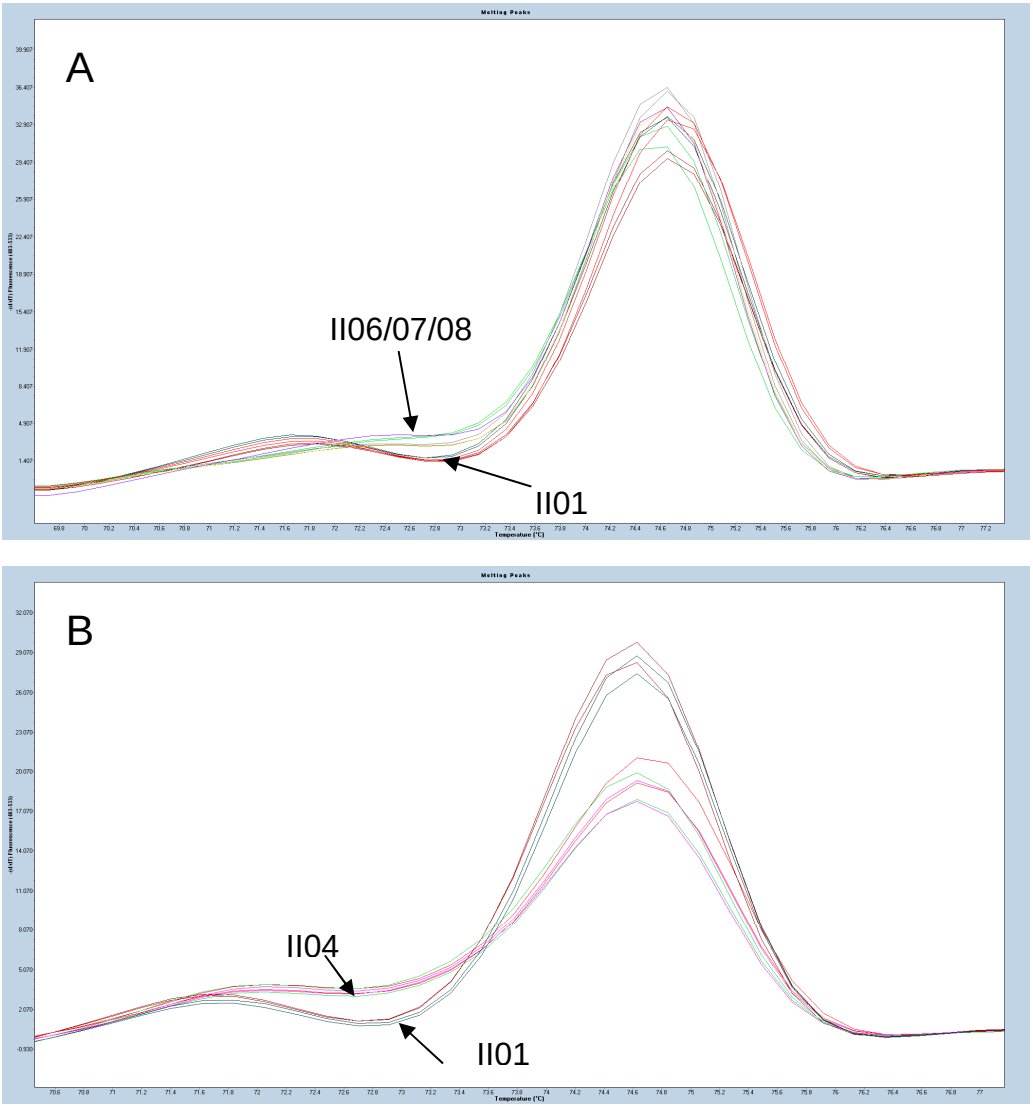


Figure 2.7 Example melting curves of the amplicon trnT-trnL*II* for haplotypes sharing the same Tm3 value at the main melting peak. **A.** The difference between *II*01 and *II*06/07/08 is shown by the position of the lower temperature melting peak ('shoulder'). **B.** The difference between *II*01 and *II*04 is shown not only by the position of the lower melting peak but also by the overall shape of the melting peak.

Table 2.13 Summary of all trnT-trnL haplotypes identified in *A. ciliata* (TLC-) and *A. norvegica* (TLN-) using HRM *in vitro* and *in silico* analysis of trnT-trnL*I* and trnT-trnL*II* amplicons

Composite trnT-trnL haplotype	trnT-trnL*I* amplotype	*In vitro* Tm1 (°C)	*In silico* Tm1 (°C)	*In vitro* Tm2 (°C)	*In silico* Tm2 (°C)	trnT-trnL*II* amplotype	*In vitro* Tm3 (°C)	*In silico* Tm3 (°C)
TLC01	*I*01	67.90	71.0	71.10	73.5	*II*01	74.65	76.5
TLC02	*I*02	67.35	70.6	70.75	73.1	*II*02	74.45	76.2
TLC03	*I*03	67.85	71.0	70.05	73.4	*II*01	74.65	76.5
TLC04	*I*04	67.75	71.0	70.10	73.4	*II*01	74.65	76.5
TLC05	*I*05***	67.75*	70.6/71.0	70.85*	73.5	*II*01	74.65	76.5
TLC06	*I*05***	68.25*	70.6/71.0	71.30*	73.5	*II*03	74.85	76.6
TLC07	*I*06	68.45	71.8	71.10	73.5	*II*01	74.65	76.5
TLC08	*I*07	67.75	71.0	70.75	73.2	*II*04	74.65	76.2
TLC09	*I*08	67.10	70.3	71.10	73.5	*II*01	74.65	76.5
TLC10	*I*09	67.25	70.6	70.30	72.7	*II*05	74.65*	76.0
TLC11	*I*10	67.30	70.6	70.75	73.1	*II*02	74.45	76.2
TLC12	*I*11	67.20	70.6	70.75	73.1	*II*02	74.45	76.2
TLC13	*I*12	68.25	71.4	71.10	73.5	*II*06	74.65	76.4
TLC14	*I*13	68.25	71.4	71.10	73.5	*II*06	74.65	76.4
TLC15	*I*14	68.25	71.4	71.10	73.5	*II*06	74.65	76.4
TLC16	*I*15	68.25	71.4	70.75	73.1	*II*06	74.65	76.4
TLC17	*I*16**	68.70*	71.4/71.7	71.80*	73.9	*II*07	74.65	76.3
TLC18	*I*17**	68.70*	71.4/71.7	71.80*	73.9	*II*08	74.65	76.3
TLC19	*I*18	67.70	71.2	70.25	72.8	*II*02	74.45	76.2
TLC20	*I*19	67.45	70.6	70.50	73.0	*II*02	74.45	76.2
TLN01	*I*20	67.20	70.6	70.20	72.8	*II*02	74.45	76.2
TLN02	*I*21	67.20	70.6	70.30	72.9	*II*02	74.45	76.2
TLN03	*I*22	67.20	70.6	69.70	72.3	*II*02	74.45	76.2
TLN04	*I*23	67.20	70.6	70.75	73.1	*II*02	74.45	76.2

* These *in vitro* Tm values are abnormal because they came with low amplification efficiency (see the lower-in-height melting peaks in Figure 2.8); therefore their Tm values may not be comparable with those of other amplotypes.

** I05, I16 and I17 have mutations in the forward primer region, which should have been masked as long as they were amplified successfully during *in vitro* HRM. Their *in silico* Tm1 values here are shown as when such mutations are included/masked.

*** I05 is different from I01 only in the forward primer's region. The *in vitro* Tms here varied among different runs as they are affected by PCR efficiency each time.

Table 2.14 Polymorphic nucleotide sites within the trnT-trnL*I* region aligned among the revealed haplotypes of *Arenaria* species. The shaded nucleotides show the sites where a SNP is the sole difference between two amplotypes.

Haplotype of TL*I*	Nucleotide sites (370bp aligned)																								
	26	31	37	50	54	59	79	88	100-106	130	146	158	171	177	181	196-205	210	211	218	220-241	260	261	286	323	333-341
*I*01	C	C	C	A	C	T	-	A	-------	T	A	T	-	-	T	----------	C	T	A	----------	A	G	A	G	---------
*I*02	C	C	C	A	C	T	-	A	ATTATTA	A	A	T	-	-	T	----------	A	T	A	AA	A	G	A	G	AA
*I*03	C	C	C	A	C	T	-	A	-------	T	A	T	-	-	T	----------	C	T	A	A	A	G	A	G	---------
*I*04	C	C	C	A	C	T	-	A	-------	T	A	T	-	-	T	----------	C	T	A	A	A	G	A	G	A
*I*05	A	T	C	A	C	T	-	A	-------	T	A	T	-	-	T	----------	C	T	A	----------	A	G	A	G	---------
*I*06	C	C	C	A	C	G	-	A	-------	T	A	T	-	-	T	----------	C	T	A	----------	A	G	A	G	---------
*I*07	C	C	C	A	C	T	-	A	-------	T	A	T	-	-	T	----------	C	T	A	A	A	T	A	G	A
*I*08	C	C	C	A	A	T	-	A	-------	T	A	T	-	-	T	----------	C	T	A	----------	A	G	A	G	---------
*I*09	C	C	C	A	C	T	-	A	ATTATTA	T	A	T	-	-	T	TATATA	A	T	A	A	A	G	A	G	AAA
*I*10	C	C	C	A	C	T	-	A	ATTATTA	T	A	G	-	-	T	TATATA	A	T	A	A	A	G	A	G	AAA
*I*11	C	C	C	A	C	T	-	A	ATTATTA	A	A	T	-	-	T	----------	A	T	A	AA	A	G	A	G	A
*I*12	C	C	C	C	C	T	A	A	-------	T	A	T	T	-	G	----------	C	T	A	A	T	G	A	G	AAAA
*I*13	C	C	C	C	C	T	A	A	-------	T	A	T	T	-	G	----------	C	T	A	A	T	G	A	G	AAAAAA
*I*14	C	C	C	C	C	T	A	A	-------	T	A	T	T	-	G	----------	C	T	A	A	T	G	A	G	AAAAAAA
*I*15	C	C	C	C	C	T	A	A	-------	T	A	T	T	-	G	AAAAAATATA	C	T	A	A	T	G	A	G	AAAAAA
*I*16	A	C	A	A	C	G	A	A	-------	T	G	T	-	G	G	TCTATCTTA	C	T	A	AA	T	G	C	A	AAAAAAAAA
*I*17	A	C	A	A	C	G	A	A	-------	T	G	T	-	G	G	TCTATCTTA	C	T	A	AA	T	G	C	A	AAAAAAA
*I*18	C	C	C	A	C	T	-	C	ATTATTA	T	A	T	-	-	T	TATATA	A	T	A	A	A	G	A	G	AAAAA
*I*19	C	C	C	A	C	T	-	A	ATTATTA	T	A	T	-	-	T	TATATA	A	T	G	A	A	G	A	G	AAAA
*I*20	C	C	C	A	C	T	-	A	ATTATTA	T	A	T	-	-	T	TATATA	A	T	A	A	A	G	A	G	AAA
*I*21	C	C	C	A	C	T	-	A	ATTATTA	T	A	T	-	-	T	TATATA	A	A	A	A	A	G	A	G	AAA
*I*22	C	C	C	A	C	T	-	A	ATTATTA	T	A	T	-	-	T	TATATA	A	A	A	ATATATAA AAT(+A*11)	A	G	A	G	AAA
*I*23	C	C	C	A	C	T	-	A	ATTATTA	A	A	T	-	-	T	----------	A	T	A	A	A	G	A	G	AA

Table 2.15 Polymorphic nucleotide sites within the trnT-trnL*II* region aligned among the revealed haplotypes of *Arenaria* species. The shaded nucleotides show the sites where a SNP is the sole difference between two amplotypes.

Haplotype of TL*II*	Nucleotide sites (270bp aligned)						
	25	35	157-159	167-178	181	233	245
*II*01	C	G	---	----------	G	A	T
*II*02	C	A	---	----------	T	A	C
*II*03	C	G	---	----------	G	C	T
*II*04	C	G	---	TAAAATAAGATA	G	A	T
*II*05	T	A	---	----------	T	A	C
*II*06	C	A	---	----------	G	A	C
*II*07	C	A	AA	----------	G	A	C
*II*08	C	A	AAA	----------	G	A	C

Aside from Tm values and melting shoulders, amplification efficiency served as an extra feature for amplotype identification. The amplotypes I05, I16 and I17 rendered a low yield during amplification probably due to the imperfect binding with the forward primer, and thus showed significantly lower-in-height melting peaks compared to other amplotypes (Figure 2.8). The low PCR yield itself indicated that these individuals may carry different amplotypes from others, which was then confirmed by DNA sequencing, although the relationship between Tm value and PCR yield remains complicated and unclear.

One difficulty occurred in distinguishing TLN01 and TLN02. The two haplotypes are different only in trnT-trnL*I* with a T-to-A SNP between *I*20 and *I*21. Upon *in vitro* HRM analysis they showed 0.10°C difference in their Tm2 values, which is under the proposed credible threshold in this study. As verified by DNA sequencing, the two haplotypes co-occur in two populations of *A. norvegica*, from Rum island and Inchnadamph in Scotland. It remains an unsolved issue to determine which samples from the two populations belong to which of the two haplotypes here.

The HRM analysis with trnT-trnL provided more complex results than those based on rps16. However it showed that HRM analysis is capable

of identifying most haplotypes within different DNA regions with similar amplicon sizes (270-390bp aligned), except the case of *II*06, *II*07 and *II*08 which cannot be distinguished from each other, as they differ at a single microsatellite site.

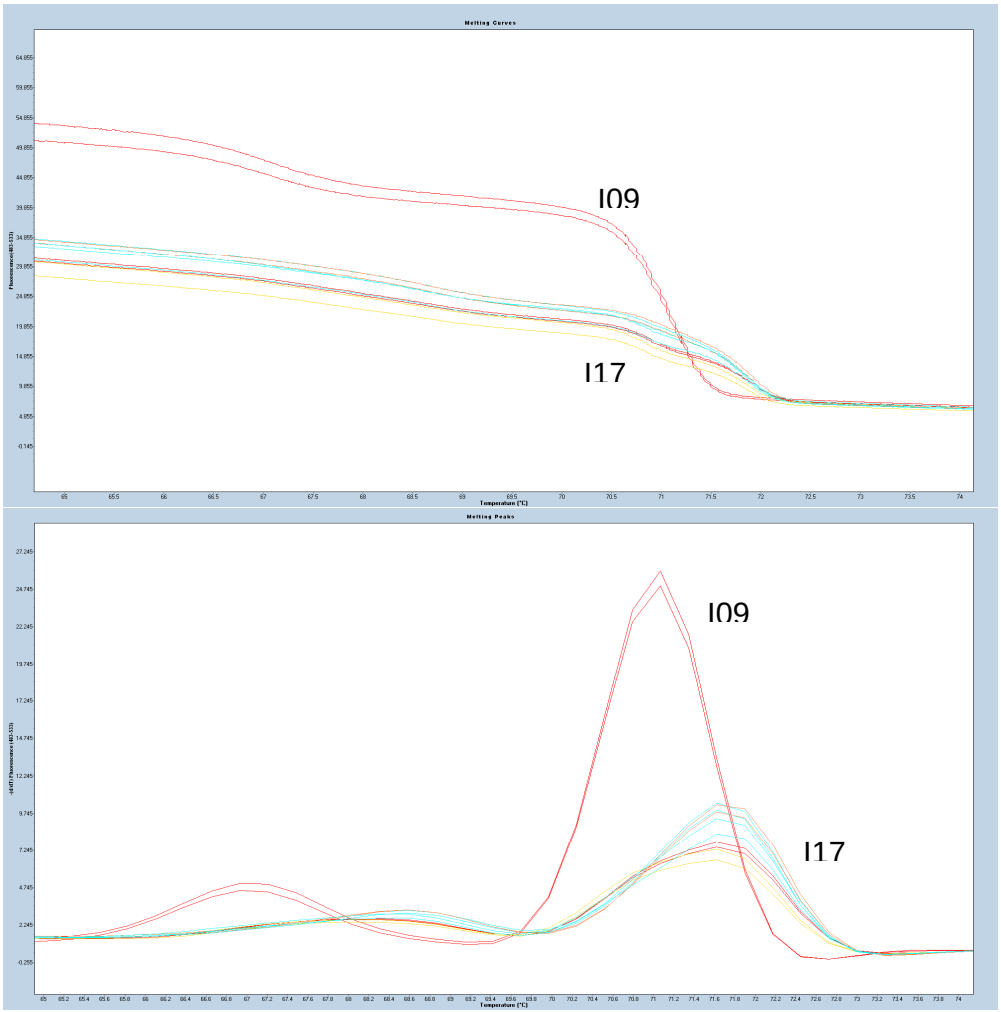


Figure 2.8 an example of low PCR yield during HRM analysis, which leads to lower aptitude melting peaks. The top panel shows the melting curves of two amplotypes of trnT-trnL*I*, *I*09 and *I*17. The PCR yield is shown by the height at the beginning of the melting curve. It is clearly seen that *I*17 has a lower PCR yield than *I*09. The lower panel shows the melting peaks of the two amplotypes. While *I*09 shows melting peaks comparable to those seen in Figure 2.7, *I*17 has lower aptitude melting peaks, the Tm values of which are considered unsuitable to be compared with the Tm values of other amplotypes as they may have been impacted by low PCR yield. However, the low PCR yield itself is an indicator of possible sequence variation.

2.4.1.5 Concatenated haplotypes

Based on the HRM analysis with rps16 and trnT-trnL, all the 484 individual samples from 23 populations of *A. ciliata* and *A. norvegica* were assigned to 34 composite haplotypes by concatenating the rps16 and trnT-trnL regions. The concatenated haplotypes and their distribution are shown in Table 2.16.

Table 2.16 The concatenated chloroplast DNA haplotypes from *A. ciliata* and *A. norvegica* and their distribution among the sampled populations.

rps16	trnT-trnL	Concatenated	Occurrence
rpsC01	TLC01	RTC01	Ir1-4, Pi2
rpsC02	TLC02	RTC14	Ir1, 3, 4, Sv2
rpsC03	TLC02	RTC15	Ir2
rpsC02	TLC12	RTC16	Au1
rpsC04	TLC03	RTC02	Ir3, Sw2, Sv1
rpsC05	TLC01	RTC03	Ir3
rpsC06	TLC01	RTC04	Py1
rpsC06	TLC04	RTC05	Py1, 2
rpsc06	TLC08	RTC06	Sw1, 2
rpsC14	TLC01	RTC07	Pi1, Sw1
rpsC14	TLC06	RTC08	Pi1
rpsC14	TLC04	RTC09	Py2
rpsC14	TLC07	RTC10	Py2
rpsC14	TLC09	RTC11	Au2
rpsC07	TLC05	RTC12	Py1
rpsC08	TLC10	RTC17	Pi2, Py1
rpsC09	TLC13	RTC22	It1, Au2
rpsC09	TLC14	RTC23	It2
rpsC09	TLC15	RTC24	It3, Sw2
rpsC13	TLC15	RTC25	Fr1
rpsC16	TLC15	RTC18	Pi2
rpsC10	TLC16	RTC26	It3
rpsC11	TLC17	RTC27	Fr1
rpsC12	TLC19	RTC19	Py1, 2
rpsC12	TLC20	RTC20	Sw2
rpsC15	TLC11	RTC21	Pi1-3
rpsC17	TLC08	RTC13	Au1
rpsC18	TLC18	RTC28	Au2
rpsN01	TLN01	RTN01	NB, NR, NIn
rpsN01	TLN02	RTN02	NE, NR, NIn, NS, NIc
rpsN01	TLN03	RTN03	NS
rpsN02	TLN04	RTN04	NR

2.4.2 Results for *Minuartia recurva*

2.4.2.1 Haplotype detection with rps16

Within the 250 individual samples of *M. recurva* from 13 localities, four haplotypes of rps16 were revealed by HRM analysis and posterior sequencing confirmation. The nucleotide differences among the haplotypes are listed in Table 2.17.

Table 2.17 Polymorphic nucleotide sites within the rps16*I* region aligned among the revealed haplotypes of *Minuartia recurva*. The shaded nucleotides show the sites where a SNP is the sole difference between two amplotypes.

Haplotype of rps16	Nucleotide sites (696bp for rps16)		
	rps16*I* (1-354bp)	rps16*II* (355-696bp)	
	300-316	468	497
rpsM01	TATGATTAGATTCTTTG	G	T
rpsM02	TATGATTAGATTATTTG	G	T
rpsM03	TATGATTAGATTATTTG	A	T
rpsM04	----------	G	C

It is seen that a single C/A substitution within rps16*I* marks the sole difference between rpsM01 and rpsM02/03, while a long indel of 17bp, also within rps16*I*, shows difference between rpsM04 and the remaining haplotypes.

Three melting peaks were revealed for rps16*I*, while only the peak between 74.0 and 75.5°C was indicative of the variation between the detected amplotypes. The approximate *in vitro* Tm values are recorded and listed with corresponding *in silico* Tms in Table 2.18, and the profiles of melting peaks for three *M. recurva* samples are shown in Figure 2.9 to illustrate the observed difference between the rps16*I* amplotypes.

Table 2.18 Summary of all rps16 haplotypes identified in *M. recurva* using HRM *in vitro* and *in silico* analysis of rps16*I* and rps16*II* amplicons

rps16 haplotype	rps16*I* amplotype	*In vitro* Tm1 (°C)	*In silico* Tm1 (°C)	*In vitro* Tm2 (°C)	*In silico* Tm2 (°C)	*In vitro* Tm3 (°C)	*In silico* Tm3 (°C)	rps16*II* amplotype	*In silico* Tm4 (°C)
rpsM01	*I*01	75.1	77.3	77.0	80.0	81.8	82.0	*II*01	77.9
rpsM02	*I*02	74.2	76.4	77.0	80.0	81.8	82.0	*II*01	77.9
rpsM03	*I*02	74.2	76.4	77.0	80.0	81.8	82.0	*II*02	77.7
rpsM04	*I*03	absent	77.5	77.0	80.0	81.8	82.0	*II*03	78.1

Note: *In vitro* HRM analysis was not done with rps16*II*, so the *in vitro* Tm4 values are not available. However it is seen from the *in silico* Tms that the three amplotypes of rps16*II* can be distinguished by HRM analysis.

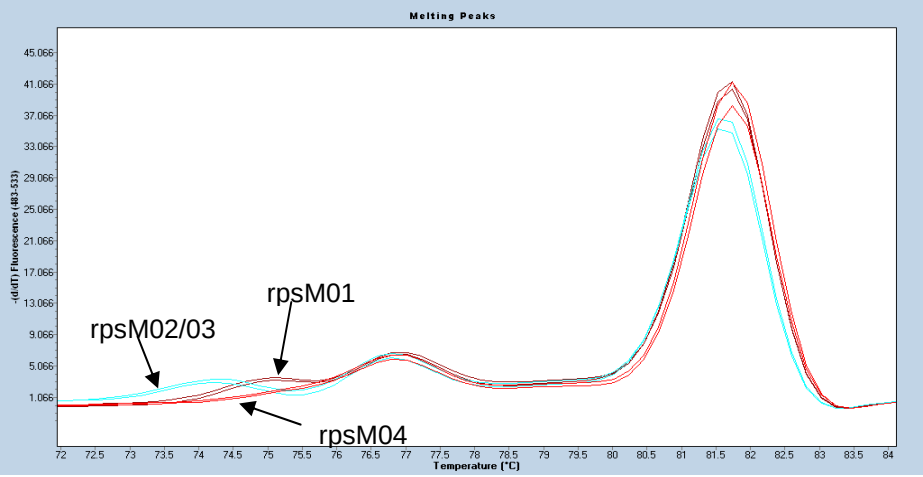


Figure 2.9 Melting peaks of three amplotypes of rps16*I* found in *Minuartia recurva*. The names of the corresponding rps16 haplotypes are used instead of amplotype names. The haplotypes rpsM02 and 03 share the same amplotype of rps16*I*.

For the amplicons of rps16*II*, it is seen from Table 2.17 that a G/A SNP at site 468 differentiated between haplotypes rpsM02 and rpsM03. These two haplotypes had already been distinguished by posterior DNA sequencing of putative haplotypes during HRM analysis of rps16*I*, because they did not co-occur in the same population (see Table 2.21). Also a SNP at site 497 within rps16*II* differentiated between rpsM04 and the other haplotypes. Overall, the amplotypes of rps16*II* did not identify any new haplotypes compared to the analysis of rps16*I*.

2.4.2.2 Haplotype detection with trnT-trnL

Within the studied individual samples of *M. recurva*, eight haplotypes of trnT-trnL were revealed by HRM analysis and posterior sequencing confirmation. The nucleotide differences among the haplotypes are shown in Table 2.19 (next page).

Table 2.19 Polymorphic nucleotide sites within the trnT-trnL region aligned among the revealed haplotypes of *Minuartia recurva*.

Haplotype of trnT-trnL	Nucleotide sites (534bp for trnT-trnL and 1-310bp for trnT-trnL*I*)									
	32-37	38	51	96	123-125	126-127	129	130	131	137
TLM01	ACAATT	A	T	-	A--	TT	T	T	C	A
TLM02	ACAATT	C	A	-	AA-	TT	T	T	C	A
TLM03	------	C	A	-	A--	-T	T	T	C	A
TLM04	ACAATT	T	A	-	A--	-T	G	A	A	C
TLM05	ACAATT	C	A	T	A--	-T	G	A	A	C
TLM06	ACAATT	A	T	-	---	TT	T	T	C	A
TLM07	ACAATT	C	A	-	AA-	-T	T	T	C	A
TLM08	ACAATT	C	A	-	AAA	-T	T	T	C	A

All the polymorphic nucleotide sites of trnT-trnL fell within the amplicon trnT-trnL*I*, which generated two melting peaks; whereas trnT-trnL*II*, which generated a single melting peak, did not display any variation among the studied samples.

Individuals representative of each of the four rps16 haplotypes were also sequenced at the trnT-trnL locus to determine if differences were shared between the two loci. The result showed that samples in population MR6 fell into two haplotypes of trnT-trnL, which are recorded as TLM05 and TLM08. HRM analysis of trnT-trnL*I* distinguished between these two haplotypes and identified each individual within the population MR6 to the appropriate amplotype. The melting peaks of representative trnT-trnL*I* amplotypes are shown in Figure 2.10-12. The approximate *in vitro* Tm

values are listed in Table 2.20, where *in silico* Tm values are also included.

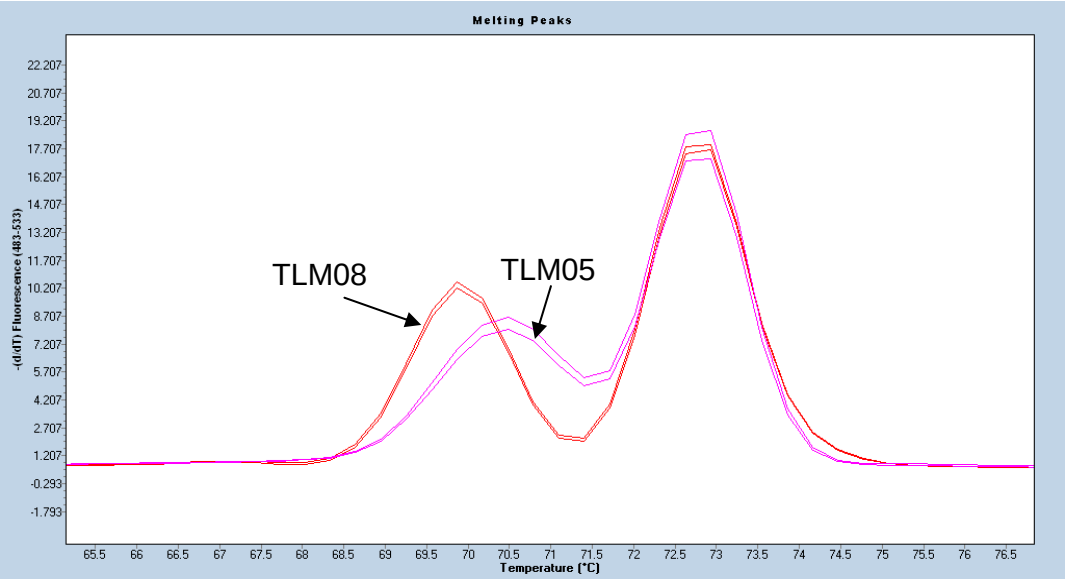


Figure 2.10 Melting peaks with trnT-trnL*I* of the amplotypes corresponding to haplotypes TLM05 and TLM08, which were found to co-occur in the population MR6. The temperatures of the first peak show difference between the two amplotypes.

Haplotypes TLM03 and TLM04 were also identified co-occurring in population MR12, in accordance with haplotypes rpsM03 and rpsM04 respectively based on the result of rps16. HRM analysis distinguished them based on the melting peaks with trnT-trnL*I* as shown in Figure 2.11.

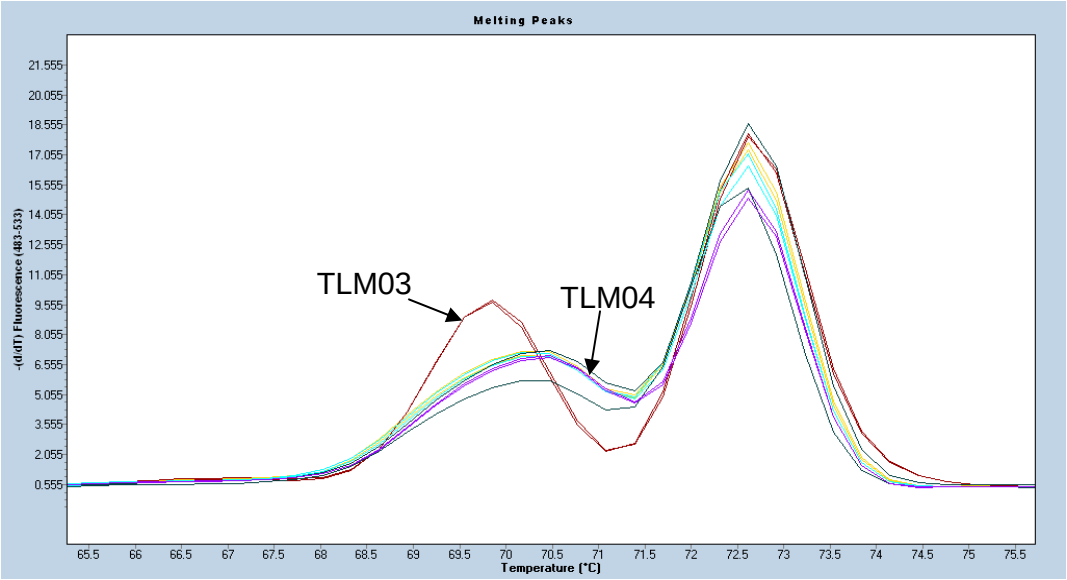


Figure 2.11 Melting peaks with trnT-trnL*I* of the amplotypes corresponding to haplotypes TLM03 and TLM04, which were found to co-occur in the population MR12. The temperatures of the first peak show difference between the two amplotypes.

HRM analysis with trnT-trnL*I* also distinguished TLM01 from TLM07, where the temperatures of the first melting peak are different between the corresponding amplotypes (Figure 2.12).

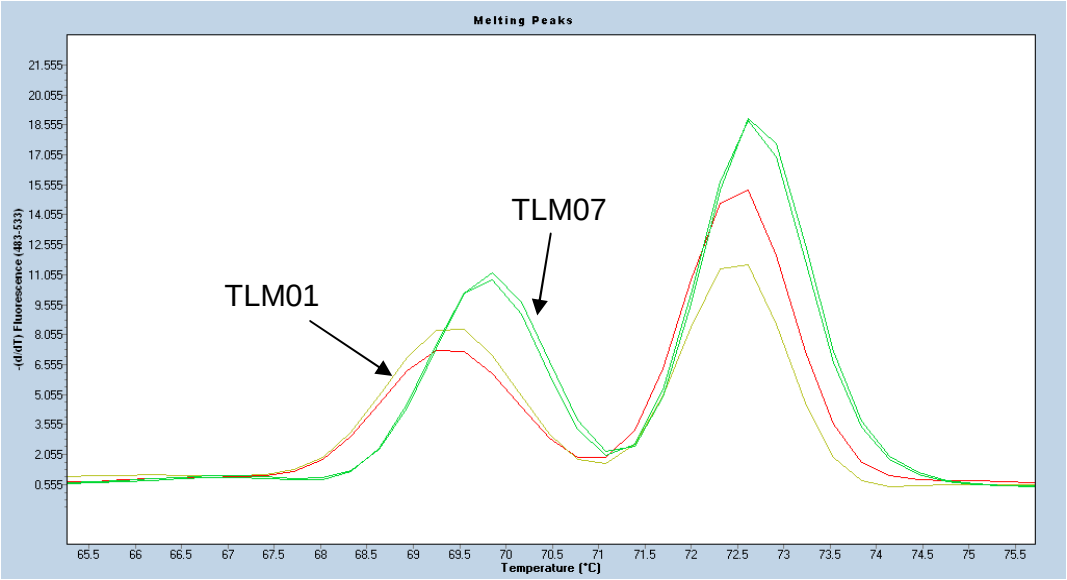


Figure 2.12 Melting peaks with trnT-trnL*I* of the amplotypes corresponding to haplotypes TLM01 and TLM07. The temperatures of the first peak show difference between the two amplotypes.

Table 2.20 Summary of all rps16 haplotypes identified in *M. recurva* using HRM *in vitro* and *in silico* analysis of rps16*I* and rps16*II* amplicons

trnT-trnL haplotype	trnT-trnL *I* amplotype	*In vitro* Tm1 (°C)	*In silico* Tm1 (°C)	*In vitro* Tm2 (°C)	*In silico* Tm2 (°C)	trnT-trnL *II* amplotype	*In silico* Tm3 (°C)
TLM01	*I*01	69.5	71.7	72.7	74.9	*II*01	76.6
TLM02	*I*02	69.9	absent	72.7	74.9	*II*01	76.6
TLM03	*I*03	69.9	absent	72.7	74.9	*II*01	76.6
TLM04	*I*04	70.6	71.9	72.7	74.9	*II*01	76.6
TLM05	*I*05	70.6	72.6	72.7	74.9	*II*01	76.6
TLM06	*I*06	69.5	71.7	72.7	74.9	*II*01	76.6
TLM07	*I*07	69.9	absent	72.7	74.9	*II*01	76.6
TLM08	*I*08	69.9	absent	72.7	74.9	*II*01	76.6

Note: *In vitro* HRM analysis was not done with trnT-trnL*II*, so the *in vitro* Tm3 values are not available.

As seen from Table 2.18, TLM01 and TLM06 vary only in the single A repeats at 123-125bp, while TLM02, 07 and 08 vary only in the single A repeats at 123-125bp and/or single T repeats at 126-127bp.

Polymorphisms in the two microsatellite sites were not amenable to HRM analysis based on the present result, and these haplotypes were only revealed by following the protocol of sequencing two or more samples of each putative amplotype from each population.

It is seen from Figure 2.10-11 that while TLM04 and TLM05 are different by a single T/C mutation and a single T indel, they have shown similar profile of melting peaks. Also it is seen from Figure 2.10-12 that while TLM03 is different from TLM02 (and TLM07, 08) by a 6-bp deletion near the binding region of the forward primer, but shares a similar melting peak profile with the latter. However their seemingly similar melting peak profiles are from different batches of runs and thus may contain inter-batch errors. They are predicted to be distinguished by HRM analysis by the *in silico* analysis (Table 2.20). Fortunately, the seemingly undistinguishable haplotype pairs, except for those varying in SSRs, were not found to co-occur and thus were not missed by HRM analysis plus posterior validation by DNA sequencing.

As to the repeatability of HRM analysis in the case of *M. recurva*, it was found that in two populations, MR2 and MR3, carriers of the same amplotype may render different melting curves. The melting peaks

produced by MR2.30 and MR3.10 are different from those produced by the other samples in the two populations, as shown in Figure 2.13, however posterior DNA sequencing confirmed that MR2.30 shares the same haplotype (TLM02) with part of other samples in MR2 and MR3.10 shares the same haplotype (TLM08) with other samples in MR3.

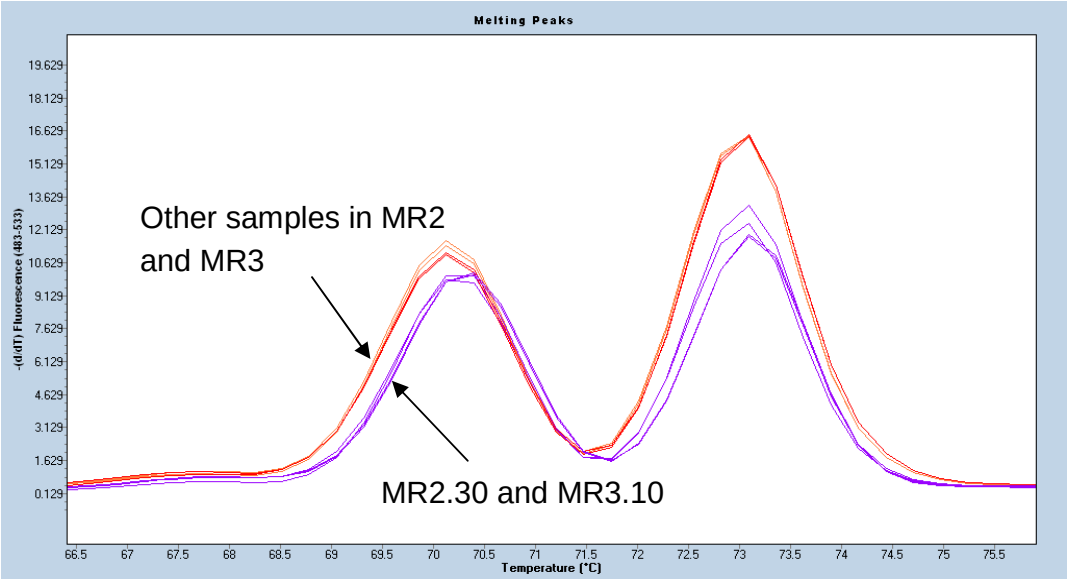


Figure 2.13 Representative melting peaks with trnT-trnL*/* of the samples from MR2 and MR3. MR2.30 and MR3.10 shared the peak curves in purple while all other samples shared the ones in red. The first peak shows around 0.2°C difference between the two groups, however, this Tm difference was not due to sequence variation as revealed by DNA sequencing validation.

2.4.2.3 Concatenated haplotypes

Based on the designation of individual samples into different haplotypes, nine composite haplotypes were obtained by concatenating the rps16 and trnT-trnL sequences. The correspondence relationship between the

haplotypes and their distribution are listed in Table 2.21. It is seen that RTM02 and RTM09 are the only pair that cannot be distinguished by HRM and meanwhile coexist in the same population (MR2).

Table 2.21 The concatenated chloroplast DNA haplotypes found in *M. recurva* and their correspondence to the rps16 and trnT-trnL haplotypes.

rps16	trnT-trnL	Concatenated	Occurrence
rpsM01	TLM01	RTM01	MR1, MR8, MR10
rpsM01	TLM06	RTM07	MR7
rpsM02	TLM02	RTM02	MR2
rpsM02	TLM07	RTM08	MR4, MR5, MR13, MR15
rpsM02	TLM08	RTM09	MR2, MR3, MR11, MR14
rpsM03	TLM03	RTM03	MR12
rpsM04	TLM04	RTM04	MR12
rpsM04	TLM08	RTM05	MR6
rpsM04	TLM05	RTM06	MR6

2.5 Discussion

2.5.1 Discussion based on the HRM analysis with the *Arenaria* species

The protocols of HRM analysis as tested in our study has generated highly detailed haplotype identity and frequency data for *A. ciliata* and *A. norvegica*. Initially it was concerned that the chosen amplicons (350-400 bp) might be too long compared to established HRM norms; however the accuracy of the analysis was sustained over these amplicon size intervals, in particular due to the presence of multiple melting domains in the double-stranded amplified DNA.

With the two targeted amplicons rps16*I* and *II*, HRM analysis revealed all except one of the 20 haplotypes that were finally confirmed by DNA sequencing, allowing discrete haplotype identification in 189 of 190 possible pairwise haplotype comparisons, including 31 of 32 observed co-occurrences of haplotypes within single populations (Figure 2.6). The haplotypes rpsC06 and rpsC14 are the only co-occurring haplotypes that have no difference in their amplicons' melting profiles (Tm values). Low Tm variation (ΔTm <0.2°C) was found between rpsC01 and rpsC05 co-occurring within the population Ir3, between rpsC06 and rpsC07 within Py1, and between rpsC08 and rpsC12 within Py1. However these co-occurring haplotypes were successfully identified through our protocol to verify all suspected variation by posterior sequencing, thus missing detection was avoided in such cases. Also because the haplotype discrimination was made within each batch of HRM analysis, where usually a single population was assayed each time, those haplotypes with similar Tm values but distributed in different populations could also be well distinguished after confirmation via DNA sequencing.

Based on the same protocols, HRM analysis with trnT-trnL*I* and *II* also revealed all of the 24 haplotypes that were finally confirmed by DNA sequencing, except the difficulty in distinguishing between TLN01 and

TLN02. Cases where different haplotypes shared similar or the same Tm values were also found. The haplotypes TLC02, TLC11, TLC12 and TLN04 were predicted by *in silico* simulation to share the exactly the same Tm values, and did show similar Tm values during the *in vitro* HRM analysis (Table 2.13), which made it potentially difficult to distinguish between them. However the four haplotypes did not co-occur in any of the populations, and the fact was validated by DNA sequencing of at least two individuals of each suspected haplotype in each population. The same situation was found between TLC03 and TLC04, among TLC13, TLC14 and TLC15, and between TLC17 and TLC18.

Sensitivity of the analysis may have been impacted by the quality of the DNA template. This work utilized a DNA sample set that had been extracted using a modified CTAB method, which doesn't guarantee a uniform chemical composition for the extracted DNA. The potentially lower quality of the DNA template may thus have had an impact on subsequent melting analysis, generating system errors between Tm readings of different batches of HRM assays. It is suspected that inter-batch and inter-population errors in Tm values among individuals sharing the same haplotype are caused by this effect. As a result, the *in vitro* Tm values in Table 2.10 and 2.13, which were adjusted mean values

for each haplotype via the use of reference samples between different runs, may still contain part of inter-batch and inter-population errors. However, the *in vitro* ΔTm values between each pair of amplotypes are much as predicted by the *in silico* simulation, which was validated by Mantel correlation test between the *in vitro* and *in silico* ΔTm values based on the rps16 data (Table 2.22).

Table 2.22 Results of Mantel correlation tests between observed *in vitro* inter-haplotype ΔTm, modelled *in silico* inter-haplotype ΔTm, and calculated inter-haplotype genetic distance (Tajima-Nei's *D*). This table is cited from Dang *et al.* 2012.

Inter-amplotype ΔTm *in vitro* vs *in silico*	ΔTm1 *in vitro* v ΔTm1 *in silico*		*0.6247*	*0.0001*
	ΔTm2 *in vitro* v ΔTm2 *in silico*		0.9257	0.0001
	ΔTm3 *in vitro* v ΔTm3 *in silico*		*0.7964*	*0.0001*
Inter-haplotype ΔTm vs Tajima-Nei Genetic Distance D between haplotypes	|ΔTm1+ΔTm2| vs. D_{rps16I}	*in vitro*	0.2168	0.0858
		in silico	*0.4468*	*0.0021*
	ΔTm3 vs. $D_{rps16II}$	*in vitro*	0.2118	0.0886
		in silico	*0.5088*	*0.0015*
	|ΔTm1+ΔTm2+ΔTm3| vs. D_{rps16}	*in vitro*	0.3593	0.0122
	|Δ*Tm1*+Δ*Tm2*+Δ*Tm3*+Δ*TmX*| *vs.* D_{rps16}	*in silico*	*0.4189*	*0.0012*

Also it is clear that where two amplotypes can be distinguished by *in silico* simulation (by 0.1°C) using uMelt^{SM}, in most cases they can also be distinguished by the *in vitro* HRM analysis on the LightCycler 480 system. One exception occurred between rps16*I*04 and *I*11, where the *in silico* simulation predicts that they should show different Tm1 and Tm2 values (ΔTm1=0.1°C, ΔTm2=0.1°C) but the *in vitro* HRM analysis

provided the same Tm1 and Tm2 values from the two amplotypes (Table 2.10). The other exception was presented among trnT-trnL*II*04, *II*06 and *II*07/08, where the *in silico* simulation predicts that they should show different Tm3 values while the *in vitro* HRM analysis provided the same Tm3 values for them (Table 2.13). Based on the consideration that the *in vitro* Tm values may contain errors, the *in silico* simulation did provide a reasonable prediction of what results the *in vitro* HRM analysis may show to us. Occasionally *in vitro* analysis produces even greater ΔTm values than the *in silico* simulation predicts, e.g. rps16*I* amplotypes *I*09 and *I*10 shared the same *in silico* Tm1=77.5°C but have different *in vitro* Tm1 values (0.35°C difference, see Table 2.10). While estimates of precise °C values of Tm peaks differs between the two methods as we have applied them, the underlying pattern of inter-haplotype ΔTm identification is equivalent between the two (mantel test of correlation between pairwise ΔTm matrices is significant at $p<.001$ Table 2.17). Overall this data affirms both that *in silico* simulation is a valid support for *in vitro* HRM work, and that the model parameters applied in uMelt^{SM} generate slightly conservative estimates of HRM curve difference between amplotypes.

As mentioned above, the limited resolution of HRM means that it is not

an equivalent to template sequencing, however a more general correlation between inter-haplotype ΔTm values and inter-haplotype genetic distance is a possibility. Mantel tests on the *in vitro* and *in silico* data here (Table 2.17) showed a significant positive correlation between the compounded differences in ΔTm1, 2 and 3 between haplotypes and the corresponding pairwise Tajima-Nei genetic distance *D* between haplotypes. Thus while certain haplotype pairs in the study with large pairwise *D* values did have very similar melting-curve profiles, these correlation tests on our data support the case that in general a greater ΔTm value between haplotypes indicates a greater level of evolutionary divergence.

Compared to the alternative approach of sequencing every single individual sample to generate a total count of haplotype frequency data across the population samples, the resources required for the HRM method were significantly reduced. Although HRM analysis was not exhaustive, we are confident that the method effectively identified haplotypes in the sampled populations. With the HRM analysis of rps16, in total over 80 sequence identities were obtained, and in only one case was an amplotype sequence returned that was not discernable by HRM analysis. This contrasts with the results of the *in silico* RFLP analysis (Dang *et al.* 2012), where only 35% of the recorded haplotypes could be

detected uniquely.

2.5.2 Discussion based on the HRM analysis with *M. recurva*

Four haplotypes of rps16 were found within the sampled populations of *M. recurva*, which are composed of three amplotypes of rps16*I* and three amplotypes of rps16*II*. HRM analysis performed with rps16*I* successfully revealed all the three amplotypes. Posterior DNA sequencing revealed the difference between rpsM02 and rpsM03 in their sequences of rps16*II*. The two haplotypes were not found to co-occur and rpsM03 is found to be carried by only one sample from the population M12. For each possible haplotype identified by HRM analysis in each population, two samples were subjected to DNA sequencing and no more variation was revealed. Thus it is considered HRM analysis with rps16*I* solely was sufficient to reveal all existent rps16 haplotypes from the sampled populations.

Eight haplotypes of trnT-trnL were found from the studied populations, which were different by mutations within trnT-trnL*I* only. HRM analysis with trnT-trnL*I* was not able to distinguish the haplotypes varying only in the poly-A and/or the poly-T sites. The low sensitivity of HRM analysis

to polymorphisms in SSRs was also seen for trnT-trnL in the case of the Arenaria species in this study (see section 2.4.1.4). Besides this, the in vitro HRM analysis seemed unable to reveal the difference between TLM02 and TLM03 although they were not analyzed together in a single batch of run. The two haplotypes vary by a 6-bp indel near the binding zone of the forward primer, and were predicted to be distinguishable by HRM analysis based on the in silico simulation. However it has been reported by Reed and Wittwer (2004) that mutations near the primer-binding sites may be more difficult to be detected than those near the centre of the amplicon, and the present result has provided an example of just this situation. In the present case HRM analysis with trnT-trnL may also have failed to distinguish between TLM04 and TLM05, which differ by a single T/C mutation and a T indel, which may also be explained by the adjacency of the mutations to the primer binding region. However, as with TLM02 and TLM03, TLM04 and TLM05 were not found to co-occur in the same populations and thus were successfully revealed by posterior sequencing. Thus when variation in SSRs is not considered, the described protocol was able to designate all the samples into the revealed haplotypes as shown in Table 2.21 (section 2.4.2.3).

While missed detection was assumed to happen occasionally, false

positive detection can also be a problem as described in 2.4.2.2, where MR2.30 and MR3.10 rendered unexpected different melting curves compared to other samples from the same populations sharing the same haplotypes. This situation was also seen with the Arenaria species where different carriers of the same amplotype may show different melting curves, which has been discussed in section 2.5.1. Human errors have been excluded in this case by repeating the assay with the same samples. The unexpected melting curve variation may be caused by chemical impurity in the DNA template solution and inter-template variation in their chemical content. With the species M. recurva, these are the only cases of false positive signals, as rare as the cases found in the Arenaria species. It is likely that such false positive results would be reduced in future work where DNA templates are prepared with higher purity.

2.6 Conclusion

Based on the results detailed in this chapter, HRM analysis is suitable for a wide range of phylogeographic studies where polymorphism is evident in commonly used 400+ bp spacer regions in chloroplast and mitochondrial DNA (e.g. Shaw et al. 2005). The limitations of the method are similar to direct sequencing in that individual loci may not

contain informative polymorphisms below the species level. However, in any studied organism where discrete organelle DNA differentiation is evident between population sub-groups, or where cryptic speciation has occurred, HRM has the capacity to greatly increase the scope and sensitivity of haplotype analysis.

Bearing in mind the sensitivity limitations (regarding SNPs and microsatellites) and the need for error-minimization as shown above, there is considerable scope for improvement in this technique, for example in the optimization of amplicon design and selection protocols, and evaluation of large loci and nuclear loci such as ITS (potentially heterozygous). Besides, for any organism where comparative sequence data has already been generated, *in silico* simulation can be carried out to evaluate the utility of HRM analysis and to optimize the HRM protocols, e.g. by selecting the amplicons that generate multiple melting domains. With wider availability of real-time PCR equipment, and the superior sensitivity of HRM to mutational differences compared to PCR-RFLP, the method provides an untapped resource in phylogeographic studies.

Chapter 3

Phylogeography, population genetics and demographic history analysis of *Arenaria ciliata* and *A. norvegica* in Europe

3.1 Introduction

Based on the chloroplast DNA data obtained for the species *Arenaria ciliata and A. norvegica*, as described in Chapter 2, a phylogeographic study has been conducted on the populations of the two carnation species.

As described by Wyse Jackson and Parnell (1987), *A. ciliata* and *A. norvegica* are two closely related species falling within the *A. ciliata* L. Complex. There are several subspecies of *A. ciliata*, varying in their morphology, chromosome number, ploidy level and distribution across Europe, however there are ambiguities in distinctions among them. The subspecies of *A. ciliata* L. include *A. ciliata* subsp. *ciliata*, *A. ciliata* subsp. *hibernica* (included in subsp. *ciliata* in Flora Europaea by Tutin *et al*. 1993), *A. ciliata* subsp. *pseudofrigida*, *A. ciliata* subsp. *moehringioides* (referred to as *A. multicaulis* in Atlas Florae Europaeae by Jalas & Suominen 1983) and *A. ciliata* subsp. *bernensis*. There is another related species included the species Complex, *A. gothica* Fries, which is considered to be closely related to *A. ciliata* subsp. *moehringioides*. Also two subspecies are recorded in *A. norvegica*, including *A. norvegica* subsp. *norvegica* and *A. norvegica* subsp. *anglica*. The two species (*senso lato*) both occur in open habitats on basic soils in

high latitude or mountainous areas, usually on clefts in limestone, but with different distributions. The species *A. norvegica* is generally a quasi-arctic plant found in Iceland, Ireland, England, Scotland, the Faroe Islands, Norway and Sweden, while the species *A. ciliata* is an arctic-alpine species widely distributed in Pyrenees, Alps and other mountainous area across Europe at a high altitude (1700-2500m), with isolated populations in northwest Ireland (460-540m), Scandinavia and Svalbard. To date there has been no phylogenetic investigation of relations within this broad species complex, and while morphemetric analysis shows some taxon-correlated variation in multivariate analysis, no discrete morphological characters have been identified that unambiguously distinguish between the various subspecies (Wyse Jackson and Parnell, 1987).

The chloroplast DNA data obtained for the species *A. ciliata and A. norvegica*, as described in Chapter 2, provide the basis for a phylogeographic study on the populations of the two species across this European distribution. As samples have been collected from most of the above localities (see Table 3.1, which contains the same populations as Table 2.1, but shows them according to different localities) and the chloroplast haplotype identities known for each collected sample (except

for the designation some of *A. norvegica* individuals from Inchnadamph and Rum Island, Table 3.2), it is possible to clarify the relationship among the relevant subspecies and to interpret the migratory history of the populations.

In the case of the ambiguous *A. norvegica* samples from Scotland, based on the trnT-trnL DNA data as described in Chapter 2, both haplotypes RTN01 and RTN02 of *A. norvegica* are found at Inchnadamph and on the island of Rum. The haplotypes differ by only a single A/T point mutation, however this mutational change cannot be detected by HRM analysis, and as we have not carried out exhaustive sequencing, it is not possible to know the exact frequencies of the two haplotypes within the two populations, other than that both haplotypes are recorded from these sites.

In this chapter it is assumed that the two haplotypes co-occur with equal frequency in each of the above populations. The possibility that one haplotype dominates the other in each population has been considered and checked throughout the presented analyses, to identify situations where these different frequencies may impact the overall conclusions.

Table 3.1 Localities of the sampled populations of *Arenaria ciliata* (Ac) and *A. norvegica* (An) in the study.

Species and Localities	Population code	Latitude/ Longitude	Sample size
Ac, Ben Bulben, Co. Sligo, Ireland	Ir1	N54º 20.672' W08º 27.373'	30
	Ir2	N54º 21.353' W08º 27.290'	30
	Ir3	N54º 21.549' W08º 24.044'	30
	Ir4	N54º 21.879' W08º 25.705'	30
Ac, Niedere Tauren, Steiermark, Austria	Au1	N47º 16.270' E14º 21.210'	27
Ac, Karawanken, Karten, Austria	Au2	N46º 30.200' E14º 29.120'	25
Ac, Piemonte Italy	It1	N44º 09.864' E07º 47.201'	26
	It2	N44º 27.227' E06º 55.313'	1
	It3	N44º 40.709' E06º 59.484'	4
Ac, Col D'Agnel, Provence- Alpes, France	Fr1	N44º 46.880' E06º 40.638'	5
Ac, Picos de Europa, Cantabria, Spain	Pi1	N43º 09.533' W04º 49.302'	19
	Pi2	N43º 10.644' W04º 49.967'	15
	Pi3	N43º 09.374' W04º 48.213'	19
Ac, Valle de Benasque, Aragón, Pyrenees, Spain	Py1	N42º 40.957' E00º 36.177'	20
	Py2	N42º 40.957' E00º 36.177'	16
Ac, Gemmipass, Leukerbad, Switzerland	Sw1	N46º 23.883' E07º 34.596'	8
	Sw2	N46º 25.165' E07º 37.478'	29
Ac subsp. *pseudofrigida*, Midtre Lovénbreen, Svalbard	Sv1	N78º 54.48' E12º 04.70'	2
Ac subsp. *pseudofrigida*, Bohemanflya, Svalbard	Sv2	N78º 23.42' E14º 44.23'	5
An, Black Head, Co. Clare, Ireland	NB	N53º 08.243' W09º 16.048'	30
An, Yorkshire, England	NE	N54º 17.000' E02º 33.010'	19
An, Eldgja gorge, Herobreio, Iceland	NIc	N64º 24.680' W18º 42.252'	2
An, Inchnadamph, Highlands, Scotland	NIn	N58º 07.493' W04º 55.374'	30
An, Rum, Western Isles, Scotland	NR	N56º 59.647' W06º 18.863'	29
An, Shetland Islands, Scotland	NS	N60º 30.835' W01º 21.674'	29

Table 3.2 Distribution of the identified haplotypes among the studied populations.

Individuals	rps16	trnT-trnL	composite
Ir1.1, 1.2, 1.4, 1.10, 1.14, 1.15, 1.17, 1.24, Ir2.2-13, 2.15-30, Ir3.12, 13, 15, 16, 17, 20-27 Ir4.9 Pi2.7, 12, 15	rpsC01	TLC01	RTC01
Ir1.3, 5-9, 11-13, 16, 18-23, 25-30, Ir3.1-6, 9-11, 14 IR4.1-8, 10-30 Sv2.1-5	rpsC02	TLC02	RTC14
Ir2.1, 14	rpsC03	TLC02	RTC15
Au1.2, 4-11, 13, 23, 24, 27	rpsC02	TLC12	RTC16
Ir3.7, 8, 18, 19, 28, 29 Sw2.5, 8, 9, 13, 16, 21, 23, 29, 30 Sv1.1, 1.2	rpsC04	TLC03	RTC02
Ir3.30	rpsC05	TLC01	RTC03
Py1.3, 11, 13, 20	rpsC06	TLC01	RTC04
Py1.1, 6, 18 Py2.7	rpsC06	TLC04	RTC05
Sw1.1-7, Sw2.1-4, 6, 7, 10, 11, 14, 17-18, 20, 22, 24-27	rpsc06	TLC08	RTC06
Pi1.1-4, 6, 8-14, 16-18 Sw1.8, Sw2.15	rpsC14	TLC01	RTC07
Pi1.19, 20	rpsC14	TLC06	RTC08
Py2.1, 4, 5, 8, 10, 12, 13	rpsC14	TLC04	RTC09
Py2.2, ,3, 6, 9, 11, 15,16	rpsC14	TLC07	RTC10
Au2.1, 4, 9, 25	rpsC14	TLC09	RTC11
Py1.4, 7	rpsC07	TLC05	RTC12
Py1.2, 5, 8, 9, 12, 14, 16, 17 Pi2.3, 6	rpsC08	TLC10	RTC17
It1.1-26 Au2.3, 6, 10, 12-14, 16, 18	rpsC09	TLC13	RTC22
It2.1	rpsC09	TLC14	RTC23
It3.1, 2, 4 Sw2.28	rpsC09	TLC15	RTC24
Fr1.3, 4, 5	rpsC13	TLC15	RTC25
Pi2.8	rpsC16	TLC15	RTC18
It3.3	rpsC10	TLC16	RTC26
Fr1.1, 2	rpsC11	TLC17	RTC27
Py1.10, 15, 19 Py2.14	rpsC12	TLC19	RTC19
Sw2.12	rpsC12	TLC20	RTC20
Pi1.7, 15 Pi2.1, 2, 4, 5, 9, 10, 11, 13, 14, 16 Pi3.1-18, 20	rpsC15	TLC11	RTC21
Au1.1, 3, 12, 14-22, 25, 26	rpsC17	TLC08	RTC13
Au2.2, 5, 7, 8, 11, 15, 17, 19-23, 24	rpsC18	TLC18	RTC28
NB1-30, NR21, NIn1 (and others in NR and Nin*)	rpsN01	TLN01	RTN01
NE1-19, NR4, NIn7 (and others in NR and Nin*), NS1-29 (except NS21, 27) NIc1, 2	rpsN01	TLN02	RTN02
NS21, 27	rpsN01	TLN03	RTN03
NR1-3, 5-17, 28, 29	rpsN02	TLN04	RTN04

* The haplotype identities of part of NIn and NR samples are undetermined between RTN01 and RTN02

With the overall aim of improving our understanding of the potential historical processes that have lead to the current biodiversity of the two species, three objectives have been set for this section of the study:

1. To complete a phylogeographic analysis of *A. cilicata* and *A. norvegica* across the sampled European localities based on phylogenetic and haplotype network analysis of all the revealed haplotypes.

2. To evaluate the population genetic structure of *A. ciliata* and *A. norvegica* based on the haplotype polymorphisms within and across the populations from different localities.

3. To investigate the likely demographic history of the populations of *A. ciliata* and *A. norvegica* based on molecular dating with the haplotype genealogy and the frequency data.

3.2 Methods

3.2.1 Haplotype phylogeny in *A. ciliata* and *A. norvegica*

In total 20 haplotypes of the rps16 intron and 24 haplotypes of the trnT-trnL intergenic spacer were revealed by HRM analysis (Chapter 2), comprising 32 concatenated haplotypes from the sampled populations of *A. ciliata* and *A. norvegica*. The sequences of the haplotypes were checked and aligned within BioEdit 7.1.3 (Hall 1999) using the function of Clustal*W* (Thompson *et al.* 1994). The haplotypes of *A. ciliata* were named as RTC01-28 and the haplotypes of *A. norvegica* were named as RTN01-04, which is consistent with the description in Chapter 2. The rps16 sequences of all the *A. ciliata* and *A. norvegica* haplotypes were aligned with a length of 777bp and the trnT-trnL sequences were aligned into 611bp, totalling 1388bp when the two aligned sequences were concatenated. With the sequence of *A. serpyllifolia* included as outgroup the concatenated haplotypes were aligned into 1410bp. Variation in simple sequence repeats (SSRs or microsatellites) was removed for phylogenetic analysis because it is evolutionarily labile and generates homoplasious characters, potentially providing misleading information (Small *et al.* 1998), and many researchers omit SSRs from

phylogeographic analyses for this reason (Mast *et al.* 2001; Guggisberg *et al.* 2006; Garcia *et al.* 2011). With polymorphic SSRs removed, the concatenated sequences were aligned into 1390bp.

Prior to analysis, the best-fit evolutionary model of nucleotide substitutions was selected with the software jModeltest (Posada 2008) from the 88 candidate models. Both the Akaike's information criterion (AIC) (Akaike 1973) and the -lnL values suggested the GTR+G model to be the best fit model for the concatenated rps16 and trnT-trnL sequences (-lnL=2551.0039; base frequencies: freqA = 0.4038, freqC = 0.1122, freqG = 0.1433, freqT = 0.3408; gamma distribution shape = 0.1560; substitution rates of different categories: R[AC] = 1.0631, R[AG] = 0.5694, R[AT] = 0.3576, R[CG] = 1.1870, R[CT] = 1.1451, R[GT] = 1.0000).

The phylogenetic relationship among the concatenated rps16+trnT-trnL haplotypes was inferred using the maximum-likelihood method within the software phyML 3.0 (Guindon *et al.* 2009), with the GTR+G model and the substitution rates suggested by jModeltest. The default settings were used for other parameters. A bootstrap test with 1000 replicates was used to provide the support values for the clustering of branches (Felsenstein

1985). The linearized maximum likelihood phylogenetic tree was constructed within the software MEGA 5.05 (Tamura *et al.* 2011) based on GTR+G model with six discrete Gamma categories for Gamma distributed substitution rates among sites. Maximum likelihood phylogenetic trees were also made for rps16 and trnT-trnL separately, and no contradictory clustering was seen between the trees for the two loci.

The statistical parsimony network (Templeton *et al.* 1992) of the identified haplotypes was inferred with the software TCS 1.21 (Clement *et al.* 2000). Gaps were treated as a fifth state in order that both nucleotide substitutions and indels were taken into account. Each indel of multiple contiguous nucleotides was treated as a single mutation event (except the indels in SSRs are ignored for the reason described above), as is normally done in phylogeographic analysis (Jakob & Blattner 2006; Sosa *et al.* 2009). The connection limit was set at >31 steps so that a single complete network was generated, otherwise the haplotypes are separated to two isolated networks if the percentage connection limit was set above 95% by default.

As assessed by Woolley *et al.* (2008), the statistical parsimony method outperforms the minimum spanning method (Rohlf 1973; Excoffier &

Smouset 1994) in constructing genealogical network with less errors among different haplotypes of the same species, especially when the substitution rate is high as seen from the *Arenaria* species in our study. The statistical parsimony method is favoured also because it shows hypothetical intermediate haplotypes between the extant ones, providing a more complete idea of the genealogical history. To evaluate the inferred networks of the two methods, minimum spanning networks were also constructed using the software Arlequin 3.5.1.3 (Excoffier & Lischer 2010), and this generated a similar topology in agreement with the statistical parsimony method (Appendix Figure 2).

3.2.2 Population genetic diversity

For population genetic analysis, the two sample sites from Pyrenees (Py1 and Py2) were merged as a single population as they were collected 1km apart on the same river floodplain and shared the same geographical position information. Also, the two sample sites from Switzerland (Sw1 and Sw2) were considered as a single population as they were from adjacent sites less than 500m apart. In total 23 populations were defined including the Italian population, It2, with only one individual. The distribution information of the detected haplotypes among the 23

populations was then used to analyze the genetic diversity with the software Arlequin 3.5.1.3 (Excoffier & Lischer 2010), wherein the populations were divided into eight groups based on taxonomic status (dividing *A. norvegica* from *A. ciliata*) and then regional geographic distributions as shown in Table 3.3.

Calculation of standard genetic diversity indices, pairwise F_{ST} values among the populations and the analysis of molecular variance (AMOVA) (Excoffier *et al.* 1992) were conducted in Arlequin to evaluate the genetic variation within and among populations and groups. The permutation number was set at 1000 and pairwise difference was used to compute inter-population genetic distance for AMOVA.

In addition to the AMOVA analysis based on our own grouping of the populations, a spatial analysis of molecular variance with the software SAMOVA (Dupanloup *et al.* 2002) was performed to explore possible natural grouping scenarios (without any *a priori* bias). Geographic location information of each population was inputted to the software, and the K value that determined the potential number of groups in each permutation test was set from 2 to 20 to generate fixation index values representing inter-group variation (F_{CT}). The K number with the highest

significant value of F_{CT} would be considered an optimal number of groups. The other two fixation indices, F_{SC} and F_{ST} representing the inter-population genetic variation within groups and across groups are also informative in that lowest F_{SC} and stabilized F_{ST} are expected with the optimal K number.

To consider the possibility of discrete migratory histories for individual haplotype clades, SAMOVA analysis was also performed when only clade I or clade II/III haplotypes were included, considering that intra-clade generic variation may provide clearer pattern of population structures. When only clade I haplotypes were included for analysis, 10 of the populations listed in Table 3.3 were involved. K values from 2 to 8 were thus tested in the SAMOVA analysis to find an optimal grouping plan. Also when only clade II and III haplotypes were included, 18 of the above populations were involved and K values from 2 to 12 were tested in SAMOVA analysis.

Besides calculation of pairwise F_{ST} values, the exact test of population differentiation (Raymond & Rousset 1995) was also carried out within the software Arlequin, with 100,000 steps of Markov chain and 10,000 dememorization steps.

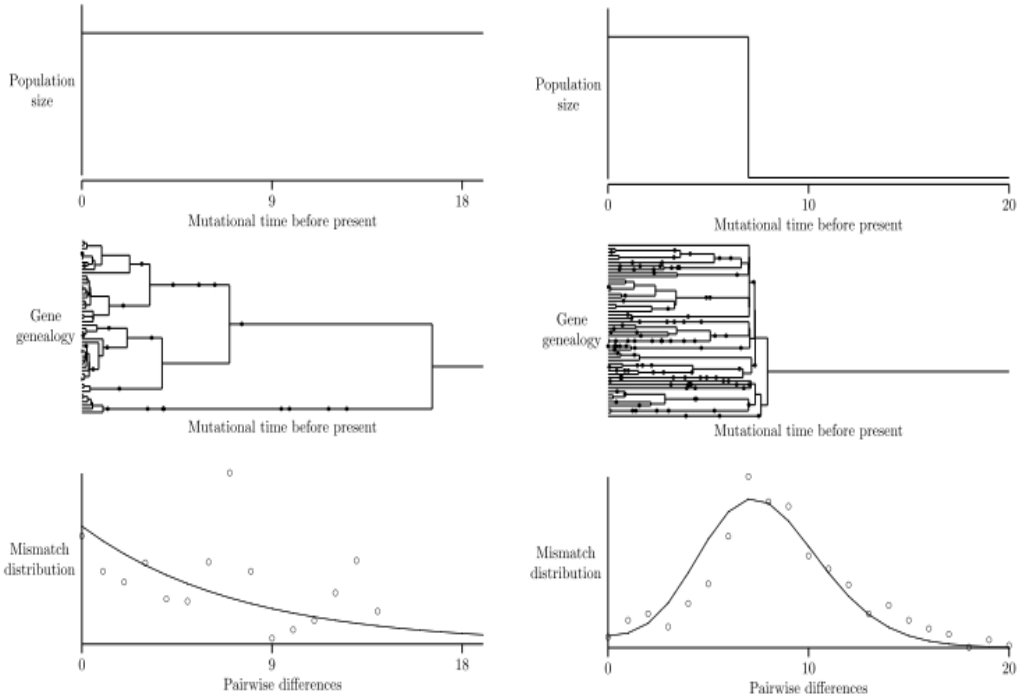
3.2.3 Demographic history

Spatial and/or demographic expansion of populations can often (but not always) leave distinctive patterns of genetic diversity among individuals within the population. It is worthwhile to test for these patterns as they can help describe more clearly the most likely history for each population. Mismatch distribution analysis (Harpending 1994) was thus carried out to investigate if they experienced historical expansion events and, if so, when the expansion events occurred.

The mismatch distribution comes from a simple calculation, where random pairs of individual samples are taken from a target population and the sum difference in DNA sequence at a particular haploid locus is recorded as *i*; the number of all possible pairs of individuals that differ by *i* nucleotide sites is recorded as F_i, and for all possible *i* differences from 0 to the largest observed value, the corresponding F_i values are calculated, so that the distribution of F_i over the *i* values is obtained, which is termed the mismatch distribution (Rogers & Harpending 1992).

The shape of the mismatch distribution with an extant population is determined by the historical demographic events it has experienced,

including expansion and decline. If a population has maintained a constant size during its history, the pairwise differences (mismatches) are expected to obey an equilibrium distribution while the observed mismatch distribution usually appears to be ragged and erratic (Rogers & Harpending 1992; Harpending 1994). However if a population has experienced a sudden expansion in its history, the mismatch distribution will appear to be a smooth unimodal curve with a single peak, the position of which is positively correlated to the expansion time before the present. Box 3.1 illustrates the difference between a stationary population and an expanded population in mismatch distribution analysis, where the Figures are from Rogers' Lecture Notes on Gene Genealogies (from Rogers 2004).



Box 3.1 An illustration of mismatch distribution in different situations. The left panel shows the situation when the population maintains a constant size, where the mismatch distribution is expected to be an equilibrium curve (solid line) and the observed data usually give a ragged distribution (circles). The right panel shows the situation when the population experienced a sudden expansion at 7 units of mutational time before the present, where the mismatch distribution appears to be a smooth curve with a peak at 7bp of pairwise differences. (Cited from Rogers' Lecture Notes on Gene Genealogies, 2004 with kind permission from the author.)

Excoffier (2004) proposed that spatial range expansion may produce the same pattern of mismatch distribution as sudden demographic expansion when migration rate between subpopulations is reasonably large, as is highly probable during dispersal events in postglacial colonization.

However, the mismatch distribution assumes the population is not under selective pressure when the demographic history is under investigation,

which may not be true in a given realistic case. Also insufficient sampling and inefficient haplotype detection may cause misleading results of mismatch distribution analysis, thus cautions should be taken when the hypothesis of an expansion model is accepted.

To evaluate whether any populations containing clade I and/or clade II/III haplotypes showed evidence of historical expansion events, a mismatch distribution test (Harpending 1994) was performed. Populations from different localities were analyzed both separately and pooled as meta-populations to investigate the expansion patterns that may be evident at different geographic and taxonomical scales. Haplotypes from clades IV and V were removed from all the populations for this test, as they are distinct lineages that may not be suitable for analyses at the population genetics level.

The mismatch distribution test with the frequency distribution of pairwise nucleotide difference between individuals was performed with the software Arlequin 3.5.1.3 (Excoffier & Lischer 2010). Parameters in both models of demographic expansion (Schneider & Excoffier 1999) and spatial expansion (Excoffier 2004) were estimated, the pairwise difference was used for molecular distance and 1000 bootstrap replicates

were used for the mismatch distribution analysis. The significance (p value) for the sum of squared deviation (SSD) is used to judge if the expansion model is accepted or rejected. For the populations or pooled units that emerged as being subject to expansion, the expansion time was estimated from the moment estimator τ via the formula $\tau=2\mu t$, where t is the number of generations after the historical expansion up to the present and μ is the mutation rate of the whole DNA locus used for study (Rogers & Harpending 1992). The per nucleotide substitution rate is estimated to be $2.9\text{-}4.8\times10^{-9}$ per nucleotide per year (see result in section 3.3.1), and as the DNA locus was aligned into 1351bp for the analysis, the value of μ is estimated as $3.9\text{-}6.5\times10^{-6}$ per generation.

Because the above test requires the DNA locus used to be evolutionarily neutral, Tajima's *D* (Tajima 1989) and Fu's Fs (Fu 1996) were also calculated in each case to see if the locus was under selection, although the rps16 intron and trnT-trnL spacer are assumed to be neutral because they are non-coding DNA regions. In all the cases of this section the *D* and Fs statistics were not significant so that the neutral hypothesis was not rejected.

3.3 Results

3.3.1 Haplotype phylogeny in *A. ciliata* and *A. norvegica*

The inferred phylogeny among the haplotypes is illustrated in Figure 3.1 and Figure 3.2. In Figure 3.1 the branch lengths are shown to represent the accumulated substitutions of each haplotype, and in Figure 3.2 the phylogenetic tree is linearized to show the relative time of divergence between different clades.

Both Figure 3.1 and 3.2 show the same topology of the phylogeny among the detected haplotypes. It is seen that the *A. ciliata* haplotypes are not monophyletic, and include *A. norvegica* haplotypes within the various observed clades. The *A. norvegica* haplotypes are not monophyletic either, falling into two different clades, each clustered with a different group of *A. ciliata* haplotypes, i.e. RTN01-03 form a single clade clustered with RTC 17 and 19-21 (with 98% confidence) while RTN04 forms a distinct clade clustered with RTC14-16 (with 93% confidence).

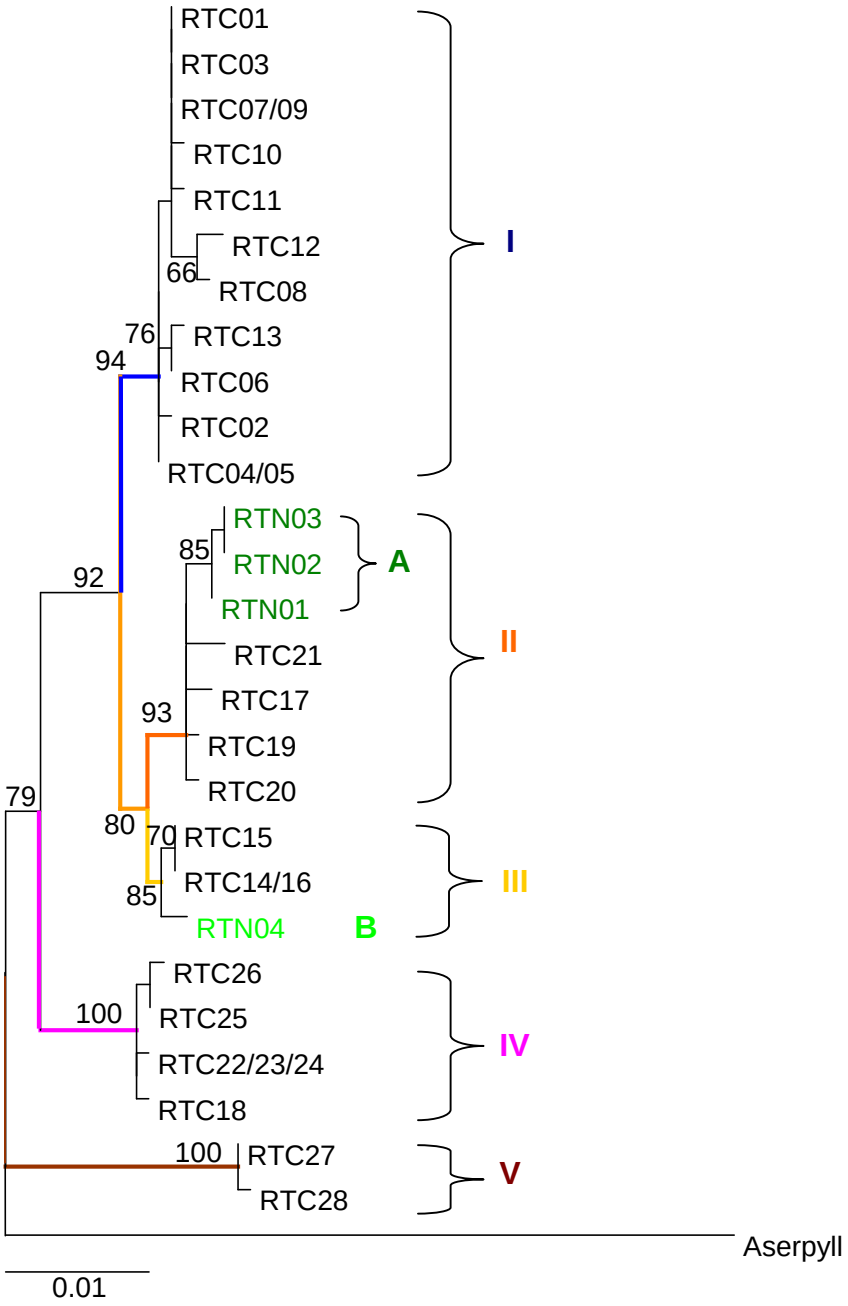


Figure 3.1 The inferred phylogenetic tree based on concatenated rps16+trnTL haplotypes (1390bp aligned after indels in SSRs are removed) using the maximum likelihood method with branch lengths showing the substitution rate. The numbers beside the branches are support values based on bootstrap analysis with 1000 replicates (only those >50% are shown); the sequence of *A. serpyllifolia* is used as outgroup (shown as Aserpyll) to root the tree. All the haplotypes are designated into five clades (I to V) with the branches in different colours. The *A. norvegica* haplotypes fall into two clades (texts in green and marked as clades A and B), each falling within a clade group of *A. ciliata*.

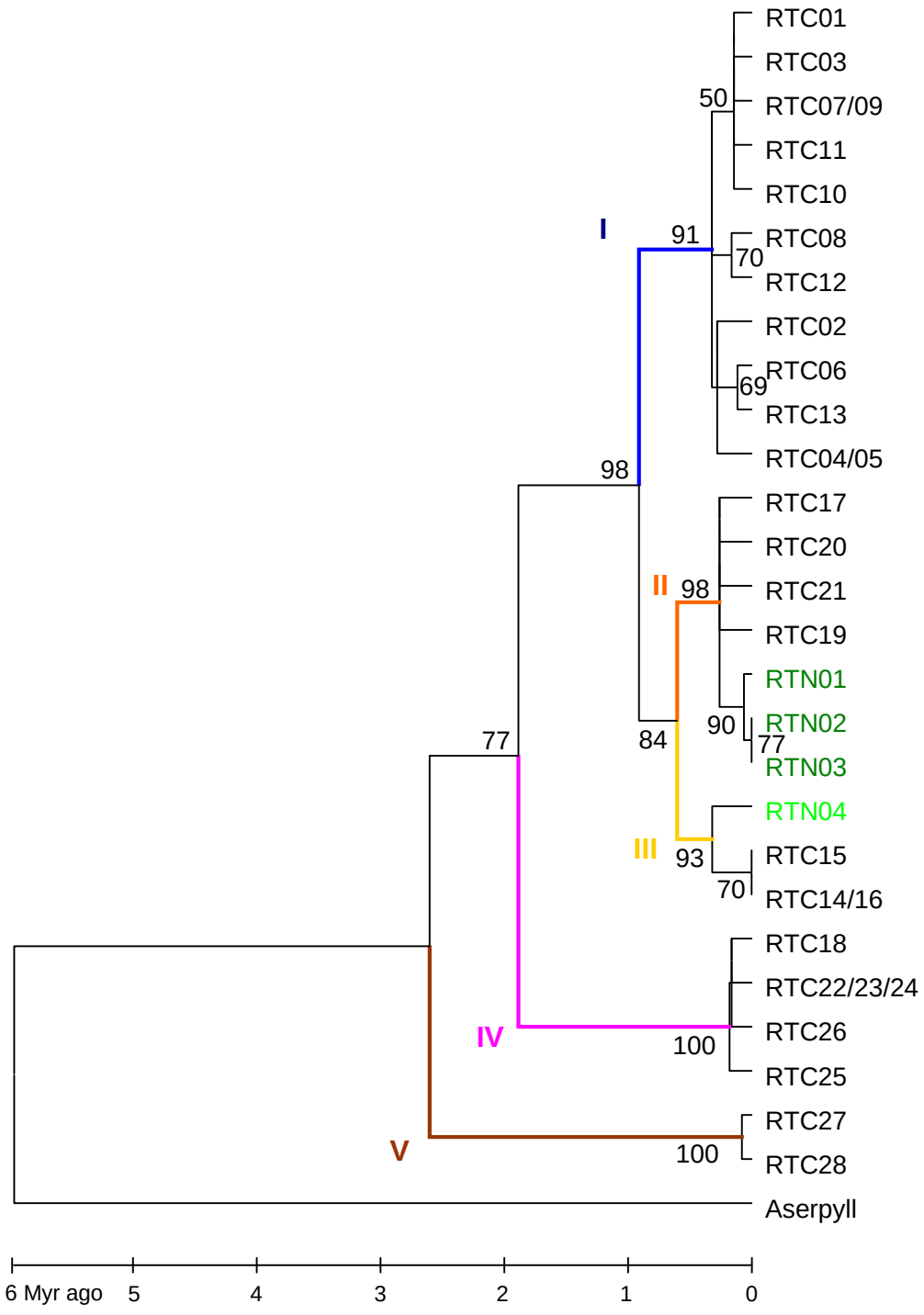


Figure 3.2 Linearized Maximum likelihood tree based on concatenated rps16+trnTL haplotypes (1390bp aligned after indels in SSRs are removed) with support values based on bootstrap analysis with 1000 replicates (only numbers greater than 50% are shown). *A. serpyllifolia* is used as the outgroup (shown as Aserpyll). The scale line below the tree shows the lower estimation of divergence time.

The haplotypes of the two studied species are divided into five distinct clades with the clustering of each group supported by 85-100% confidence based on bootstrap test (clade I to V in Figure 3.1 and 3.2). It is interesting that the *A. norvegica* haplotypes are divided into two clades, with clade A falling within clade II and clade B within clade III.

The divergence between *Arenaria serpyllifolia* and the lineages falling within the *A. ciliata* Complex is estimated around 6-10 million years (Myr) ago based on the fossil evidence from Valente *et al.* (2010) (Appendix Figure 1). The mutation rate of chloroplast genome in the studied taxa is meanwhile estimated at $2.9\text{-}4.8\times10^{-9}$ per nucleotide per generation (and thus per year because *A. ciliata* is an annual herb) assuming constant mutation rate of the sequences under investigation, which is close to the estimation by Wolfe *et al.* (1987) at 1.0 to 3.0×10^{-9} per nucleotide per year for chloroplast genomes of plants. From the linearized tree in Figure 3.2, it is estimated the divergence of clade V from other clades occurred around 2.6-4.3 Myr ago, the divergence of clade IV occurred 1.9-3.2 Myr ago, the divergence between clade I and II/III occurred 0.9-1.5 Myr ago, and the divergence between clades II and III occurred 0.6-1.0 Myr ago.

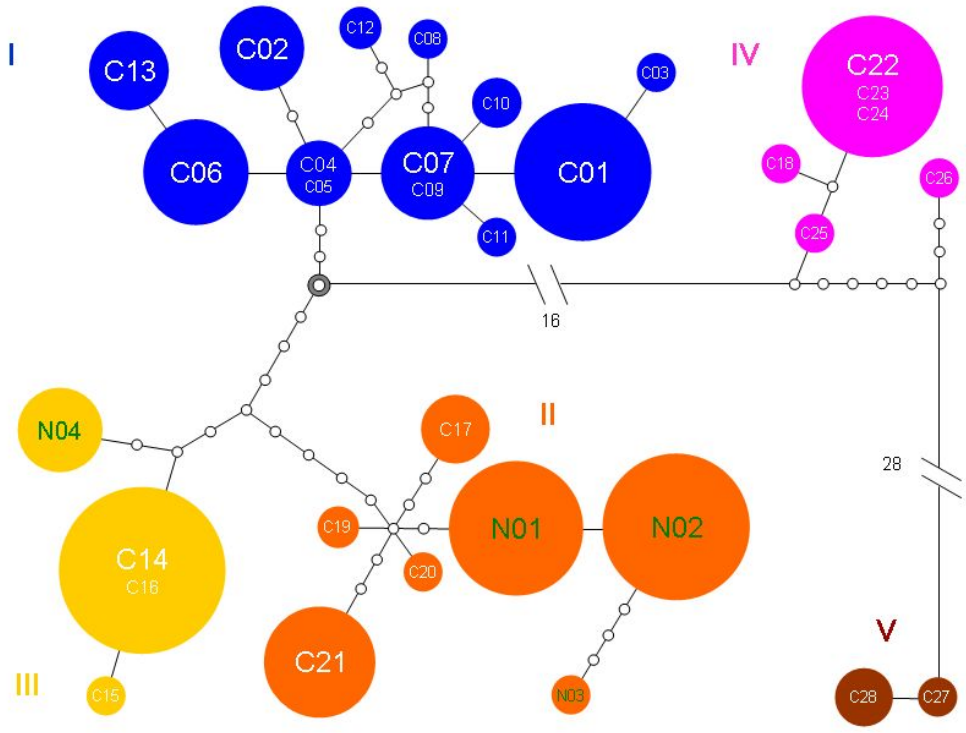


Figure 3.3 Statistical parsimony network based on concatenated rps16+trnTL haplotypes. Each circle with text represents a haplotype or a few haplotypes varying only in microsatellites, the area of which is semi-proportional to the number of individuals sharing the haplotype(s), as the extant haplotypes with <10 carriers are represented by circles of equal size. The small circles without text represent hypothetical intermediate haplotypes varying by one mutational step from the aside ones. In agreement with the phylogenetic tree in Figure 3.1 and 3.2, the haplotypes are clearly divided into five clades shown in different colours. The text for RTN01-04 is shown in green while the text for RTC01-28 in white, to indicate their identities as different previously recognized species. It is noted that the numbers are undetermined between the carriers of N01 and N02 in the populations from Rum Island and Inchnadamph in Scotland and the two haplotypes were assumed to be equally represented in the two populations in preparation of this network. The ring-shaped small circle is inferred as the diverging node for all the clades, while the haplotype RTC07/09 is identified as the most ancestral haplotype by the software TCS 1.21.

The statistical parsimony network of the haplotypes is generated from TCS 1.21 and then redrawn for clarity (Figure 3.3). Each haplotype is

indicated by a circle with text, i.e. C01 for the haplotype RTC01, while the small circles without text indicate hypothetical intermediate haplotypes. All haplotypes as nodes are connected via straight lines showing their relationship. Each node is different from its nearest neighbour by one mutational step.

The haplotype network in Figure 3.3 shows clearly that the haplotypes are divided into five clades as recognized in the phylogenetic trees from Figure 3.1 and 3.2. The two clades of *A. norvegica* haplotypes are included in clades II and III of *A. ciliata* respectively. The overall topology of the network suggests that clade II and III are closest sister lineages which in turn are clustered with clade I. While there is a hypothetically common ancestor haplotype (shown by the ring-shaped node in Figure 3.3) for all the five clades, clade IV and V appear to be more distinct lineages. The link from the clades IV and V to other clades may not reflect the true genealogy here because of the deep divergence between them and other clades and lack of identified intermediate haplotypes. Clade IV and V were assigned to isolated network motifs when the default connection limit is used within the software TCS (both clades isolated at 95% and clade V isolated at 90%), indicating their divergence beyond intra-specific level.

Within clade II, the haplotypes of *A. norvegica* form a monophyletic lineage (clade A), but share the same hypothetically ancestral haplotype with the *A. ciliata* haplotypes by similar mutational steps, except RTN03 is 7 steps away from the ancestral haplotype. In clade III, the RTN04 haplotype of *A. norvegica* also has a co-ancestor with RTC14, 15 and 16 of *A. ciliata*. Both the phylogenetic tree and the haplotype network suggest that the previously identified *A. norvegica* contains two different chloroplast lineages under the species name of *A. ciliata*, without sufficient accumulated mutations separating them from the latter species.

3.3.2 Phylogeography of *A. ciliata* and *A. norvegica*

The distribution of the five clades is illustrated on the map of Europe as shown in Figure 3.4. In order to distinguish the previously identified species *A. norvegica*, its occurrence is indicated in green colours (dark green for clade A and light green for clade B). Populations from the same locality as defined in Table 3.1 are pooled as a single geographic unit shown by the corresponding circle on the map. One individual of *A. norvegica* from Norway is also indicated on the map, the rps16 sequence of which was obtained from Genbank (Accession No. HM772117;

Westergaard *et al.* 2011) indicating that the sample belongs to RTN01 or RTN02 in clade A.

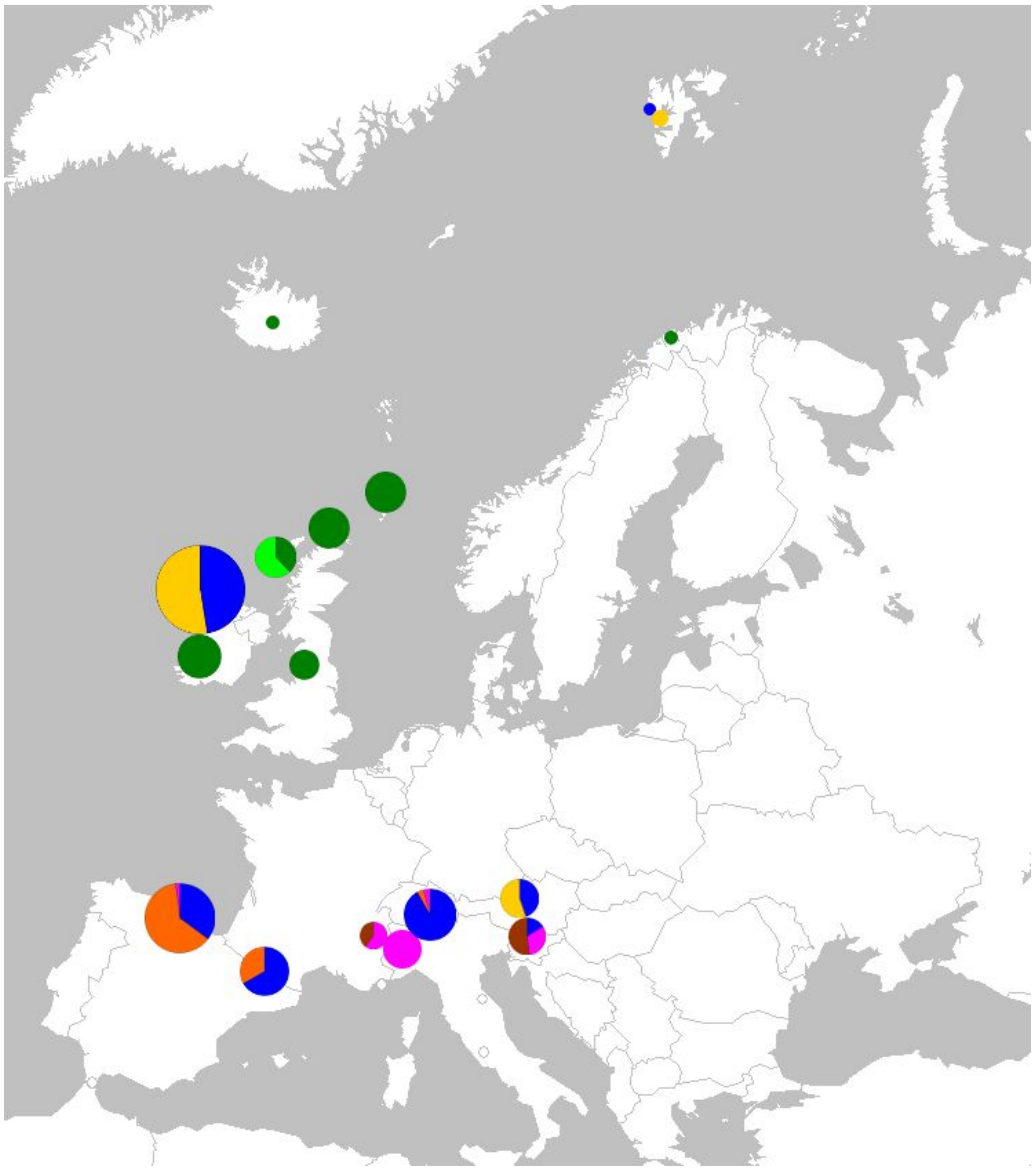


Figure 3.4 Distribution of the five haplotype clades among the populations from across Europe. Each circle indicates a locality according to Table 3.1. The pie diagram shows the haplotype clades and their frequencies in each locality. The colour coding for haplotype clades remains the same as in Figures 3.1-3.2, where the *A. norvegica* clades are shown in greens. Note that the green circle in Norway is based on an rps16 record of *A. norvegica* in Genbank (Accession No. HM772117, from Westergaard *et al.* 2011).

The haplotypes in clade I (shown in blue) are the most common identified in *A. ciliata*, found from seven localities including the Ben Bulben mountains in northwest Ireland, the mountainous area of Picos de Europa in the north side of Spain, the Pyrenees Mountains, the Alps mountains in Austria and Switzerland, and the islands of Svalbard, which cover almost the entire distribution of the species. Clade II (shown in orange) is found to coexist with clade I in the populations from Picos, Pyrenees and part of Alps in Switzerland, but is replaced by clade III (shown in golden yellow) in Ireland and the east side of Alps in Austria, where clade I is also found. Clades I and III are both found in Svalbard from different populations. Considering the small sampling size and the short distance between the two populations from Svalbard, it is possible that clades I and III could coexist in some populations in Svalbard as well.

Clade IV (in pink) is found to grow on both east and west sides of Alps, with rare occurrence on the Picos mountains (Pi2.8 as the only carrier of RTC18 in the population) and in one of the Swiss populations (Sw2.28 carrying RTC24). So in two populations (Pi2 and Sw2) clades I, II and IV are found to coexist. In populations Au2 and Fr1 from the east and west sides of Alps, clade IV is found to coexist with clade V (in brown), which gives two more cases where distinct clades co-occur in the same

population.

Based on the phylogenetic trees in Figure 3.1-3.2 and the statistical parsimony network in Figure 3.3, clade I contains the biggest number (11) of haplotypes which diverged by relatively fewer mutations. As clade II and III share a common ancestor and are not found to coexist in any locality, in this analysis they form a single phylogenetic cluster that includes all *A. norvegica* haplotypes (and this grouping is applied in later analysis below). Clade II and III jointly contain 10 haplotypes, which vary by more mutations compared to those in clade I. Due to the evident divergence between clade I and clades II/III, it may help if we analyze clade I and clades II/III both separately and jointly to better understand their phylogeographic history.

The geographic distribution of the haplotypes in clade I is illustrated beside the statistical parsimony network among the haplotypes in Figure 3.5. The distribution of the haplotypes in clade II and III is illustrated beside the statistical parsimony network among the involved haplotypes in Figure 3.6. The extent to which the two haplotype clusters overlap each other in their geographic distribution is clear on the two maps.

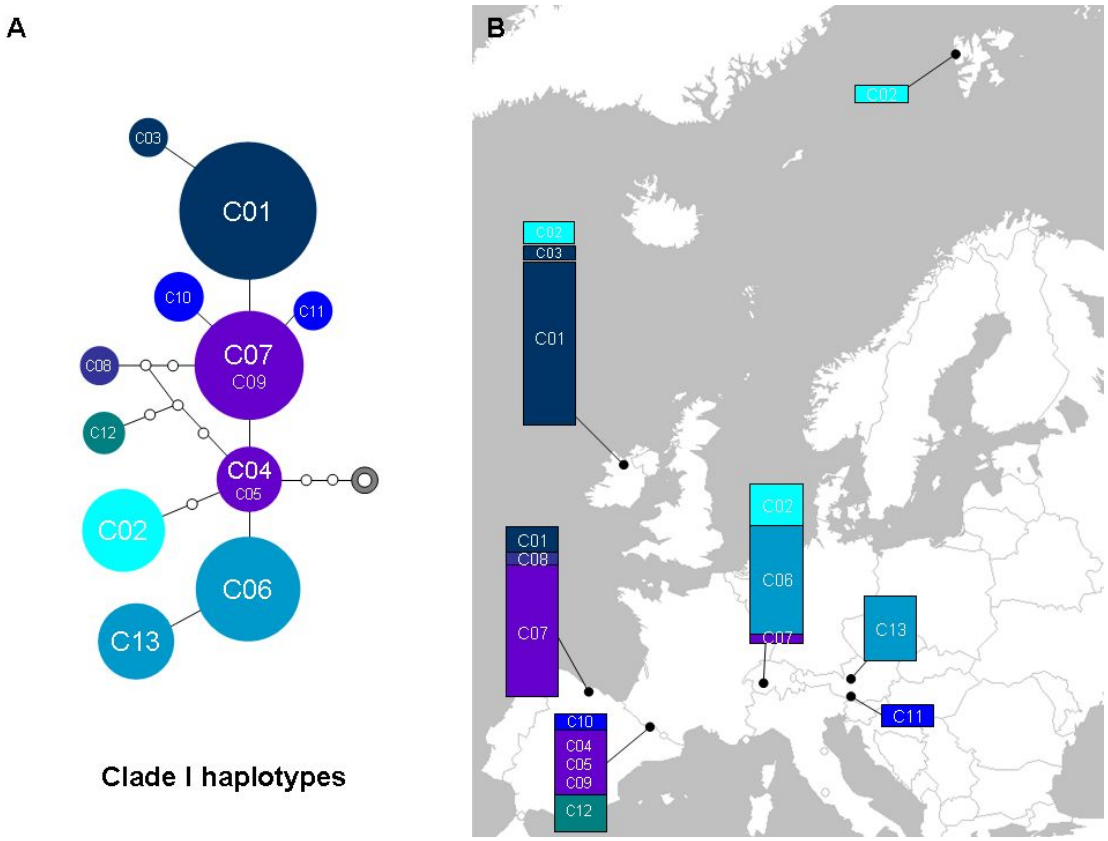


Figure 3.5 The statistical parsimony network among the haplotypes in group I as inferred from TCS (A) and the geographic distribution of the haplotypes (B).

A. The solid line between each pair of circles indicates one mutational step. Each circle with text indicates an extant haplotype, e.g. C01 represents the haplotype RTC01, the area of which indicates the number of individuals sharing this haplotype. Haplotypes RTC07 and 09 (and C04 and 05) vary only in SSRs and are treated as a single haplotype in all analyses. The small circles without text indicate hypothetical intermediate haplotypes, and the ring-shaped circle is the diverging node as shown in Figure 3.3. RTC04/05 and C07/09 are shown in the same colour as they are inferred as the most ancestral extant haplotypes; C10 and C11 are shown in the same colour as they are both one-step mutants of RTC07/09.

B. On the right panel, each column represents the samples from each locality, as described in Table 3.1 and Figure 3.4. The small dots indicate the sampling sites. The proportion of different haplotypes within each locality is shown by different colours in the column, in accordance with the colour coding in panel A.

As shown in Figure 3.5, the clade I of *A. ciliata* contains 11 haplotypes when SSRs are not taken into account, which makes clade I the biggest clade among all the five clades. The most common haplotype is RTC01, which is found both from the Ben Bulben mountain tops in northwest Ireland and from the Picos mountains on the north seaside of Spain. However, the abundance of this haplotype is due to the large sampling size in Ireland (50 carriers of RTC01 out of 120 individuals from four populations). The haplotype RTC03 is one mutational step from RTC01, in fact by a single deletion of six contiguous nucleotides, which is found only in Ireland. The second most common haplotype is RTC07 (including RTC09), which is also the most widely distributed haplotype found from the Picos mountains, the Pyrenees in Spain and the Gemipass mountains in Switzerland. Another most abundant haplotype is RTC06 found only in Switzerland, with its one-step mutant RTC13 found only from one population on the north side of Alps in Austria.

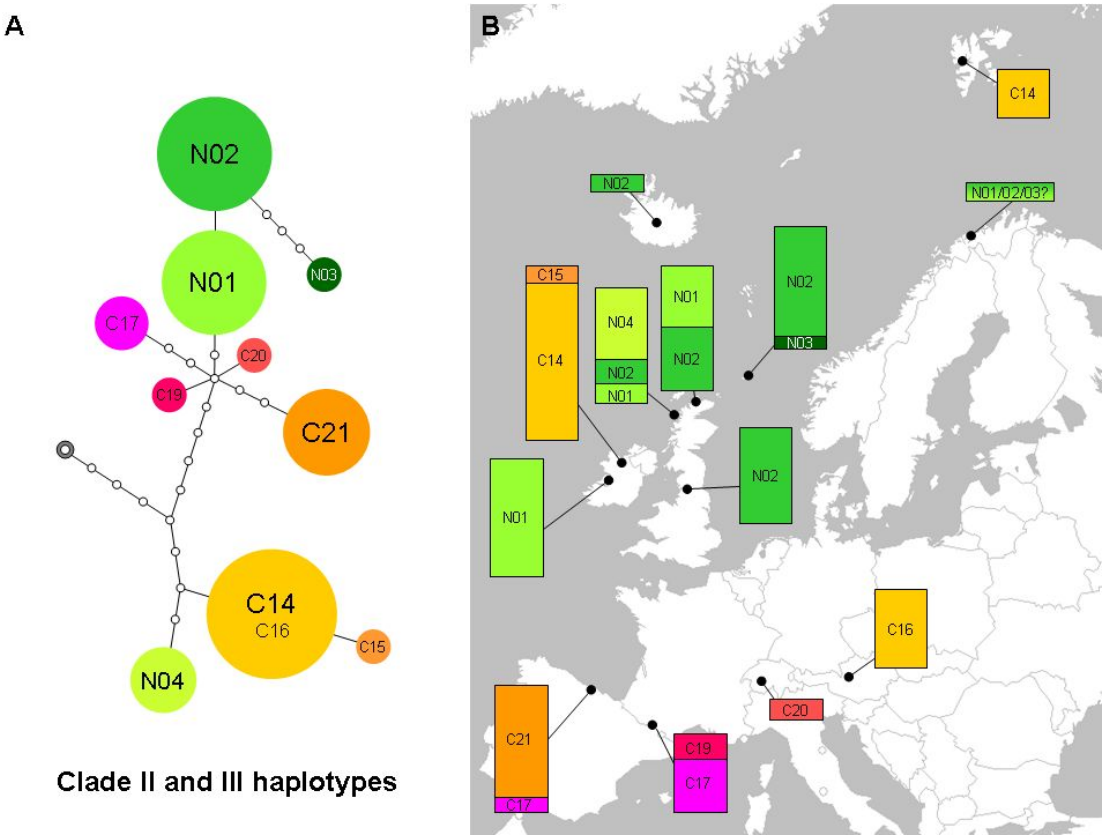


Figure 3.6 The statistical parsimony network among the haplotype in clade II and clade III as inferred from TCS (A) and the geographic distribution of the haplotypes (B).

A. The solid line between each pair of circles indicates one mutational step. Each circle with text indicates one haplotype, e.g. C15 represents haplotype RTC15, the area of which indicates the number of individuals sharing this haplotype. Haplotypes RTC14 and 16 vary only in SSRs and are treated as a single haplotype in all analyses. The small circles without text indicate hypothetical intermediate haplotypes, and the ring-shaped circle is the diverging node as shown in Figure 3.3.

B. On the right panel, each column represents the samples from each locality, as described in Table 3.1 and Figure 3.4. The small dots indicate the sampling sites. The proportion of different haplotypes within each locality is shown by different colours in the column, in accordance with the colour coding in panel A. It is noted that 1) The information of the Norwegian sample is obtained from Genbank, where only rps16 sequence is available so that it is unknown if the sample belongs to RTN01, RTN02 or RTN03; 2) In the two populations from northwest Scotland (Inchnadamph and Rum Island), the proportion between RTN01 and RTN02 is unknown and assumed to be 1:1 approximately.

There are 10 haplotypes in clades II and III when SSRs are not considered, including the four haplotypes of *A. norvegica*. As seen from the haplotype network and the phylogenetic trees (Figure 3.1-2), there is a deep divergence between clades II and III, which is estimated 0.6-1.0 Myr ago. The two clades do not co-occur except in the population from Rum Island where RTN01 and RTN02 in clade II are found to coexist with RTN04 in clade III.

The genealogical network shows a hypothetical ancestral haplotype of clade II which is not found in the sampled populations. This haplotype may have extinguished or become too rare to be sampled in the current populations. One of its one-step mutant haplotype, RTC19, is found in Pyrenees while the other of its one-step mutant, RTC20, is found from Switzerland, both with very few carriers (four of RTC19 and one of RTC20). One of its three-step mutant haplotypes, RTC17, is found from both Pyrenees and Picos with ten carriers of the haplotype while the other three-step mutant haplotype, RTC21, found only in Picos with 30 carriers.

Besides the four mutant haplotypes from the hypothetical ancestor, there is another two-step mutant haplotype, RTN01, which is found in at least three populations from Ireland and Scotland, as one of the most common

haplotypes of *A. norvegica*. The other most common haplotype of *A. norvegica*, RTN02 as the one-step mutant from RTN01, is widely distributed in England, north Scotland, the Rum Island, Iceland, and an isolated population from the islands of Shetland. In Shetland an endemic haplotype, RTN03, is found as a four-step mutant of RTN02, indicating a long history since the isolation of the population. In summary, the three haplotypes in clade A of *A. norvegica* consist of a single lineage distinct from the four other lineages in clade II of *A. ciliata*. These four lineages are restricted to continental refugia across the Picos, Pyrenees and Alps, while the lineage recognized as *A. norvegica* has a distinct northern distribution scattered in a number of northwestern islands and Scandinavia. The remaining *A. norvegica* haplotype, RTN04 has a different history and is most closely associated geographically and genetically with the RTC14 haplotype extant on Ben Bulben in Ireland.

3.3.3 Population genetic diversity and structure

The average gene diversity (h) and nucleotide diversity (π) within each population are listed in Table 3.3 to show the genetic diversity at the intra-population level.

It is seen that the populations from Pyrenees, East Alps, Picos and Ireland in turn showed the highest genetic diversity in terms of haplotype polymorphisms. The nucleotide diversity largely agreed with the gene diversity but gave the highest values for the east and southwest Alps, the two places that contain most of the carriers of the deeply divergent haplotypes in clades IV and V.

Table 3.3 The gene diversity (*h*) and nucleotide diversity (π) for each population to show intra-population genetic variation.

Group	Population	Gene diversity (*h*)	Nucleotide diversity (π)
Ireland	Ir1	0.4046	15.3747
	Ir2	0.1287	6.5655
	Ir3	0.6828	20.2989
	Ir4	0.0667	2.5333
Picos	Pi1	0.3743	7.7544
	Pi2	0.6190	24.2095
	Pi3	0.0000	0.0000
Pyrenees	Py	0.8587	18.0143
Southwest Alps	It1	0.0000	0.0000
	It2	--	--
	It3	0.5000	20.0000
	Fr1	0.6000	43.2000
North side of middle-west Alps	Sw	0.5300	13.8453
East Alps	Au1	0.5185	21.7778
	Au2	0.6267	41.8267
Svalbard	Sv1	0.0000	0.0000
	Sv2	0.0000	0.0000
A. norvegica, arctic and sub-arctic islands	NB	0.0000	0.0000
	NE	0.0000	0.0000
	NR*	0.5616	10.5813
	NIn*	0.5172	0.5172
	NS	0.1330	2.7931
	NIc	0.0000	0.0000

* The values of NR and Nin are inaccurate as the frequencies of RTN01 and RTN02 are undetermined in the two populations.
-- The values are unavailable for It2 as it contains only one individual sample.

Based on our own grouping of the populations as shown in Table 3.3, the result of AMOVA (Table 3.4) shows that 41.81% of the genetic diversity can be explained by the inter-group variation, which is higher than that explained by intra-population variation (29.71%) and inter-population/intra-group variation (28.47%). All the values of variation percentage are significant to the $p<0.01$ level.

The result of SAMOVA based on all detected haplotypes is shown in Figure 3.7, where the values of fixation indices are plotted against the K numbers. The results of SAMOVA based on separate analyses of clade I and clade II/III haplotypes are shown in Figure 3.8A and B respectively. The p value for every fixation index was less than 0.01 so that all the values in Figure 3.7 and 3.8 are significant.

Table 3.4 Results of the AMOVA on the *A. ciliata* and *A. norvegica* populations by dividing the 23 populations into the eight groups as shown in Table 3.3.

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation	Fixation indices
Among groups	7	4014.632	7.84866	41.81**	F_{CT}=0.4181
Among populations within groups	15	1587.344	5.34488	28.47**	F_{SC}=0.4893
Within populations	457	2549.095	5.57789	29.71**	F_{ST}=0.7029
Total	479	8151.071	18.77143		

** Significant at $p<0.01$.

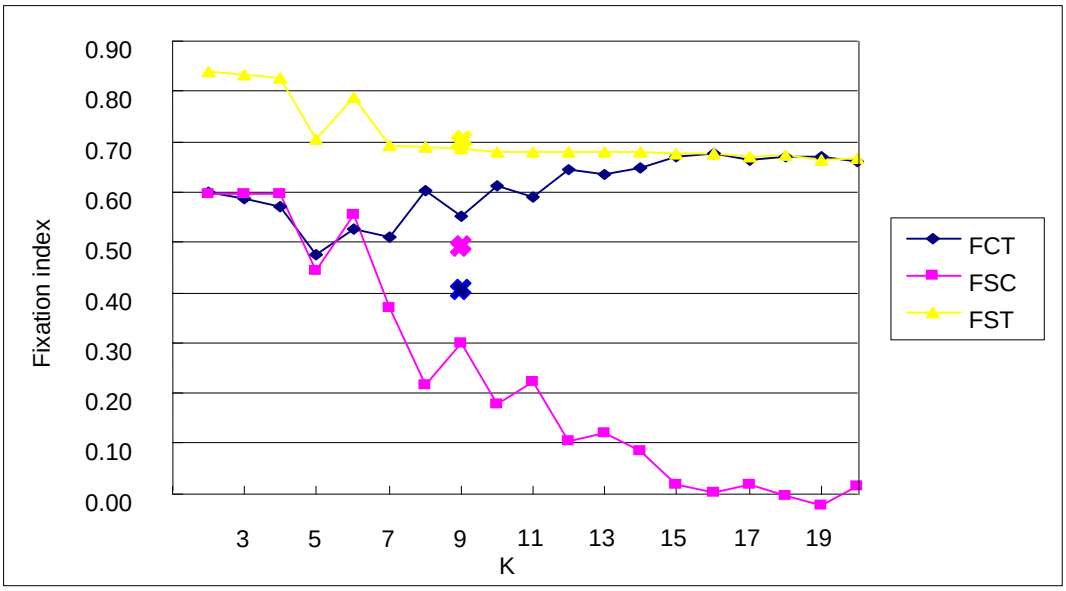


Figure 3.7 Fixation index values plotted against the K numbers of groups, as calculated in SAMOVA. The F_{CT} values (blue) indicate genetic variation among groups. The F_{SC} values (pink) indicate genetic variation among populations within groups. The F_{ST} values (yellow) indicate genetic variation among populations across groups. The K value indicates the number of groups into which the populations are divided. The three individual crosses of each colour indicate the corresponding fixation index values when the populations are divided into the 8 *a-priori* groups described in Table 3.3.

As illustrated in Figure 3.7, the F_{CT} value decreases initially from 0.6 when K increases beyond 2 groups, but then begins to increase when K is greater than 5. The larger K values beyond 6 generate progressively increasing F_{CT} values which reach a plateau of 0.67 when K=16. However, assigning populations into 16 groups seems to dissolve group structure beyond the level where any meaningful grouping of the populations is seen (Table 3.5). The structuring of populations into 8 groups (Table 3.5) results in a different optimal arrangement compared to the initial *a-priori*

grouping based on taxonomic and geographic criteria (Table 3.3). It is also seen from Figure 3.7 that the optimal SAMOVA-suggested grouping of 8 yields a greater F_{CT} value than that of our initial grouping, indicating the new grouping plan explains the genetic diversity better in terms of inter-group variation than the initial *a-priori* grouping.

Table 3.5 Grouping plan based on SAMOVA when K=8 and K=16. All detected haplotypes were included for grouping of the 23 populations listed in Table 3.3.

Group		Populations
K=8	K=16	
1	1	NB, NE, NIn, NS, NIc
	2	Pi2
	3	Pi3
2	4	Ir1, Ir3, Ir4
	5	NR
3	6	Au1
	7	Sv2
4	8	Py
	9	Pi1
5	10	Au2
6	11	It1
	12	It2, It3
	13	Fr1
7	14	Sw
	15	Sv1
8	16	Ir2
F_{SC}: 0.2164	F_{SC}: 0.0028	
F_{ST}: 0.6890	F_{ST}: 0.6757	
F_{CT}: 0.6031	F_{CT}: 0.6748	

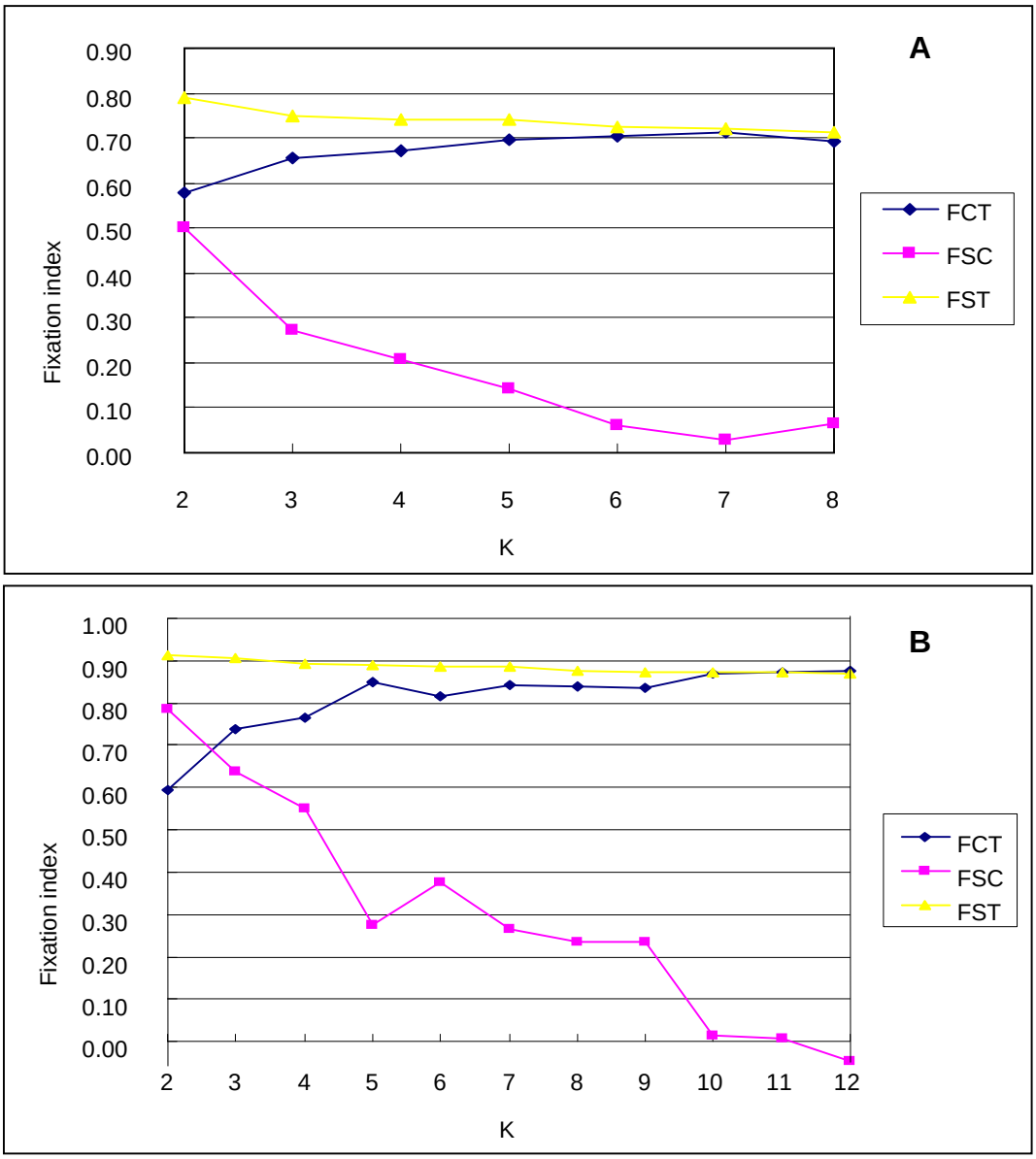


Figure 3.8 Fixation index values plotted against the K number of population groups, as calculated in SAMOVA, for clade I (A) and clades II/III (B), treated separately. The F_{CT} values (blue) indicate genetic variation among groups. The F_{SC} values (pink) indicate genetic variation among populations within groups. The F_{ST} values (yellow) indicate genetic variation among populations across groups. The K value indicates the number of groups into which the populations are divided.

With only clade I haplotypes included in the analysis, it is seen from Figure 3.8A that the F_{CT} value rises to a plateau when K is beyond 5 and

reaches a maximum when K=7. The result suggests the ten populations containing clade I haplotypes should be optimally divided into 5 to 7 groups. With only clade II and III haplotypes included, it is seen from Figure 3.8B that the FCT value plateaus when K reaches 5 achieving a maximum when K=12, suggesting the 18 relevant populations be divided into between 5 and 12 groups optimally. The optimal grouping plans for separate treatment of the clades as indicated by these analyses are listed in Table 3.6 for comparison

It is seen from Table 3.6 for clade I populations, that although K=7 yielded a higher F_{CT} value than K=5, this increase in group number separates Ir3 from other Irish populations and separates Au2 from the group containing Pi1 and Py compared to the grouping suggested for K=5. For the populations containing clade II and III haplotypes, increasing K from 7 to 12 separates i) Ir2 from other Irish populations; ii) Pi2 from other Picos populations; iii) Sw from Py, and iv) NB and NIn from other *A. norvegica* populations (except NR). On the other hand, compared to K=7 for clade II/III, reducing K to 5 merges Au1 and Sv2 with the Irish populations as a single group.

Table 3.6 Grouping plan based on SAMOVA when only clade I or clade II/III haplotypes are included. Different groups are shown in different colours and populations within each group are shown in the same colour.

	Clade I		Clade II and III		
	K=5	K=7	K=5	K=7	K=12
	Ir1	Ir1	Ir1	Ir1	Ir1
	Ir2	Ir2	Ir2	Ir2	Ir2
	Ir3	Ir3	Ir3	Ir3	Ir3
	Ir4	Ir4	Ir4	Ir4	Ir4
	Pi2	Pi2	Pi2	Pi2	Pi2
			Pi3	Pi3	Pi3
	Pi1	Pi1	Pi1	Pi1	Pi1
	Au2	Au2			
	Py	Py	Py	Py	Py
	Sw	Sw	Sw	Sw	Sw
	Au1	Au1	Au1	Au1	Au1
	Sv1	Sv1			
			Sv2	Sv2	Sv2
			NR	NR	NR
			NB	NB	NB
			NE	NE	NE
			NIc	NIc	NIc
			NS	NS	NS
			NIn	NIn	NIn
F_{CT}	0.6962	0.7121	0.8479	0.8435	0.8760
F_{SC}	0.1437	0.0302	0.2743	0.2665	-0.0478
F_{ST}	0.7398	0.7208	0.8896	0.8852	0.8700

By looking at the difference between the grouping plans of clade I and clade II/III, it is seen that the major difference is focused on the populations Pi1, Pi2 and Py. The Picos population Pi2 is grouped with the Irish populations based on the genetic data from clade I, while it is grouped with other Picos populations based on clade II/III. Also the data based on clade I suggests that Pi1, Py and Au2 from Picos, Pyrenees and

Austria respectively fall within a single group, while they are separated based on clade II/III data, in that Pi1 is grouped with other Picos populations and Py is grouped with the Swiss population Sw.

Pairwise F_{ST} values between populations and the results of the population differentiation test are listed in Table 3.7. Most of the F_{ST} values are significantly positive and most of the pairwise population differences are significant, indicating strong isolation among the populations and lack of gene flow. One exception is between Sv2 and Ir4, where the F_{ST} value is negative but close to zero, reflecting their genetic similarity. Indeed according to the exact differentiation test, Sv2 is not significantly differentiated from Ir4. This is in accordance to the fact that the haplotype RTC14 is shared by Sv2 and the Irish populations Ir1, Ir3 and Ir4. The same situation is seen between the *A. norvegica* populations from Iceland and Shetland, NIc and NS, reflecting the fact they both contain the haplotype RTN02. Also it is seen that the Italian population It2 is not significantly differentiated from It3 and the French population Fr1, whereas It3 and Fr1 are significantly differentiated but with the pairwise F_{ST} close to zero.

Table 3.7 Pairwise F_{ST} values between all samples populations (below the diagonal) and the significance of population differentiation (above the diagonal). '+' indicates significantly different and '-' indicates not significantly different.

	Ir1	Ir2	Ir3	Ir4	Py	Pi1	Pi2	Pi3	Sw	Au2	Au1	Sv1	Sv2	NB	NE	NR	NIn	NS	NIc	It1	It2	It3	Fr1
Ir1		+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+
Ir2	0.60		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Ir3	0.21	0.22		+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	-	+	+
Ir4	0.17	0.87	0.54		+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	-	+	+
Py	0.36	0.40	0.15	0.62		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Pi1	0.53	0.43	0.19	0.85	0.13		+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+
Pi2	0.37	0.62	0.34	0.63	0.23	0.49		+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+
Pi3	0.67	0.90	0.66	0.93	0.58	0.88	0.20		+	+	+	+	+	+	+	+	+	+	+	+	-	+	+
Sw	0.59	0.54	0.36	0.78	0.35	0.39	0.57	0.80		+	+	-	+	+	+	+	+	+	+	+	-	+	+
Au2	0.54	0.54	0.44	0.67	0.42	0.45	0.40	0.62	0.52		+	+	+	+	+	+	+	+	+	+	+	+	+
Au1	0.25	0.54	0.23	0.48	0.28	0.39	0.38	0.65	0.27	0.49		+	+	+	+	+	+	+	+	+	+	+	+
Sv1	0.56	0.64	0.16	0.93	0.24	0.45	0.45	1.00	0.31	0.33	0.36		+	+	+	+	+	+	-	+	-	-	-
Sv2	0.08	0.85	0.43	-0.11	0.52	0.81	0.45	1.00	0.72	0.53	0.35	1.00		+	+	+	+	+	+	+	-	+	+
NB	0.71	0.92	0.70	0.94	0.61	0.91	0.41	1.00	0.82	0.67	0.70	1.00	1.00		+	+	+	+	+	+	+	+	+
NE	0.67	0.90	0.66	0.93	0.58	0.88	0.37	1.00	0.80	0.62	0.65	1.00	1.00	1.00		+	+	-	-	+	+	+	+
NR	0.29	0.78	0.48	0.44	0.47	0.70	0.33	0.62	0.70	0.60	0.44	0.74	0.33	0.60	0.57		+	+	-	+	+	+	+
NIn	0.70	0.91	0.70	0.93	0.61	0.90	0.41	0.94	0.82	0.66	0.69	0.99	0.98	0.48	0.43	0.59		+	-	+	+	+	+
NS	0.68	0.89	0.68	0.89	0.60	0.86	0.39	0.78	0.80	0.65	0.67	0.94	0.90	0.43	0.01	0.56	0.15		-	+	+	+	+
NIc	0.52	0.85	0.51	0.89	0.41	0.79	0.02	1.00	0.71	0.42	0.48	1.00	1.00	1.00	0.00	0.37	0.19	-0.30		+	-	-	-
It1	0.88	0.94	0.83	0.98	0.84	0.95	0.86	1.00	0.88	0.58	0.85	1.00	1.00	1.00	1.00	0.93	1.00	0.98	1.00		+	+	+
It2	0.78	0.89	0.68	0.97	0.73	0.88	0.65	1.00	0.79	0.14	0.71	1.00	1.00	1.00	1.00	0.86	0.99	0.97	1.00	1.00		-	-
It3	0.76	0.84	0.67	0.94	0.70	0.83	0.64	0.96	0.77	0.26	0.70	0.76	0.88	0.97	0.96	0.84	0.96	0.94	0.80	0.69	-0.90		+
Fr1	0.69	0.77	0.59	0.88	0.61	0.72	0.53	0.87	0.69	0.02	0.62	0.48	0.71	0.91	0.87	0.77	0.90	0.87	0.54	0.72	-0.38	0.02	

3.3.4 Demographic history

The τ values and corresponding p values for each pooled population under both the sudden demographic expansion model and the spatial expansion model are listed in Table 3.8.

Table 3.8 Mismatch distribution analysis of population demographic history as inferred from Arlequin. τ values are indicated for each population, each with its lower and upper bounds at the 95% confidence level under the spatial expansion model. Some of the populations are examined under conditions when either clade I or clade II/III is included only. Ri indicates raggedness index and Rp indicates raggedness p value.

	Demographic expansion				Spatial expansion					
	τ	SSD-p	Ri	Rp	τ	Low. τ	Upp. τ	SSD-p	Ri	Rp
Ir	0.00	0.00	0.56	0.96	**38.12**	29.59	45.50	0.09	0.56	0.13
Ir clade I	3.00	0.01	0.67	0.60	**14.21**	0.00	19.64	0.36	0.67	0.77
Ir clade III	3.00	0.05	0.89	0.87	**13.89**	0.00	170.50	0.10	0.89	0.85
Py	36.38	0.05	0.08	0.00	**32.19**	2.33	40.64	0.22	0.08	0.21
Py clade I	**3.16**	0.15	0.08	0.26	**2.83**	0.56	4.27	0.34	0.08	0.49
Py clade II	0.00	0.00	0.75	0.95	**6.67**	0.00	74.00	0.09	0.75	0.41
Pi	0.00	0.00	0.46	0.97	**38.25**	27.11	42.96	0.17	0.46	0.39
Pi clade I	**3.25**	0.06	0.49	0.40	**5.33**	0.00	8.36	0.58	0.49	0.59
Pi clade II	**3.00**	0.09	0.80	0.78	**6.67**	0.00	85.25	0.22	0.80	0.80
Au1	0.00	0.00	0.77	0.96	43.45	29.11	341.00	0.00	0.77	0.26
Au1 clade I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Au1 clade III	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Au2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Au2 clade I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sw	0.00	0.00	0.50	0.94	**19.62**	13.25	25.18	0.37	0.50	0.38
Sw clade I	0.00	0.00	0.56	0.94	**20.70**	0.00	170.50	0.13	0.56	0.54
Sw clade II	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sv1 (clade I)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sv2 (clade III)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A. norvegica (clade II and III)	0.77	0.01	0.16	0.00	**20.69**	0.57	26.18	0.24	0.16	0.50
Clade I pooled	21.12	0.10	0.05	0.00	**14.21**	3.48	23.78	0.21	0.05	0.36
Clade II/III pooled	22.76	0.01	0.14	0.00	20.13	12.04	25.93	0.00	0.14	0.14
All pooled	344.00	0.01	0.04	0.00	**33.93**	26.15	44.49	0.10	0.04	0.07

From Table 3.8 it is seen in most of the cases where at least one of the expansion models is accepted, the spatial expansion model has stronger support than the sudden demographic expansion model. In a few cases both models are accepted with SSD-p values greater than 0.05, and in these cases the spatial expansion model still has stronger support based on larger p values. While the raggedness index (Ri) value in each of the above cases is large, the raggedness p (Rp) value is not significant under the spatial expansion model, so that the expansion model is not rejected in these cases. In the other cases neither of the models is accepted for the populations with SSD-p values lower than 0.05, indicating no evidence in the current genetic structures for expansion events in their recent history.

In the cases where the spatial expansion model is accepted, relevant τ values (shown in bold in Table 3.7) are used to calculate the expansion time. The lower/upper bounds of the τ values are not used for this calculation but they do provide the information of the possible range of the expansion time. With the μ value estimated at $3.9\text{-}6.5\times10^{-6}$ per generation, the expansion time is estimated for each of the relevant populations as shown in Figure 3.9. Both Tajima's *D* and Fu's FS tests resulted p values greater than 0.05, which indicated that neutral assumption is not rejected.

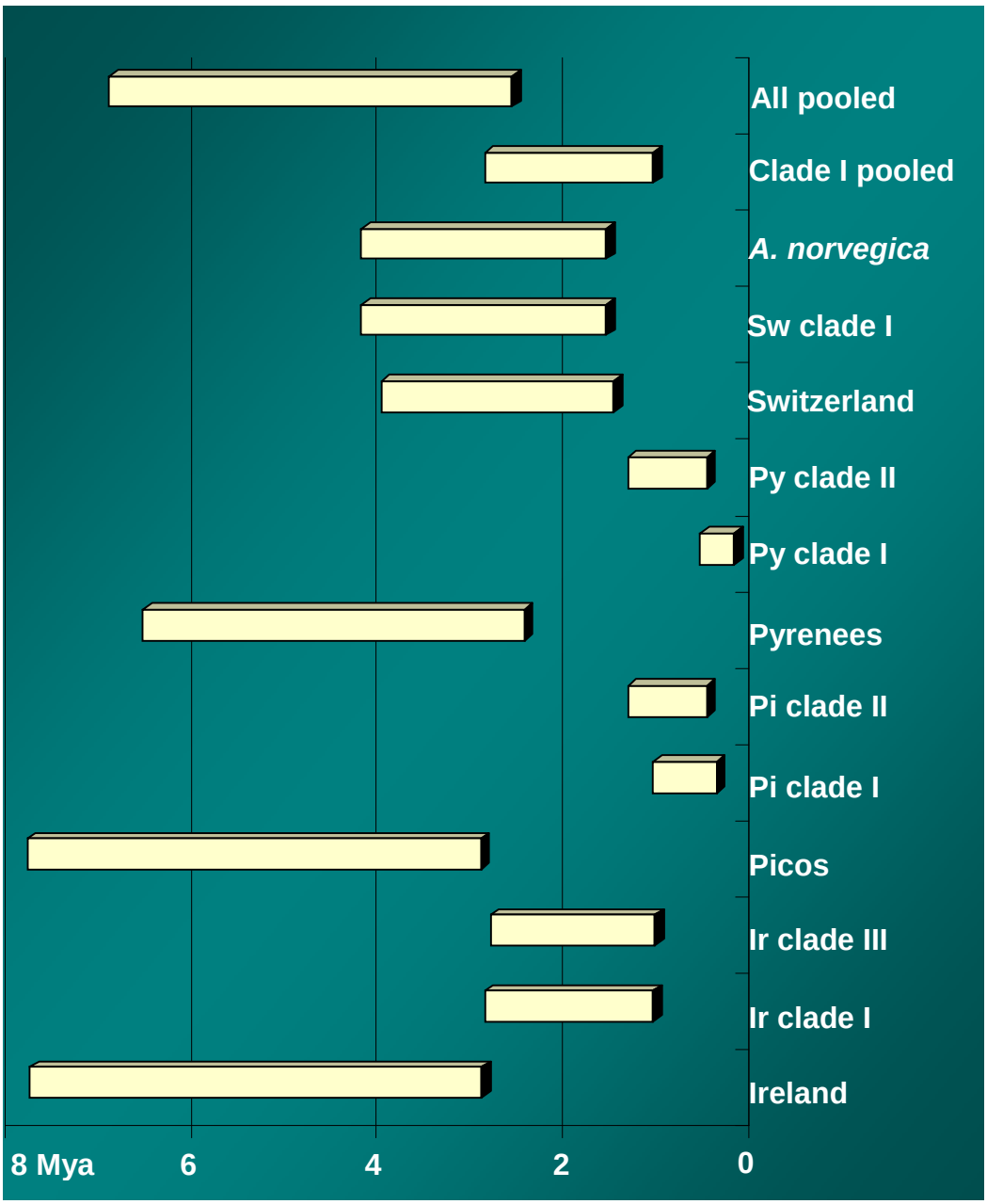


Figure 3.9 The estimated expansion time of the populations under the spatial expansion model. When the population contains haplotypes both in clade I and in clade II/III, the two clades are analyzed both separately and jointly. The time is shown in Mya (million years ago) and each bar shows the time range as the mutation rate μ ranges from 3.9 to 6.5×10^{-6} per generation.

The estimated expansion time is surprisingly early in the cases where clades I and II/III are jointly taken into account. The populations from

Ireland, Picos and Pyrenees all show an old expansion around 3-7 million years ago, and the pooled populations of the entire species shows similar expansion time.

When only clade I is taken into account, the pooled meta-population shows an expansion time around 2 million years ago. The Swiss population in this clade is indicated to have the earliest expansion while the Irish population expanded later. In contrast, the populations in Picos and Pyrenees are shown to have a much more recent expansion within 0.5-1 million years ago. Clade II/III shows a similar pattern. While the overall meta-population of clade II/III does not fit either of the expansion models, it is seen that the expansion of the Irish population is much earlier than that of the populations in Spain.

The *A. norvegica* populations are indicated to have experienced an earlier expansion event than other populations within clade II/III. However, both clade II and clade III are included in the *A. norvegica* populations while all other populations only contain one of the clades each. The oldest expansion date of *A. norvegica* populations in fact involved the early divergence between clade II and clade III.

Mismatch distribution for selected populations (Figure 3.10) show the multimodal distribution of pairwise nucleotide differences, which are plotted under the spatial expansion model.

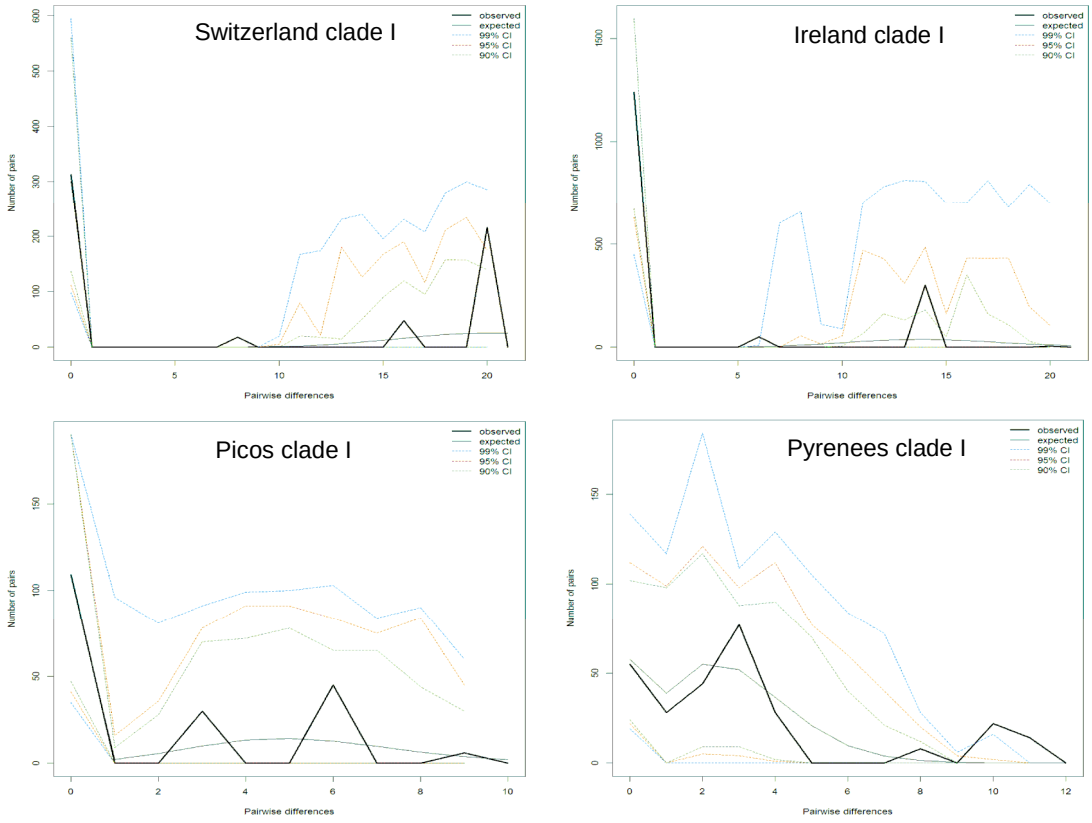


Figure 3.10 Result of mismatch distribution of the populations when only clade I is included. The x axis indicates the number of differences between pairs of haplotypes and the y axis indicates the occurrence of each number. The black solid line shows the observed distribution of pairwise nucleotide differences within the population, the grey curve is the expected distribution under the spatial expansion model, while the dashed lines with colours mark the credible intervals of the distribution at different significance levels.

With samples carrying clade I haplotypes only, it is shown in Figure 3.10

that the populations from Switzerland, Ireland, Picos and Pyrenees have different patterns of distribution. In the Swiss population, the highest occurrence is seen of 20 pairwise differences, while another two peaks are seen of 16 and 8 pairwise differences. In the Irish populations the highest occurrence is seen of 14 pairwise differences, while this number falls to 6 and 3 in Picos and Pyrenees. However, the Picos population have another peak at 3 and the Pyrenees have another peak at 10. The mismatch distribution curves for all the four localities are multimodal with high pairwise difference values, indicating a long history of all the populations. According the number of pairwise differences with the highest occurrence, the Swiss population experienced the earliest expansion and the Irish population experienced the second earliest expansion. It is noted that Switzerland and Pyrenees shared a peak at 8, Ireland and Picos shared a peak at 6 while Picos and Pyrenees shared a peak at 3, implying some common expansion events being shared by multiple populations in the history.

With samples carrying clade II or III haplotypes only, the mismatch distribution curves are also plotted for the above populations, except for the Swiss population which does not fit either of the expansion models. The results are shown in Figure 3.11, where the Irish populations with

clade III haplotypes only, the Picos and Pyrenees populations with clade II haplotypes only are included.

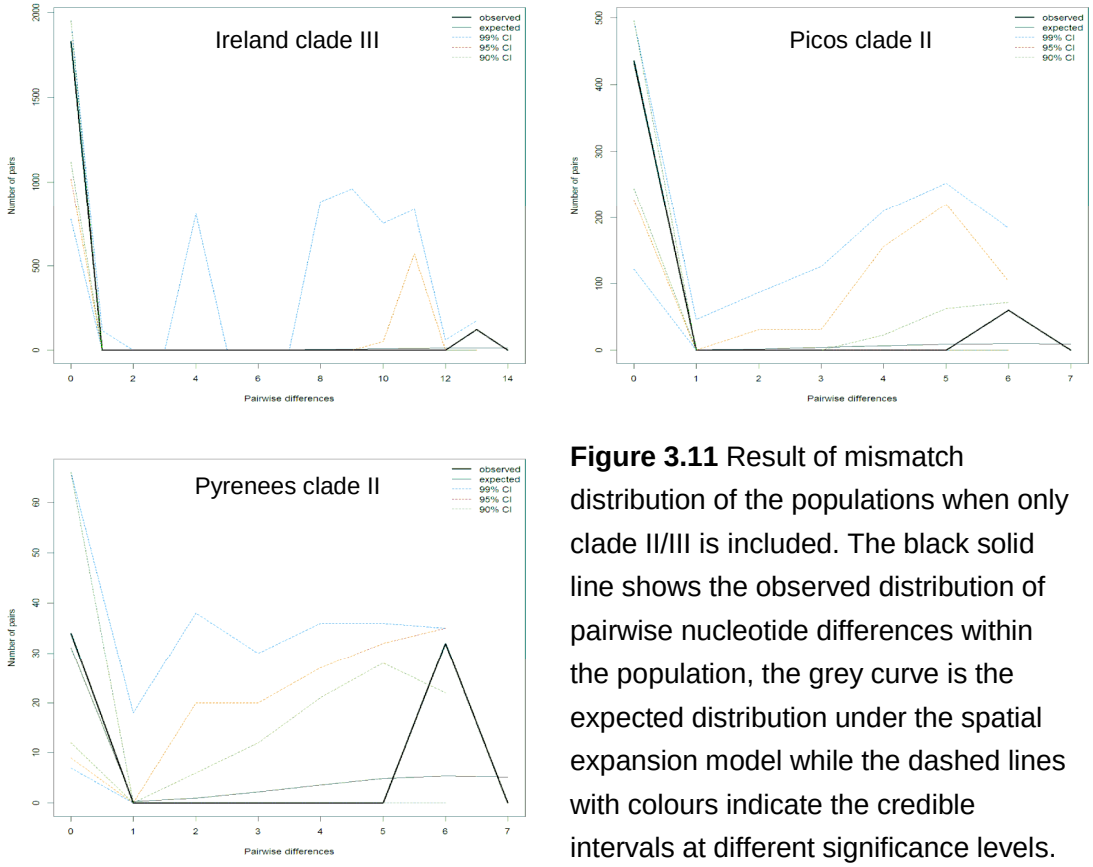


Figure 3.11 Result of mismatch distribution of the populations when only clade II/III is included. The black solid line shows the observed distribution of pairwise nucleotide differences within the population, the grey curve is the expected distribution under the spatial expansion model while the dashed lines with colours indicate the credible intervals at different significance levels.

It is seen in the Irish populations with clade III haplotypes only, the mismatch distribution has a similar pattern with that of the clade I carriers, with an occurrence peak at 13. This may suggest that the carriers of clade III and clade I haplotypes may have co-occupied Irish habitats during the same historical period. The clade II Picos and Pyrenees populations here again shared an occurrence peak at 6 pairwise differences, indicating a

common expansion event in the two places. However, in clade I the expansion event corresponding to the occurrence peak at 6 only happened in the Picos populations, while there was a more recent common expansion event corresponding to the occurrence peak at 3 shared by both places. This suggests that the historical migratory dynamics of clade I and clade II were not always the same in continental Europe. And again, the Irish populations are shown to have experienced a much older expansion than the Spanish populations.

3.4 Discussion

3.4.1 Phylogeny and subspecies status of the *Arenaria ciliata* Complex

Based on our result of the phylogeographic study with the previously recognized two species, *A. ciliata* and *A. norvegica* within the *A. ciliata* Complex, five distinct clades are revealed to compose the species complex. While the samples identified as *A. ciliata* fall into the five clades as a paraphyletic group, the *A. norvegica* samples also fall into two distinct clades. Neither of the two species forms a monophyletic group based on the data from their chloroplast genomes, however data from nuclear genomes are still needed to verify their taxonomical status. As the

divergence of clades IV and V from other clades are dated to at least 2-4 million years ago and their carriers are generally endemic to south side of Alps with only rare occurrence in Spain, they may well represent discrete taxa at or below species level but in process of speciation. There is no morphological evidence to help assign the two clades into any of the subspecies identified previously, neither does their distribution help as they do not correspond to the geographic distribution of any of the subspecies.

Based on the phylogeny of haplotypes in clades I, II and III, the subspecies status of *A. ciliata* subsp. *hibernica* and *A. ciliata* subsp. *pseudofrigida*, which were designated to the Irish populations and Svalbard population previously, are not supported by the genetic signature from chloroplast DNA in these populations. The populations presumed to be the two subspecies share haplotypes in two clades with those populations found from Switzerland and Austria.

While the phylogenetic analysis based on chloroplast DNA data suggests the division of these clades does not agree with the subspecies, further studies are need to clarify the extant lineages in the *A. ciliata* Complex. One possibility is to examine in more details populations from the

southern side of the Alps and also the Iberian Peninsula, to check if *A. ciliata* subsp. *moehringioides* and *A. ciliata* subsp. *bernensis* correspond to the genetic clades IV and V. The second possibility is to include the populations from Finland and Russia, which were identified as *A. ciliata* subsp. *pseudofrigida* in the floristic literature to see if they are genetically identical to the populations from Svalbard.

The *A. norvegica* samples are revealed to fall within two distinct lineages, each of which falls under a clade of *A. ciliata*. Thus the chloroplast data did not support the taxonomical classification of the species as a single lineage. Also, the fact that the English populations consist of the haplotype which is also found in multiple places in Scotland did not support the classification of *A. norvegica* subsp. *anglica*.

3.4.2 Phylogeography based on haplotypes in clade I and clade II/III

Overall, the divergence among the five clades is deeper than thought previously and the divergences between clades IV and V and other clades are close to the inter-specific level, as judged by taxon divergence rates within Caryophyllaceae as estimated by Valente *et al.* (2010), where the divergence time between sister species is estimated to be as recent as

1.0-2.0 Myr ago. Considering that the divergence between clade I and II/III is estimated 0.9-1.5 Myr ago, it is reasonable that they are treated as separate taxonomic units in the present study. Also because clade I and clade II/III haplotypes composed most of the populations of interest across Ireland and continental Europe, the present study is focused on the haplotypes within these clades.

The revealed coexistence of clade I and II/III in more than one place is uncommon compared to other similar phylogeographic studies (e.g. Bisconti *et al.* 2011; Gutiérrez-Rodríguez *et al.* 2011), where the genealogy among haplotypes is usually consistent with their distribution among populations. However similar situations have been seen in other Alpine plant species complexes (e.g. Dixon *et al.* 2009). Such disagreement between the genealogy and the distribution may indicate that historical migratory events were not always concurrent with the divergence of the observed haplotype clades. In order to interpret the phylogeographic processes, the intra-clade rather than inter-clade polymorphisms should probably be considered more informative.

3.4.2.1 Clade I

Within clade I, RTC07/09 is determined by the software TCS to be the most ancestral haplotype (probably because it is directly linked to the greatest number of haplotypes, 5), while RTC04/05 (with 4 linked haplotypes) is closer to the diverging node than RTC07/09. Both haplotypes can thus be considered as ancestral haplotypes that give rise to all other haplotypes in the clade. These haplotypes are found in the Picos de Europa, the northwest part of Alps and in the Pyrenees, where both RTC07/09 and RTC04/05 are found together along the Valle de Benasque, suggesting that these are the most ancient populations, possibly established during the time period when RTC04 and RTC07 diverged from each other (320 to 530 thousand years ago).

RTC06 and RTC13 are the majority haplotypes in the populations from west and east parts of the northern Alps, where they are found as endemic haplotypes. As they are more recent haplotypes diverged from RTC04/05 and not found in other populations, it is reasonable to speculate that the current Alps populations established later than the Iberian populations. Furthermore, the relationship between RTC06 and RTC13 as well as the relationship between RTC07 and RTC11 suggest the Austrian populations

on the east part of Alps established later than the Swiss population on the middle part of Alps. An eastward migration along the Alps mountains is thus a possibility, but not later than the period when RTC11 diverged from RTC07 (150-250 thousand years ago) or when RTC13 derived from RTC06 (90-150 thousand years ago), because the Austrian haplotypes appear to be both endemic and peripheral to the clade.

RTC01 is found both from the Picos de Europa and Ireland, indicating the dispersal of the clade from the Iberian Peninsula northward to Ireland after RTC01 diverged from RTC07/09. The endemic distribution of RTC03 may imply the origin of the haplotype after the establishment of the Irish populations. Thus the migration of the clade from Spain to Ireland should have occurred after the divergence between RTC01 and RTC07/09 but before the divergence between RTC01 and RTC03. However, the maximum likelihood phylogenetic tree linearized with molecular clock in Figure 3.2 does not differentiate the two divergence events in a similar order. It can only be inferred that the establishment of the Irish populations occurred between 150-250 thousand years ago and the present.

One interesting haplotype is RTC02, which is derived from RTC04/05.

This haplotype is found in three disjunct localities, including Ireland, Switzerland and Svalbard. It remains unclear how this haplotype might have dispersed to these distantly separated places. The position of RTC02 both in the haplotype network and in the linearized phylogenetic tree indicates an old origin of the haplotype, which is dated to 270-450 thousand years ago. This indicates an early occupation of the haplotype in these places. However, the lack of accumulated mutations to the RTC02 haplotype between the three populations implies that a more recent establishment event in these places is also possible. The contradiction remains unsolved.

3.4.2.2 Clade II and III

There are ten haplotypes in clade II and clade III jointly. It is noted that the haplotypes RTN01, 02 and 03 of *A. norvegica* fall within clade II, and RTN01 as the most common haplotype is derived from the hypothetical ancestral haplotype of clade II by only two mutational steps. The separation between the lineage that includes the Irish population of *A. norvegica* in the Burren and its closest relatives in the continent is thus estimated to be the time when RTN01 diverged from RTC17, 19, 20 and 21, which is dated to 250-410 thousand years ago.

In clade III, the most abundant haplotype is RTC14/16, which occurs in three distantly separated places, including Ireland, Svalbard and Austria. As the one-step mutant of RTC14/16, the haplotype RTC15 is found endemic to the populations in Ireland. The distribution pattern of RTC14/16 is similar to that of RTC02 in clade I, which is found from Ireland, Svalbard and Switzerland. As described above for clade I, the establishment of these populations should have occurred after RTC02 diverged from sister lineages, which is estimated to be 270-450 thousand years ago. Similarly within clade III, it is likely the populations in these places established after RTC14/16 had diverged from its co-ancestor with RTN04, which is dated to 310-520 thousand years ago. Thus based on the information of different lineages from both clade I and clade III, it seems possible that a single dispersal event involving haplotypes of both clades occurred some time between 270 and 520 thousand years ago and established the populations in Svalbard and part of the populations in Ireland and the Alps.

Possibly around the same time, another lineage in clade III represented solely by RTN04 settled in the Rum Island and survived the subsequent climate changes. Thus the *A. norvegica* population found in the Rum

Island is composed of two distinct lineages possibly via two historical colonization events, i.e. the settlement of RTN04 from 270-520 thousand years ago and that of RTN01 and RTN02 around 60-100 thousand years ago. There is a possibility that the Rum Island hosted the ancestral population of clade II and clade III among all the 13 localities on the map in Figure 3.6, however as the greatest diversity of clade II haplotypes is concentrated in Iberia, the supporting evidence for this Rum centre of origin is weak. An alternative hypothesis is the Rum population is composed of individuals that came to the place via multiple colonization events. It is unknown at this point whether this pattern is reflected in the nuclear data.

3.4.3 Co-occurring and co-migration of different clades

It is unexpected that distinct lineages of the same species are repeatedly found to coexist on multiple occasions. Carriers of clade I and clade II haplotypes are found to co-occur in the Picos de Europa, the Pyrenees and Switzerland, while clade I and clade III co-occur in Ireland and Svalbard. It is interesting to consider how they could have migrated together through such a long journey and then coexisted in the same habitats without mutually excluding each other.

The co-occurring distinct lineages in multiple places also provide an extraordinary scenario as there seems no correlation between the genealogy and the geographic distribution, which is unusual in phylogeographic process. The extraordinarily high genetic diversity could have been explained by multiple colonization events over recent biogeographic history, but in that case we should have seen more diversified clades than what have been revealed in this study. There might be some ecological mechanisms maintaining the two clades, either clade I with clade II or clade I with clade III, within the same population without mutual exclusion, which has functioned throughout the migratory history of the species. However, based on our demographic study above, the dynamics of the two clades do not appear to be entirely consistent with each other although there is some overlapping and agreement.

3.4.4 The structure of the populations

While all the revealed haplotypes in the five clades are included, 16 groups with the highest F_{CT} value were suggested by SAMOVA analysis, however the population groups in this case become dissolved without meaningful structure. Compared to this, K=8 enhanced the F_{CT} value

significantly compared to lower K values while dividing the populations into fewer groups. Both grouping plans as shown in Table 3.5 unexpectedly separate Ir2 from the other Irish populations, but a few interesting relationships among populations were indicated by the eight groups. Group 1 shows that the majority of *A. norvegica* populations (also the clade A-only populations) are closely related to the Picos populations of *A. ciliata*. Group 2 shows that the Rum population of *A. norvegica* has a close relationship with the Irish populations of *A. ciliata*. Groups 3 and 7 show the close relationship between the Svalbard populations and the north Alps populations which are geographically distant from each other. The latter three cases indicate longitudinal rather than latitudinal gene flow, suggestsing that North/ South migration routes predominated in the distribution of these haplotypes. This is in accordance with the recurring cycles of southward expanding and northward retreating of glaciation throughout the Quaternary (Hewitt 2000). Barriers may have prevented latitudinal gene flow from one population to another eastward or westward, among which the biggest barrier may have been the low-altitude land area between the Pyrenees and the Alps.

By looking jointly into Table 3.6 and Figure 3.8, K=5 for clade I and K=5

for clade II/III are sufficient to assign most of the genetic difference at the inter-group level. They have also revealed the difference in the grouping plan of the populations between clade I and clade II/III. Considering that clade I includes more recently diverged haplotypes than clade II and III, the disagreement of the grouping plan between the clades may indicate dispersal events at different times. Based on the data from clade I, the Picos population Pi2 is genetically similar to the Irish populations while genetic similarity is also found among Picos (Pi1), Pyrenees (Py) and southeast Alps (Au2), indicating a relatively recent genetic exchange across these places. Based on the data from clade II and III, the Picos populations are assigned within a separate group, while genetic similarity is seen between the populations from Pyrenees and Switzerland, indicating the isolation between Picos and Pyrenees but abundant genetic exchange between Pyrenees and Alps in a more remote history. The assignment of Au1 from northeast Alps and Sv2 from Svalbard within the same group as the Irish populations when K=5 based on clade II/III data implies the possibility of very ancient gene flow among these places. However, the highest F_{CT} values obtained when K=7 for clade I and K=12 for clade II/III indicate the possibility that all the populations in the present study have been isolated for a long period, and that the population structuring may fail to reveal the complex history of their establishment.

3.4.5 The origin of Irish populations and the possibility of Irish refugia

The four populations recorded as *A. ciliata* found in Ben Bulben, northwest Ireland are composed of three lineages, two in clade I and one in clade II, which may have migrated to Ireland via two colonization events. Based on the above analysis, it is inferred that RTC02 in clade I together with RTC14 in clade II probably came to Ireland from the continental Europe from some time between 270 and 520 thousand years ago, and later RTC01 in clade I came to Ireland from 150-250 thousand years ago. Two endemic haplotypes have emerged after the population had settled, including RTC03 derived from RTC01 and RTC15 derived from RTC14.

While it is unclear where the first colonization of the species came from exactly, the second colonization was likely to come from the Picos area in the north part of the Iberian Peninsula, because the two places share the common haplotype RTC01. As clade I finds no occurrence in Britain, the hypothesis is thus supported that some of the Irish plants may have come directly from the Iberian Peninsula via a land bridge without passing through the island of Britain during the postglacial migration (Wingfield

1995; Kelleher *et al.* 2004; Mitchell 2006).

The *A. norvegica* population found in the Burren, middle-west Ireland is revealed to host the oldest haplotype, RTN01, in the clade A of the species and thus may constitute one of the oldest populations in the clade, while other populations of the clade may have established later as they contain a derivative haplotype RTN02 as a one-step mutant of RTN01. However, an alternative hypothesis is that the Scottish populations are older than the Burren population because they also include RTN01 together with RTN02. In such a scenario the haplotype RTN01 may have migrated to Ireland at a later time, possibly in the postglacial period, but either never migrated south to England, or became extinct there.. This second hypothesis seems more likely as there is no evidence for a glacial refugium in the Burren. The settlement of the earliest population of this clade is inferred as early as 250-410 thousand years ago while the lineage diverged from its co-ancestor with the *A. ciliata* lineages in continental Europe. The populations of *A. norvegica* found in Iceland and Shetland are inferred to have established since 60-100 thousand years ago after RTN02 diverged from RTN01, so there is some evidence of a recent geographic expansion in the islands of the Northeast Atlantic for *A. norvegica*.

Thus in clade II the Irish population of *A. norvegica* is inferred to have a shorter history than the British populations, but the *A. ciliata* populations in Ireland and the *A. norvegica* population in Rum are almost equally ancient when only clade III is considered. Thus the hypothesis that part of the Irish flora may have come through the island of Britain (Webb 1983; Wingfield 1995) is supported by the data from clades II and III.

However, the date of the migratory events is inferred to be much older than thought before. It is usually believed the current biodiversity was shaped by the postglacial migration after Pleistocene ended 10-12 thousand years ago, while our data suggested multiple colonization events occurred since as early as 500 thousand years ago, during the four or five glacial cycles that occured in the late Pleistocene (Hewitt 2000). This indicates that the several places including Ireland, Iceland, the islands of Rum and Shetland in Scotland, the Svalbard islands and the Scandinavia Peninsula all hosted refugia during the time when glaciation dominated the northern areas of Europe. Specifically, the western surface of Ben Bulben as a refugia proposed by Synge and Wright (1969) is supported by our result that the *A. ciliata* populations there may have survived more than one glacial period.

Based on both phylogenetic analysis and demographic analysis, the Irish populations of *A. ciliata* in Ben Bulben are inferred to have a much older history than thought before. While sharing clade I haplotypes with other populations from Picos, Pyrenees, Switzerland, Austria and Svalbard, the Irish populations also contain haplotypes in clade III, which are only shared by Austria and Svalbard (and Rum Island harbouring the RTN04 haplotype of *A. norvegica*). Clade III is closely related to clade II but has diverged from the latter since 0.6-1.0 million years ago, while Ireland is one of the only three harbours of the clade. The clade in Ireland may also hold the key in understanding the origin of the unique haplotype of *A. norvegica* (RTN04) endemic to the Rum Island in Scotland. Moreover, one of the Irish populations includes an endemic haplotype in clade III (RTC15), which indicates the long history of the population in Ireland. Considering the possible Irish refugia proposed by Synge and Wright (1969) and the fact that more northern areas including the Rum Island and Svalbard also harbour haplotypes in clade III, there may have been a larger area of refugia including these islands, which helped the plants survive multiple glacial events during the last several million years, with the greatest surviving genetic diversity preserved at Ben Bulben.

3.5 Conclusion

With genetic information from chloroplast DNA, multiple lineages have been revealed in the previously recognized species *A. ciliata* and *A. norvegica*. The deep divergence between lineages is unexpected, while different lineages are found to coexist within the same populations repeatedly. The high genetic diversity and the demographic patterns indicate an early origin and unique history of the Irish populations, as they contain both haplotypes shared by Spanish populations and those shared by the Alps populations. The establishment of the Irish populations of *A. ciliata* may have been as early as two million years ago and *in situ* survival through multiple cycles of glacial oscillations has been vital to maintain the lineages in Ireland.

Chapter 4

Phylogeography, population genetics and demographic history analysis of *Minuartia recurva* in Europe

4.1 Introduction

4.1.1 Background

Based on the chloroplast DNA data obtained for the species *Minuartia recurva* as described in Chapter 2, a phylogeographic study has been conducted on the populations of the species.

The carnation species *Minuartia recurva* (All.) Schinz & Thell., commonly called curved sandwort, is a diploid (2n=30), tufted perennial herb with woody basal stems growing in non-calcareous rocks. It is an alpine plant found widely across mountainous areas in south Europe, from the Iberian Peninsula across the Alps to the Balkans and Turkey, and also on two separate sites in Ireland. Two subspecies have been described within the species, including *M. recurva* subsp. *recurva* and *M. recurva* subsp. *condensata* (or subsp. *juressi* synonymously) which differ in morphology, but subspecies with intermediate characteristics have also been reported in the Iberian Peninsula, south France and the southeast Alps. The first subspecies is found throughout the species distribution except in Sicily, while the second subspecies is found in Sicily, south and central Italy and south part of Balkans (Jalas & Suominen 1983; Tutin *et*

al. 1993). In Ireland the plant has been recorded from two areas: the Caha mountains on border between Co. Kerry and Co. Cork, and a second, only recently discovered site in the Comeragh Mountains in Co. Waterford (Green 2007). The Irish sites for this species are disjunct within Europe, and represent by far the most isolated stations for plants of these species.

4.1.2 Objectives

With the overall aim of improving our understanding of the potential historical processes that have lead to the current biodiversity of the species, three objectives have been set for this section of the study:

1. To complete a phylogeographic analysis of *M. recurva* across the sampled European localities based on phylogenetic and haplotype network analysis of all the revealed haplotypes.

2. To evaluate the population genetic structure of *M. recurva* based on the haplotype polymorphisms within and across the populations from different localities.

3. To investigate the likely demographic history of the populations of *M.*

recurva based on molecular dating with the haplotype genealogy and the frequency data.

4.2 Materials and Methods

4.2.1 Haplotype phylogeny in *M. recurva*

In the present study, chloroplast DNA data were obtained from 250 individual samples in 13 populations across the European distribution of the species, and these data are used for phylogeographic analysis in this chapter to understand the biodiversity of the species. The locations and sampling sizes of the populations are listed in Table 4.1.

With the method described in Chapter 2, the samples were designated into the revealed haplotypes as listed in Table 4.2. Based on this information it is possible to conduct phylogeographic, population genetic and demographic analysis with the sampled populations of *M. recurva*.

Table 4.1 Localities of the sampled populations of *Minuartia recurva* in the study.

Species and Localities	Population code	Latitude/ Longitude	Sample size
Comeragh Mountains, Co. Waterford, Ireland	MR1	N52º 14.163' W07º 31.202'	20
Caha Mountains, Co. Kerry, Ireland	MR2	N51º 44.368' W09º 43.161'	30
	MR3	N51º 44.592' W09º 42.314'	30
Ligurische Alpen, Piemonte, Italy	MR4	N44º 09.734' E07º 47.082'	10
Cottische Alpen, Piemonte, Italy	MR5	N44º 33.086' E07º 07.135'	3
Mountain Kopaonik, Serbia	MR6	N43º 16.375' E20º 49.132'	26
Pico El Vigia, Cantabria, Spain	MR7	N43º 03.341' W04º 44.651'	35
Huesca, valle de Benasque, Pyrenees, Spain	MR8	N42º 32.936' E00º 33.159'	24
	MR10		
Laguna Negra, Soria, Spain	MR11	N41º 59.763' W02º 50.915'	13
Kosovo	MR12	-- --	5
Gspon, Staldenried, Switzerland	MR13	N46º 13.906' E07º 55.300'	27
	MR14	** **	8
	MR15	N46º 13.381' E07º 54.945'	19

-- Geographic information unavailable; ** the sampling site of MR14 is between MR13 and MR15.

Table 4.2 Distribution of the identified haplotypes among the studied populations.

Individuals	rps16	trnT-trnL	composite
MR1.1-20	rpsM01	TLM01	RTM01
MR8.1-11, 13-18; MR10.2-8			
MR7.1-35	rpsM01	TLM06	RTM07
MR2.3, MR2.x*	rpsM02	TLM02	RTM02
MR4.1-10	rpsM02	TLM07	RTM08
MR5.3-5			
MR13.1-27			
MR15.1-19			
MR2.21, MR2.x*	rpsM02	TLM08	RTM09
MR3.1-30			
MR11.1-13			
MR14.1-8			
MR12.2	rpsM03	TLM03	RTM03
MR12.1, 3-5	rpsM04	TLM04	RTM04
MR6.1, 3, 5, 7, 11, 14, 15, 16, 19, 20, 22, 26	rpsM04	TLM08	RTM05
MR6.2, 4, 6, 8-10, 12, 13, 17, 18, 21, 23-25	rpsM04	TLM05	RTM06

* The haplotype identities of part of MR2 samples are undetermined between RTM02 and RTM09

Based on the HRM analysis with two chloroplast DNA loci as described in Chapter 2, four haplotypes of rps16 and nine haplotypes of trnT-trnL were revealed from the sampled populations of *M. recurva*, comprising nine concatenated haplotypes. The sequences of the haplotypes were checked and aligned within BioEdit 7.1.3 (Hall 1999) using the function of Clustal*W* (Thompson *et al.* 1994). The haplotypes of *M. recurva* were named as RTM01-09, which is consistent with the description in Chapter 2. An extra haplotype of an individual from Index Seminum of the species, recorded as MR0178A, was also included for phylogenetic analysis. The rps16 sequences of all the *M. recurva* haplotypes were aligned with a length of 696bp and the trnT-trnL sequences were aligned into 537bp, totalling 1233bp when the two aligned sequences were concatenated. With the sequence of *M. verna* included as outgroup the concatenated haplotypes were aligned into 1279bp. Variance in simple sequence repeats (SSRs or microsatellites) was removed for phylogenetic analysis as described for *A. ciliata* and *A. norvegica* in Chapter 3, on the basis that variation in SSRs is considered to be evolutionarily labile and generates homoplasious characters that potentially provide misleading information (Small *et al.* 1998). Many researchers omit SSRs from phylogeographic analyses for this reason (Mast *et al.* 2001; Guggisberg *et al.* 2006; Garcia *et al.* 2011). With polymorphic SSRs removed, the

concatenated sequences were aligned into 1275bp.

Prior to analysis, the best-fit evolutionary model of nucleotide substitutions was selected with the software jModeltest (Posada 2008) from the 88 candidate models. While the -lnL values suggested the GTR+G model to be the best fit model, the TVM+G model was selected according to Akaike's information criterion, with the AIC score of 4535.5048 (Akaike 1973). However, the GTR+G model was finally determined to be used as it was suggested the second best model by its AIC score (4537.5029). Under the GTR+G model for the concatenated rps16 and trnT-trnL sequences, the following parameters were calculated: -lnL=2245.7514; base frequencies: freqA = 0.3802, freqC = 0.1221, freqG = 0.1618, freqT = 0.3360; gamma distribution shape = 0.1550; substitution rates of different categories: R[AC] = 2.3791, R[AG] = 1.4076, R[AT] = 0.4582, R[CG] = 0.7506, R[CT] = 1.3835, R[GT] = 1.0000.

The phylogenetic relationship among the concatenated rps16+trnT-trnL haplotypes was inferred using the maximum-likelihood method within the software phyML 3.0 (Guindon *et al.* 2009), with the GTR+G model and the substitution rates suggested by jModeltest. The default settings were

used for other parameters. A bootstrap test with 1000 replicates was used to provide the support values for the clustering of branches (Felsenstein 1985). The linearized maximum likelihood phylogenetic tree was constructed within the software MEGA 5.05 (Tamura *et al.* 2011) based on GTR+G model with six discrete Gamma categories for Gamma distributed substitution rates among sites.

The statistical parsimony network (Templeton *et al.* 1992) of the identified haplotypes was inferred with the software TCS 1.21 (Clement *et al.* 2000). Gaps are treated as a fifth state in order that both nucleotide substitutions and indels are taken into account. Each indel of multiple contiguous nucleotides is treated as a single mutation event (except the indels in SSRs are ignored for the reason described above), as is normally done in phylogeographic analysis (Jakob & Blattner 2006; Sosa *et al.* 2009). The percentage connection limit was set above 95% by default.

4.2.2 Population genetic diversity

For population genetic analysis, the two sample sites from Caha Mountains in southwest Ireland (MR2 and MR3) were merged as a single population as they were collected on two moutain tops less than 1km

apart from each other. Also, the three sites from Switzerland (MR13, 14 and 15) were considered as a single population as they were from adjacent sites less than 1.5km apart. Considering that MR8 and MR10 were sampled at the same site so that they were in fact from the same population, in total ten populations were defined from different localities. The distribution information of the detected haplotypes among the ten populations was then used to analyze the genetic diversity with the software *Arlequin* 3.5.1.3 (Excoffier & Lischer 2010).

Calculation of standard genetic diversity indices and pairwise F_{ST} values among the populations were conducted in *Arlequin* 3.5.1.3 to evaluate the genetic variation within and among populations and groups. Besides calculation of pairwise F_{ST} values, the exact test of population differentiation (Raymond & Rousset 1995) was also carried out using *Arlequin*, with 100,000 steps of Markov chain and 10,000 dememorization steps.

In order to carry out an analysis of molecular variance (AMOVA) (Excoffier *et al.* 1992), a spatial analysis of molecular variance with the software SAMOVA (Dupanloup *et al.* 2002) was performed to explore possible natural grouping scenarios (without any *a priori* bias).

Geographic location information of each population was inputted to the software, and the K value that determined the potential number of groups in each permutation test was set from 2 to 6 to generate fixation index values representing inter-group variation (F_{CT}). It is noted that geographic information is unavailable for the population MR12 from Kosovo, thus its location was recorded as E20°45' and N42°45' (due south to the Serbian population MR6) to ease the SAMOVA analysis. The K number with the highest significant value of F_{CT} would be considered an optimal number of groups. The permutation number was set at 1000 and pairwise difference was used to compute inter-population genetic distance for AMOVA. After the AMOVA analysis, the values of F_{CT}, F_{SC} and F_{ST} were evaluated to assess the genetic variation at different spatial levels.

4.2.3 Demographic history

As described in Chapter 3 (section 3.2.3), spatial and/or demographic expansion of populations can leave distinctive patterns of genetic diversity among individuals within population. It is worthwhile to test for these patterns as they can help describe more clearly the most likely history for each population. A mismatch distribution analysis (Harpending 1994) was thus carried out to investigate if the *M. recurva*

populations experienced historical expansion events and, if so, when the expansion events occurred. Because most of the populations were revealed to consist of single haplotypes, it was unlikely that mismatch analysis could be reliably applied with each population. Thus the populations from different localities were pooled together as a single meta-population to investigate the expansion patterns that may be evident for the entire European distribution of the species.

The mismatch distribution test with the frequency distribution of pairwise nucleotide difference between individuals was performed with the software *Arlequin* 3.5.1.3 (Excoffier & Lischer 2010). Parameters in both models of demographic expansion (Rogers & Harpending 1992) and spatial expansion (Excoffier 2004) were estimated, the pairwise difference was used for molecular distance and 1000 bootstrap replicates were used for the mismatch distribution analysis. The significance (p value) for the sum of squared deviation (SSD) is used to judge if the expansion model is accepted or rejected. If the pooled populations was revealed to have been subjected to expansion, the expansion time was estimated from the moment estimator τ via the formula $\tau=2\mu t$, where t is the number of generations after the historical expansion up to the present and μ is the mutation rate of the whole DNA locus used for study . The

per nucleotide substitution rate within the *M. recurva* lineage was estimated to be 2.9-4.8×10^{-9} per nucleotide per year (as per equivalent calculations in Chapter 3, section 3.3.1), and as the DNA locus was aligned into 1229bp for the analysis, the value of μ was estimated as 3.6-5.9×10^{-6} per year. Also because the above test requires the DNA locus used to be evolutionarily neutral, Tajima's *D* (Tajima 1989) and Fu's Fs (Fu 1996) were also calculated in each case to see if the locus was under selection.

4.3 Results

4.3.1 Haplotype phylogeny and phylogeography in *M. recurva*

The inferred phylogeny among the haplotypes is illustrated in Figure 4.1 and Figure 4.2. In Figure 4.1 the branch lengths are shown to represent the accumulated substitutions of each haplotype, and in Figure 4.2 the phylogenetic tree is linearized to show the relative time of divergence between different clades.

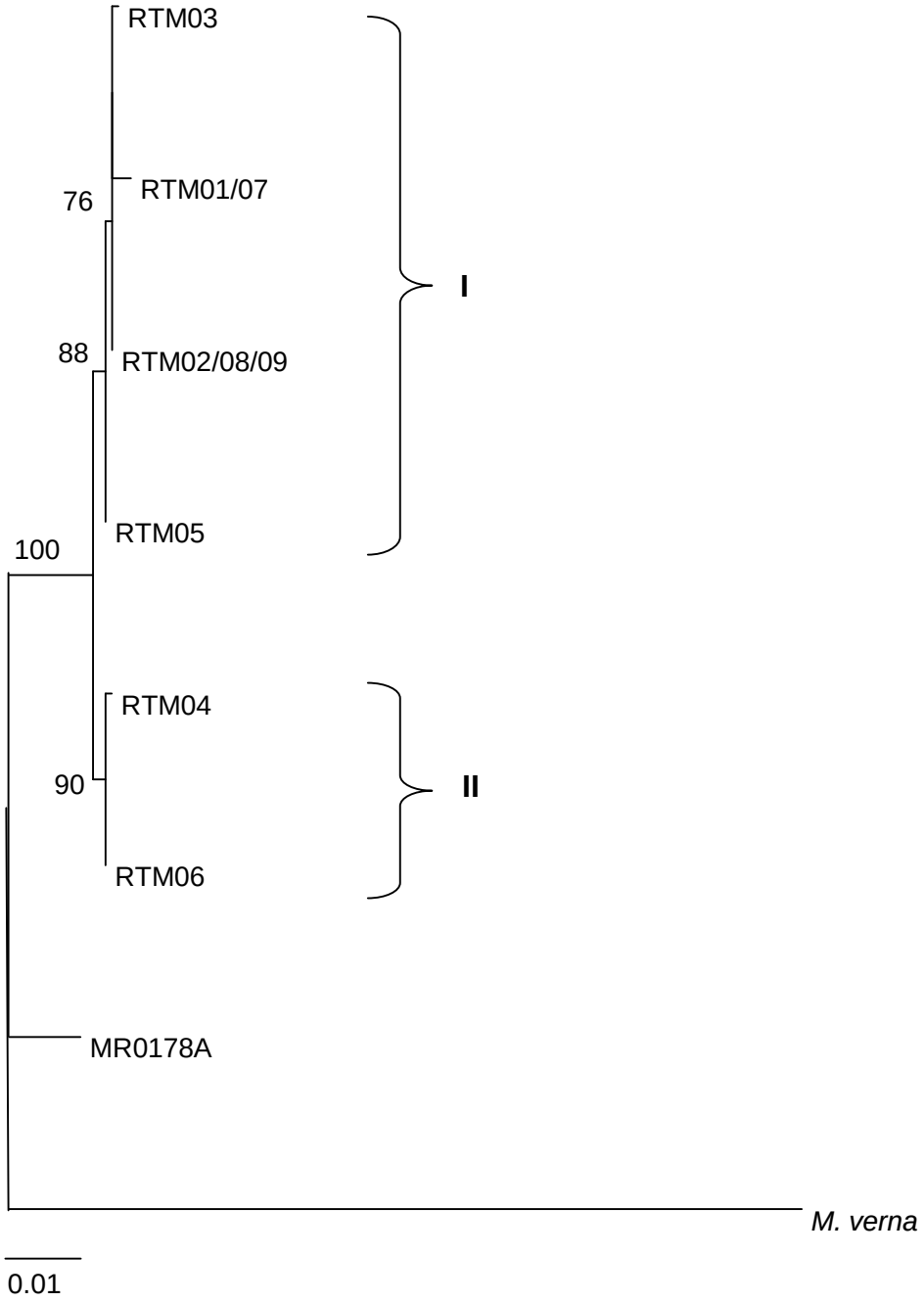


Figure 4.1 The inferred phylogenetic tree based on concatenated rps16+trnTL haplotypes (1275bp aligned after indels in SSRs are removed) using the maximum likelihood method with branch lengths showing the substitution rate. The numbers beside the branches are support values based on bootstrap analysis with 1000 replicates (only those >50% are shown). The sequence of *M. verna* is used as outgroup to root the tree. An extra voucher of *M. recurva* recorded as MR0178A was also included, the taxonomical status of which is undetermined.

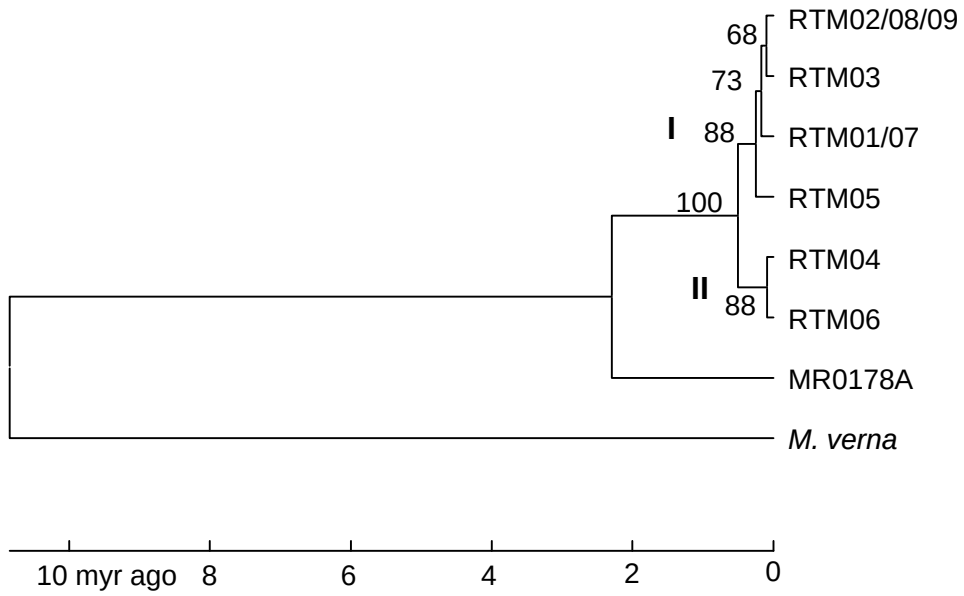


Figure 4.2 Linearized Maximum likelihood tree based on concatenated rps16+trnTL haplotypes (1275bp aligned after indels in SSRs are removed) with support values based on bootstrap analysis with 1000 replicates (only numbers greater than 50% are shown). The scale line below the tree shows the lower estimation of divergence time. The sequence of *M. verna* is used as outgroup to root the tree. An extra voucher of *M. recurva* recorded as MR0178A was also included, the taxonomical status of which is undetermined.

Both Figure 4.1 and 4.2 show the same topology of the phylogeny among the detected haplotypes, except that the clustering of RTM02/08/09 and RTM03 in Figure 4.2 (supported with 68% confidence) is rejected in Figure 4.1. It is seen in both figures that the *M. recurva* haplotypes fall into two different clades. One of the clades (I) includes RTM04 and RTM06 while in the other clade (II) RTC05 diverges from the cluster of all the rest haplotypes. The agreed clusters are supported by 73-100%

confidence based on bootstrap test.

Based on the fossil evidence from Valente *el al.* (2010) (Appendix Figure 1), the divergence between *Minuartia recurva* and *M. verna* was dated to 11-18 million years (Myr) ago and the mutation rate of chloroplast genome in the studied taxa was estimated at $2.9\text{-}4.8\times10^{-9}$ per nucleotide per year assuming constant mutation rate of the sequences under investigation. From the linearized tree in Figure 4.2, it is thus estimated the divergence of clade I from clade II occurred 0.51-0.85 Myr ago, the divergence of RTM05 from other haplotypes in clade I occurred 0.24-0.40 Myr ago, and the divergence among RTM01/07, RTM02/08/09 and RTM03 occurred 0.17-0.28 Myr ago.

The statistical parsimony network of the haplotypes is generated from TCS 1.21 and then redrawn for clarity. The geographic distribution of the haplotypes is illustrated on the map. The genealogical network among the haplotypes and their geographic distribution are shown in Figure 4.3

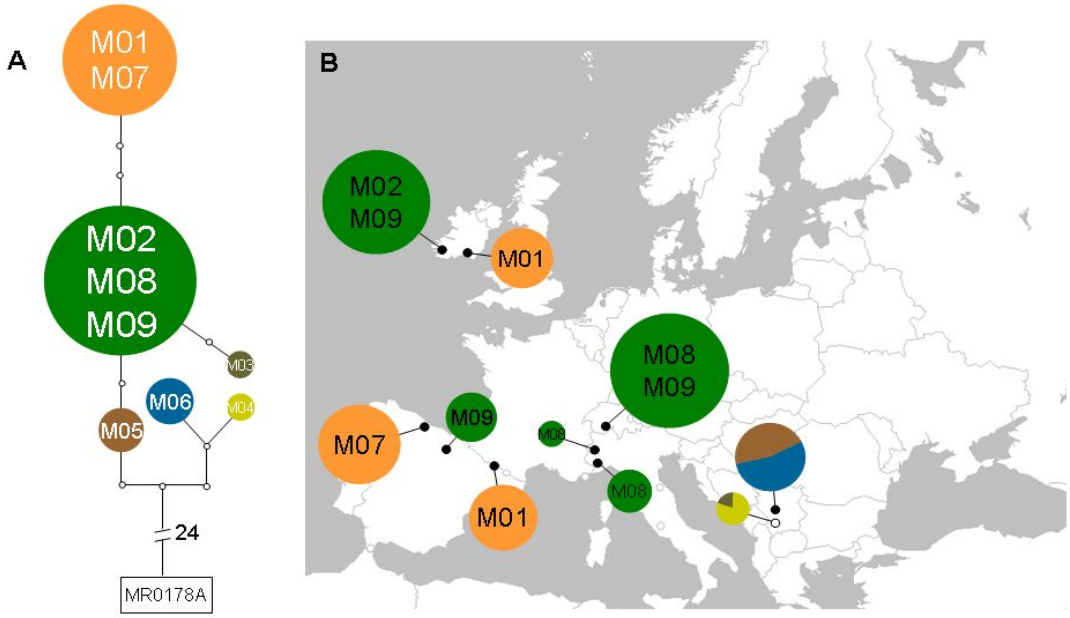


Figure 4.3 The statistical parsimony network among the haplotypes in group I as inferred from TCS (A) and the geographic distribution of the haplotypes (B).

A. The solid line between each pair of circles indicates one mutational step. Each circle with text indicates an extant haplotype, e.g. M05 represents the haplotype RTM05, the area of which indicates the number of individuals sharing this haplotype. Haplotypes RTM01 and 07 vary only in SSRs and are treated as a single haplotype in all analyses. The small circles without text indicate hypothetical intermediate haplotypes. Haplotypes are shown in different colours.

B. On the right panel, each pie represents the samples from each locality, as described in Table 4.1. The small dots indicate the sampling sites. The proportion of different haplotypes within each locality is shown by different colours in the pie, in accordance with the colour coding in panel A. It is noted that the exact geographic location of the Kosovo population is unknown so it is shown by a small circle instead of black dot.

In Figure 4.3, each haplotype is indicated by a circle with text, i.e. M01 for the haplotype RTM01, while the small circles without text indicate hypothetical intermediate haplotypes. All haplotypes as nodes are connected via straight lines showing their relationship. Each node is

different from its nearest neighbour by one mutational step. As variation in SSRs was not included in the analysis, RTM01 and RTM07 are considered as the same haplotype, and the same situation applies to the merge of RTM02, RTM08 and RTM09.

Based on the haplotype network in Figure 4.3, the haplotypes fall into two clades in accordance with the phylogenetic tree in Figure 4.1 and 4.2. RTM02/08/09 was determined by the TCS programme to be the ancestral haplotype. However, the connection between RTM01-06 and MR0178A as an out group suggests that RTM04, 05 and 06 are more ancestral haplotypes.

The distribution of the haplotypes RTM01-06 is illustrated on the map of Europe as shown in Figure 4.3B. Populations from the same locality as defined in Table 4.1 are pooled as a single geographic unit shown by the corresponding circle on the map. It is noted that the geographic information is unavailable for population MR12 from Kosovo, and its location is shown by a circle in Figure 4.3B instead of a black dot, indicating its undetermined location.

It is seen that RTM02/08/09 is the most widely distributed haplotype,

which is found from south west Ireland, the north side of Spain, the southwest side and northern part of Alps. The second common haplotype is RTM01/07, which is found in south Ireland, the north side of Spain and the Pyrenees. The two haplotypes were not found to co-occur in any single population, but their distribution is mixed within the larger geographic scale. Compared to the two above haplotypes, RTM03 and 04 are found restricted to Kosovo while RTM05 and 06 restricted to Serbia.

4.3.2 Population genetic diversity and structure

The average gene diversity (h) and nucleotide diversity (π) within each population are listed in Table 4.3 to show the genetic diversity at the intra-population level.

Table 4.3 The gene diversity (h) and nucleotide diversity (π) for each population to show intra-population genetic variation.

Population code	Gene diversity (h)	Nucleotide diversity (π)
MR1	0.0000	0.0000
MR2 and MR3	0.0000	0.0000
MR4	0.0000	0.0000
MR5	0.0000	0.0000
MR6	0.7996	13.9569
MR7	0.0000	0.0000
MR8 and MR10	0.0000	0.0000
MR11	0.0000	0.0000
MR12	0.4965	12.0000
MR13-15	0.0000	0.0000

It is suggested by both nucleotide diversity (π) and gene diversity (h) that the population MR6 from Serbia and MR12 from Kosovo showed the highest genetic diversity in terms of haplotype polymorphisms, simply because each of the two populations harbors two haplotypes while each of the other populations is composed of one haplotype.

The result of SAMOVA based on all detected haplotypes is shown in Figure 4.4, where the values of fixation indices are plotted against the K numbers. The p value for every fixation index was less than 0.01 so that all the values in Figure 4.4 are significant.

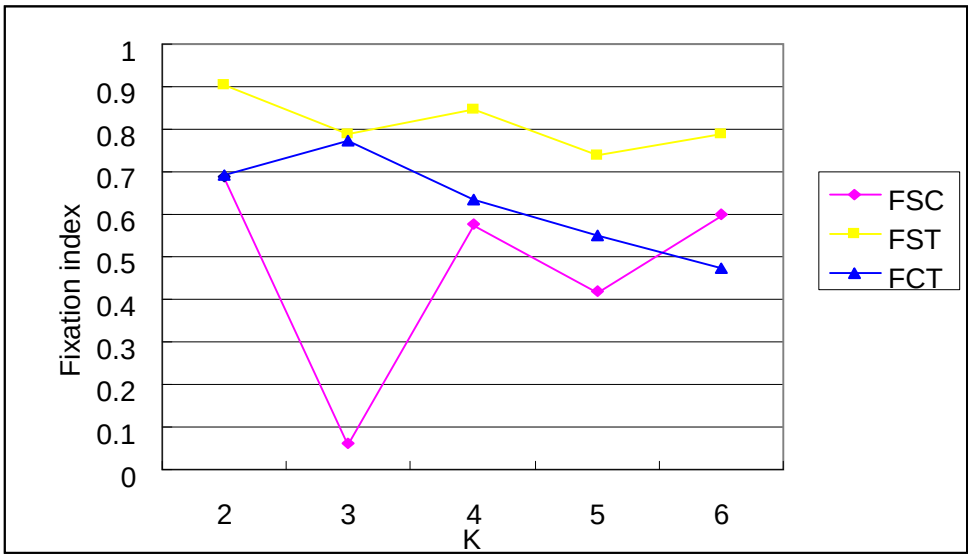


Figure 4.4 Fixation index values plotted against the K numbers of groups, as calculated in SAMOVA. The F_{CT} values (blue) indicate genetic variation among groups. The F_{SC} values (pink) indicate genetic variation among populations within groups. The F_{ST} values (yellow) indicate genetic variation among populations across groups. The K value indicates the number of groups into which the populations are divided.

As illustrated in Figure 4.4, the F_{CT} value increases immediately from 0.69 to 0.77 when K increases from two to three, and then begins to decrease when K is greater than three. The structuring of populations into three groups is thus suggested to be optimal by SAMOVA analysis, which is shown in Table 4.4. When the populations were manually grouped by their genetic composition or by their geographic distribution, they were assigned into four groups, which are also shown in Table 4.4 (on next page).

Table 4.4 Grouping plans based on SAMOVA when K=3 and K=4 and grouping plans manually based on genetics and on geography when K=4.

Group			
K=3 by SAMOVA	K=4 by SAMOVA	K=4 by genetics	K=4 by geography
Group 1	Group 1	Group 1	Group 1
MR1	MR1	MR1	MR1
MR7	MR7	MR7	MR2, 3
MR8, 10	MR2, 3	MR8, 10	
	MR4		
Group 2	MR8, 10	Group 2	Group 2
	MR11		
MR2, 3	MR13-15	MR2, 3	MR7
MR4		MR4	MR8, 10
MR5		MR5	MR11
MR11		MR11	
MR13-15		MR13-15	Group 3
	Group 2		
			MR4
	MR5		MR5
			MR13-15
Group 3	Group 3	Group 3	Group 4
MR6	MR6	MR6	MR6
MR12			MR12
	Group 4	Group 4	
	MR12	MR12	
F_{SC}: 0.0603	F_{SC}: 0.5754	F_{SC}: -0.0439 [1)]	F_{SC}: 0.5624
F_{ST}: 0.7878	F_{ST}: 0.8451	F_{ST}: 0.7888	F_{ST}: 0.7320
F_{CT}: 0.7742	F_{CT}: 0.6351	F_{CT}: 0.7977	F_{CT}: 0.3876

1) The F_{SC} value is not significant for this grouping.

It is seen from Table 4.4 that the SAMOVA suggested grouping has assigned the populations into three groups, with a high F_{CT} value (0.7742). This is largely in accordance with the manual grouping based on genetic composition of the populations, as MR1, MR7, MR8 and MR10 sharing the same haplotype RTM01/07 are assigned to the first group and MR2,

MR3, MR4, MR5, MR11 and MR13-15 sharing RTM02/08/09 are assigned to the second group. While MR6 and MR12 are genetically different as they contain different haplotypes, they were assigned to the same group by SAMOVA. Manually assigning MR6 and MR12 apart did render a higher F_{CT} value (0.7977), though a negative and non-significant F_{SC} value (-0.0439) was obtained under this grouping plan. Meanwhile it is seen that a lower F_{CT} value (0.3876) was obtained when the populations were assigned into four groups according to their geographic distribution, which indicate that geographic distance does not explain as much inter-population genetic variation.

Pairwise F_{ST} values between populations and the results of the population differentiation test are listed in Table 4.5. It is seen that the population differentiation test shows exactly the same pattern as suggested in the AMOVA analysis when the populations were assigned into four groups, as populations from different groups are shown to be differentiated while populations from the same group are shown to be undifferentiated. Pairwise F_{ST} values also agree with the pattern as they are positive when the two populations are from different groups and zero when the populations are from the same group.

Table 4.5 Pairwise F_{ST} values between all samples populations (below the diagonal) and the significance of population differentiation (above the diagonal). '+' indicates significantly different and '-' indicates not significantly different.

	MR1	MR23	MR4	MR5	MR6	MR7	MR8-10	MR11	MR12	MR13
MR1		+	+	+	+	-	-	+	+	+
MR2-3	1.0000		-	-	+	+	+	-	+	-
MR4	1.0000	0.0000		-	+	+	+	-	+	-
MR5	1.0000	0.0000	0.0000		+	+	+	-	+	-
MR6	0.4827	0.6829	0.4486	0.3400		+	+	+	+	+
MR7	0.0000	1.0000	1.0000	1.0000	0.5609		-	+	+	+
MR810	0.0000	1.0000	1.0000	1.0000	0.5065	0.0000		+	+	+
MR11	1.0000	0.0000	0.0000	0.0000	0.4745	1.0000	1.0000		+	-
MR12	0.8915	0.9569	0.8167	0.6502	0.1081	0.9318	0.9063	0.8468		+
MR13-15	1.0000	0.0000	0.0000	0.0000	0.6663	1.0000	1.0000	0.0000	0.9526	

4.3.3 Demographic history

Results from both Tajima's *D* (*D*=0.8466, p=0.82) and Fu's Fs (Fs=1309, p=0.98)) suggested that neutral assumption is not rejected. The τ values with their lower and upper bound within 95% confidence interval and corresponding p values (SSD-p) were obtained for the pooled population in the mismatch analysis under both the sudden demographic expansion model and the spatial expansion model. The corresponding ragged index (Ri) and ragged significance (Rp) were also calculated. The above results are listed in Table 4.6.

Table 4.6 Mismatch distribution analysis of population demographic history as inferred from Arlequin. τ values are indicated for the pooled population, with lower and upper bounds at the 95% confidence level under both expansion models. Ri indicates raggedness index and Rp indicates raggedness p value.

Model	parameters					
	τ	Low. τ	Upp. τ	SSD-p	Ri	Rp
Sudden (demographic) expansion	5.5566	0.2910	91.5566	0.03	0.3824	0.02
Spatial expansion	**4.3051**	1.4169	6.9082	0.17	0.3824	0.37

It is seen that both the SSD-p and raggedness p values for sudden expansion model are lower than 0.05, thus the model is not accepted. In contrast, the SSD-p and raggedness p values for spatial expansion model are higher than 0.05, which means the model is not rejected. It is thus

inferred that the total populations of *M. recurva* may have experienced spatial expansion in their history. The τ value (4.3051) is thus used to calculate the expansion time. With the μ value estimated at $3.6\text{-}5.9\times10^{-6}$ per year in the formula $\tau=2\mu t$ (see section 4.2.3), the expansion time is estimated to be 0.37-0.60 million years ago. If the lower or upper bound of τ is used, the expansion time is 0.12-0.20 or 0.59-0.96 million years ago.

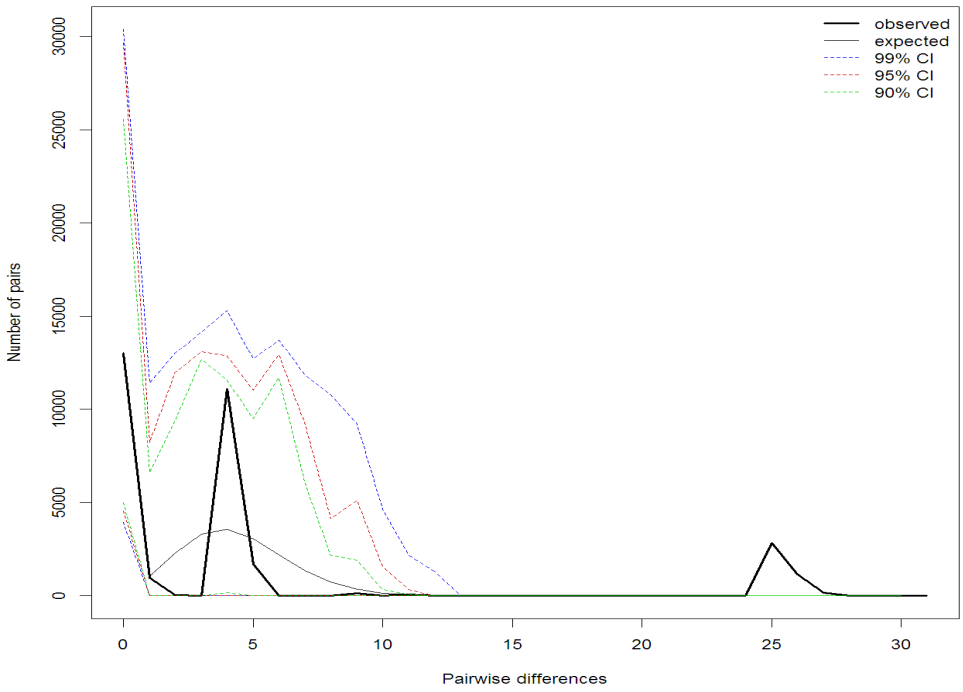


Figure 4.5 Result of mismatch distribution of the pooled populations under spatial expansion model. The x axis indicates the number of differences between pairs of haplotypes and the y axis indicates the occurrence of each number. The black solid line shows the observed distribution of pairwise nucleotide differences within the population, the grey curve is the expected distribution under the spatial expansion model, while the dashed lines with colours mark the credible intervals of the distribution at different significance levels.

Mismatch distribution analysis has rendered a multimodal profile, which is shown in Figure 4.5. There are two peaks in the mismatch distribution profile, one occurring at 4 pairwise differences and the other at 25 pairwise differences.

4.4 Discussion

4.4.1 Phylogeography of *M. recurva*

There are fewer haplotypes found in *M. recurva* across its distribution in Europe compared to those found in *Arenaria ciliata* and *A. norvegica* as described in Chapter 3. Also the divergence among haplotypes is much smaller than that found in the *Arenaria* species. The deepest divergence in *M. recurva* is seen between clade I and II, which was dated to be 0.5-0.9 Myr ago. The date is comparable to that of the divergence between clades II and III in *A. ciliata*, which was estimated to be 0.6-1.0 Myr ago. As clades II and III were jointly considered as a single taxonomic group during the analysis in Chapter 3, it is reasonable to consider clades I and II in *M. recurva* as in the same taxonomic group in this section of study.

Although there are fewer haplotypes found in the species, the geographic

distribution of the haplotypes does not simply comply with their genealogy. The most common haplotype, RTM02 (including RTM08 and 09), is found in a wide range of places including southwest Ireland, the north side of Spain, the southwest part and north side of Alps. The second most common haplotype, RTM01 (including RTM07) is found in Ireland, the north side of Spain and the Pyrenees. The distribution of the two haplotypes appears to overlap in broad geographic terms, but the two never co-occur at the same site.

Across the wide geographic range of RTM02/08/09, the difference in the two adjacent nucleotide sites of SSRs within trnT-trnL was the only DNA variation that may give extra information (Figure 4.6), from which it is seen that either RTM02 or RTM09 is one mutational step from RTM08, which seems an intermediate haplotype between RTM02 and 09.

Figure 4.6 The sequence variation in the two sites of SSRs. The red square shows the differences among RTM02, 08 and 09, while the green square shows difference between RTM01 and 07.

Due to the limited sensitivity of HRM analysis it was not possible to assign each individual in these populations to the exact haplotype. However based on DNA sequencing of representative samples it is known that RTM02 is found only in Ireland and RTM09 is found in Ireland (MR2, 3), Spain (MR11) and Switzerland (MR13-15), while their intermediate haplotype, RTM08, is found only in southwest part of Alps (MR4, 5). Thus no concordance is seen between the haplotype genealogy and the geographic distribution of the three closest haplotypes.

Similarly when RTM01 and RTM07 are considered as two haplotypes with a single A insertion/deletion between them, it is seen that RTM01 is found in Ireland (MR1) and the Pyrenees (MR8, 10) while RTM07 is found in north side of Spain (MR7). While the geographic distance is shorter between MR1 and MR7, the variation in SSRs shows a closer relationship between MR1 and MR8/10.

However, considering that DNA variation in SSRs may provide misleading information in phylogenetic analysis as described in section 4.2.1, the above discussion based on SSRs may be questionable. The lack of variation within each emerged haplotype only indicates that the dispersal of the species across its current distribution may have occurred

during a recent period.

The ancestral haplotypes RTM04, 05 and 06 are all found in Serbia or Kosovo, limited within the Balkans distribution. Another haplotype, RTM03, which is a two-step mutant of RTM02, is also found in Kosovo. The result shows that the Balkan populations may host the highest genetic diversity of the species, which may also have been the oldest refugia during the Ice age.

4.4.2 The structure of the populations

Both pairwise FST values and the exact differentiation significances suggested the grouping plan based on the apparent genetic difference among the populations. The analyses of SAMOVA suggested group structure of the populations when K=3 was in accordance with the manual grouping based on their evident genetic relationship, except that MR6 and MR12 were put into the same group. SAMOVA failed to provide the proper grouping when K=4, probably because it requires all fixation index values to be significant. However when MR6 and MR12 were manually put into separate groups, a higher significant F_{CT} value was indeed obtained. Meanwhile, the manual grouping based on

geographic localities rendered a much lower F_{CT} value, indicating that the geographic separation cannot explain the genetic difference among populations sufficiently in this case.

4.4.3 The origination of the Irish populations

The occurrence of RTM01/07 in south Ireland (MR1 from Waterford) and two places in Spain indicates a close relationship between these populations. Similarly, the occurrence of RTM02/08/09 in southwest Ireland (MR2 and 3 from Kerry), Spain and the Alps indicate that these populations are related. The Irish populations may have established after RTM01/07 diverged from RTM02/08/09, which is estimated between 0.17-0.28 Myr ago and the present. It is thus speculated that the Irish populations of *M. recurva* established during a more recent period than the populations of *A. ciliata*. However, no informative DNA polymorphisms are available at this moment to determine the exact time when the species came to Ireland, and the survival the Irish populations through the last glaciation remains a possibility.

4.5 Conclusion

With genetic information from chloroplast DNA, nine (or six is SSRs is not considered) haplotypes have been revealed in the species *M. recurva*, falling within two clades that diverged 0.5-0.9 Myr ago. The distribution of the haplotypes in the older clade is restricted to the Balkans, which also hosts one haplotype in the other clade. Thus the Balkans is considered to be the oldest refugium of the species. The most common haplotypes are in the other clade and have diverged more recently, which have occupied all the studied populations across Ireland, the Iberian Peninsula and the Alps. The study shows that all except the Balkan populations have established between 0.17-0.28 Myr ago and the present, which is more recent compared to the establishment of the *Arenaria* species as described in Chapter 3. As an alpine plant, *M. recurva* may have experienced a different phylogeographic history from that of the arctic-alpine *Arenaria* species.

Chapter 5

High Resolution Melting analysis: evaluation of its efficiency in phylogeographic research

5.1 Introduction

5.1.1 Background

In the previous chapters HRM analysis was successfully used with two chloroplast DNA regions for haplotype detection in the phylogeographic research of two carnation species. However if HRM analysis is going to be widely applied in phylogeographic research, evidence is needed to support the efficacy of the method with different DNA loci from various taxonomical groups. This chapter thus presents a more detailed evaluation of the sensitivity of HRM analysis to different DNA templates.

Based on the comparison between the *in vitro* and *in silico* HRM Tm values described in Chapter 2 (Table 2.10 and 2.13), it has been demonstrated that the software uMelt^{SM} (Dwight *et al.* 2011) is a useful tool for prediction of Tm variation between alleles of the same DNA locus. An *in silico* ΔTm between two amplotypes of greater than 0.1°C is considered the detectable threshold of Tm variation in equivalent *in vitro* analysis. Thus *in silico* HRM analysis with uMelt^{SM} is here used to predict if two given allele amplotypes from a given taxon are distinguishable via *in vitro* HRM analysis, and to evaluate the probability

of missed detection with various DNA loci of different sizes and template compositions.

5.1.2 Aims and objectives

In this section of the study, the primary aim is to understand how HRM analysis performs with a variety of template DNA loci varying in size and sequence content, both from computer-generated sequence batches, and from actual sequences recorded in the literature. With computer-generated DNA sequences, the likelihood that two allele amplicons varying by small deviations in sequence can be distinguished by HRM analysis is assessed. With actual DNA sequence data from peer-reviewed sources, it is examined how many out of the extant haplotypes revealed by exhaustive sequencing can be distinguisehd by HRM analysis. Using this dual approach, the efficacy of the method can be better understood in a broader sense, so as to fill the knowledge gap regarding the potential utility of HRM analysis for haplotype detection in future research work.

With the above aims, four objectives are proposed. First, as the basis of sensitivity assessment, the missed detection rate of HRM analysis is evaluated regarding the four different classes of single nucleotide

substitution that can occur between otherwise identical templates. Second, the sensitivity of HRM is investigated with respect to incremental increases in the size of amplicons being analysed. Thirdly the performance of HRM in detecting single and double SNP mutations between amplicons at different size intervals is evaluated. Finally, the performance of HRM analysis is evaluated with datasets from published literature in phylogeographic research, covering different DNA regions and a variety of taxa under investigation. All these data are then brought together to discuss the effectiveness and limitation of HRM analysis for haplotype detection in phylogeographic research.

Objective 1: HRM performance with different mutation classes

Although posterior DNA sequencing identified only one extra haplotype undetected by HRM analysis of rps16 in *Arenaria ciliata* and *A. norvegica* (see Chapter 2 and Dang *et al.* 2012), the question of the extent of missed detection remains. The representative amplicon rps16*I* contains two melting domains and thus shows two melting peaks, which is favourable for use in HRM analysis because multiple peaks offer increased capacity for identifying template deviation (Wittwer *et al.* 2003). However, it is not known how many of the possible mutant alleles

of such an amplicon could be distinguished from the wild type via HRM analysis, considering a single nucleotide polymorphism (SNP) mutation can occur on any nucleotide site within the amplified region.

There are four classes of SNP mutations, including class I (A/G and T/C), class II (A/C and T/G), class III (C/G) and class IV (A/T) mutations. The class IV A/T substitutions are regarded as most difficult to detect via HRM (Reed & Wittwer 2004) though G/C substitutions can also be problematic (von Ahsen *et al.* 2001; Liew *et al.* 2004). However, HRM sensitivity has not been evaluated quantitatively for the four mutation classes in the case of a single DNA region where all possible SNP mutations to the wild-type template are checked. *In vitro* HRM analysis for this kind of permutation test would be cost prohibitive and time consuming, whereas *in silico* HRM analysis allows us to do such work without the use of any wet-lab time or resources.

Thus as a case study, the locus of rps16*I* is used with *in silico* HRM analysis to quantitatively evaluate the missed detection rate of HRM analysis considering all possible SNP mutations within the amplified region, and to see how the detection rate varies among the four classes of SNPs. The missed detection rate is also evaluated regarding mutants

carrying two SNPs. Mutations with more than two nucleotide substitutions are not investigated in the present study because they require significantly increased amount of computation, which is time prohibitive at present.

Also the focus is only on substitutions and not on insertions and deletions (indels) for two reasons. The first reason is that large indels (>3bp) usually cause significant Tm variation in HRM analysis (examples are shown by the case study of *Arenaria* species in Chapter 2), so identification of this kind of mutation is unlikely to be problematic for HRM. The second reason is that small indels as microsatellite polymorphisms are usually omitted from phylogeographic analyses because they are evolutionarily labile and potentially provide misleading information (Small *et al.* 1998), while it should be born in mind that both variation in simple sequence repeats and other small indels can be difficult to characterise by HRM analysis (shown by the trnT-trnL HRM analysis in Chapter 2).

Objective 2. The effect of amplicon size on the number of melting domains

In order to assess the effect of the amplicon size on HRM sensitivity, the first priority is to assess how likely amplicons of different sizes may generate multiple melting domains. Although it is known HRM analysis usually exhibits lower sensitivity with longer amplicons when only a single melting domain is present, longer amplicons are more likely to yield multiple melting domains, which may improve the performance of HRM analysis in mutation detection (Wittwer *et al.* 2003). Here it is investigated quantitatively how amplicons at different size intervals are likely to generate multiple melting domains. For this aim, random DNA sequences generated on computer with a range of designated sizes are tested through the *in silico* HRM analysis.

As discussed in Chapter 2, amplicons of 250-400bp are the most favourable to be used for HRM analysis in phylogeographic studies as a balance between HRM sensitivity and locus coverage. Thus to cover an appropriate range of size intervals around this optimal size group, this analysis will include amplicon sizes between 50 and 650bp.

Objective 3. The effect of amplicon size on detection rate

Besides the probability of generating multiple melting peaks, the exact detection rate is also assessed for amplicons at different size intervals. Although larger amplicons are in theory more likely to carry multiple melting domains, the exact detection rate regarding all possible mutations within an amplicon may still vary with different amplicon sizes. In particular, when it is known that two melting domains are present for each amplicon, the larger amplicons are still expected to show lower detection rates of mutations. This expectation is being quantitatively tested in this part of study. In order to cover the amplicon sizes of 250-400bp which are the most favourable for HRM analysis in phylogeographic studies, the amplicons between 100 and 550bp are assessed in this section of the study.

Reed and Wittwer (2004) have assessed the performance of HRM analysis with a range of amplicon sizes between 50 and 1,000bp, and showed that mutations in <300bp amplicons were 100% detectable while mutations in 400-1000bp amplicons were partly missed with 96.1% detectable. However, their study was non-exhaustive based on limited numbers of mutations at specific nucleotide sites within the amplicons.

With the aim in the present study of detecting unknown haplotypes in *de*

novo sample groups, rather than identifying known haplotypes among samples, the discrimination rate between haplotype templates is investigated in terms of mutations on all possible nucleotide sites within the amplified DNA region. However, only class III (C/G) SNPs instead of all the four classes of SNPs are investigated for comparison of the detection rate between different amplicon sizes, as these represent the class of mutations that is difficult to detect via HRM (Chapter 2, Reed & Wittwer 2004). As done with the detection rate test of rps16*I* (5.2.1), mutations with two SNPs are also investigated in this part of study. However these are restricted to double class III mutations. Mutations with more than two SNPs and those with indels are not included here for the same reason as described in 5.2.1. The reason for including only one class of mutations (class III) is that inclusion of all the four substitution classes would be time-prohibitive at this point, and the inclusion of one class is sufficient in terms of comparing the amenability of different amplicons to HRM analysis.

Objective 4. *Post-hoc* HRM analysis of published data

Several recent phylogeographic studies are considered for this *post-hoc* HRM analysis where DNA sequencing of all available individual samples

was conducted to get full information of the extant polymorphisms presented by the collected samples of specific taxonomic groups. The sequence data in these studies cover both chloroplast and mitochondrial DNA regions, which are normally used in phylogeographic studies of plants and animals (Doellman *et al.* 2011; Gutiérrez-Rodríguez *et al.* 2011; Bisconti *et al.* 2011; Garcia *et al.* 2011; Westergaard *et al.* 2011). These data serve as an ideal resource for post-hoc *in silico* HRM analysis, to demonstrate how many of the revealed haplotypes would have been distinguished if HRM analysis had been used and also how the screening efficiency would have impacted on the number of samples needing full sequencing analysis. This section of study will provide a guide of how well HRM analysis can be applied in future phylogeographic research.

5.2 Methods

5.2.1 Evaluation of HRM performance with different mutation classes

The amplotype *I*01 of rps16*I* found in *Arenaria ciliata* was chosen as wild type, based on which mutant amplotypes were generated on computer, each with a single nucleotide substitution from the wild type. Class I-IV

mutations were made on all possible nucleotide sites within the amplicon except the primer-binding regions (sites 1-21 and 328-352). For each A or T nucleotide on the wild-type template, a class I (A to G or T to C), class II (A to C or T to G) and class IV (A to T or T to A) mutation was made, giving 3 mutant amplotypes for each position included. Similarly for each C or G nucleotide, a class I (C to T or G to A), class II (C to A or G to T) and class III (C to G or G to C) mutation was also made. Though specific evolutionary models of substitution rates may apply in this and other cases, all classes of mutations are herein considered to be equally likely to happen on all possible nucleotide sites as an ideal condition to estimate the detection rate.

All of these mutant amplotypes were inputted to uMelt^{SM} and their melting peaks and corresponding Tm values were obtained and tabulated on a spreadsheet. The settings within the software were the same as those described in section 2.2.6, where the thermodynamic set of Huguet et al. (2010) was used, the Mg2+ and mono-valent cation concentrations were set at 2.5mM and 22mM respectively, and DMSO concentration was set at 10%. As demonstrated in 2.3.1.7, the two amplotypes were considered to be indistinguishable by *in vitro* HRM analysis if the *in silico* ΔTm between two haplotypes is less than 0.1°C. This approach facilitated an *in*

silico prediction of the theoretical detection rate during *in vitro* HRM analysis for all possible occurrences of the four classes of mutations to the wild-type template.

Within the detection rate test for each class of mutations, it arises that all the mutants generated from the same wild type are mutually different by two SNPs, so the detection rate for double nucleotide mutations can also be estimated, i.e. if two alleles of rps16*I* are different by two SNPs, how likely they are to be distinguished by HRM analysis. This serves as an estimate of how many out of all the possible DNP mutants from rps16*I*01 can be distinguished from it by HRM analysis.

5.2.2 Evaluation of the effect of amplicon size on the occurrence of multiple melting domains

Random DNA sequences were generated via the online tool at http://users-birc.au.dk/biopv/php/fabox/random_sequence_generator.php, with the GC fractions of 0.2 and AT fractions of 0.3. Considering that the DNA regions used in our study on the *Arenaria* species and other phylogeographic studies (see section 5.3.4) have their GC content between 25% and 45%, it is reasonable that we set the GC content of all

the generated sequences around 40% to mimic realistic situations. Fifty random sequences were generated for each amplicon size, which range from 50bp to 650bp in 50bp intervals. These sequences were inputted into uMelt^{SM} and their *in silico* melting profiles were obtained. The parameter settings within the software were the same as described in section 5.2.1. The sequences showing more than one melting peak were counted and their frequency was recorded for each amplicon size.

5.2.3 Evaluation of the effect of amplicon size on the sensitivity of HRM analysis

To observe the impact of amplicon size on HRM sensitivity, seven sequences with sizes of 100, 150, 257, 354, 456, 533 and 550bp were selected as wild types to generate mutants. The seven wild type amplicons all show two melting domains with the two corresponding melt peaks well separated (by 3.5-9.5°C, shown in Figure 5.2). Class III (C/G) mutations were made on all possible nucleotide sites (except 20bp zones at both the beginning and the end of the sequence, assumed as the primer binding regions) to generate mutant sequences each carrying a single C/G substitution from the wild type. All the generated mutants were inputted into uMelt^{SM} for *in silico* HRM analysis. Tm values were recorded for all

the mutant sequences as well as the wild types, and mutants with Tm difference (≥0.1°C) in at least one of the melting peaks from the wild type were considered as detectable by *in vitro* HRM analysis. The detection rate was thus counted for each amplicon size.

The detection rate of DNP mutations was also calculated for each of the above templates in the same way as described in section 5.2.1, as all the mutants generated from the same wild type are mutually different by two SNPs. However, only double class III mutations were taken into account for this calculation.

Aside from the seven amplicons with double melting peaks, two additional amplicons of 150bp and 250bp with single melting peak were also used as wild types to check HRM sensitivity. The missed detection rates were compared with those from the 150bp and 257bp double-peak amplicons to confirm that amplicons with multiple melting domains yield higher HRM sensitivity than the amplicons with single melting domain of the same size.

5.2.4 *In silico* HRM analysis of published data

A selection of recently published works in phylogeography were chosen for *post hoc* HRM analysis, where DNA sequences of multiple haplotypes were found for the target loci in chloroplast or mitochondrial genomes. Some of the chosen studies also included sequence data from nuclear genomes, however the analysis of diploid genotypes is not being investigated in the present study because of the different and more complex melting profiles generated by possible heterozygous compared to homozygous genotypes (Wittwer *et al.* 2003). As such, although nuclear loci do represent a further category of potential HRM genotypeing analysis, they are not investigated in the present study. In this section the amplified organelle DNA regions described in these publications (both mitochondria and chloroplast) have been subjected to *post-hoc in silico* HRM analysis using the software uMelt^{SM}, based on the same parameter settings as used in section 5.2.1. As per the protocol of using amplicon sizes which optimize HRM sensitivity, each amplified sequence has been split into 200-400bp amplicons to facilitate HRM analysis, but occasionally a target region of c.500bp has been entirely put into the *in silico* assay with reliable sensitivity results. Tm values for all detected amplotypes were recorded and allele amplotypes sharing an

identical combination of Tm values were considered as indistinguishable. Similarly haplotypes of each original DNA locus (comprising 2+ concatenated amplotypes) that shared the same combination of Tm values were considered indistinguishable. The observed frequency of missed detection for HRM was determined by counting the indistinguishable haplotypes relative to the total number of haplotypes detected by exhaustive DNA sequencing. Furthermore, where precise geographic information was available, it was checked whether the indistinguishable haplotypes co-occurred within the same populations. The likelihood of missed detection is considered to be reduced if the indistinguishable haplotypes have discrete distributions because the assignment of samples into putative haplotypes is performed within each population during *in vitro* HRM analysis. When geographic information is unavailable in a particular literature, the phylogeny of the sampled haplotypes is evaluated to determine what mistakes could arise due to missed detection in HRM analysis.

It is noted that the observed frequency of missed detection in this section is different from the pairwise rate of missed detection in sections 5.2.1 and 5.2.3. In the above sections the focus was how likely two amplotypes are to be distinguished by HRM analysis. In this section of the study the

concern is how many out of all the extant haplotypes within the target populations can be detected by HRM analysis and how many are still unrevealed.

5.3 Results

5.3.1 HRM performance with different mutation classes

Within the 352bp region of rps16*I*, class I (A/G or C/T) mutations and class II (A/C or G/T) mutations were generated on all 306 nucleotide sites between site 22 and 327, thus 306 mutant amplotypes were obtained for each of class I and II mutations. Class III (C/G) and IV (A/T) mutations were generated on all possible sites within the same region, resulting in 95 class III mutants and 211 class IV mutant amplicons (Dang *et al.* 2012). Those with Tm values different from that of the wild type (Tm1=77.7°C and Tm2=80.6°C) are considered as detectable, the numbers of which are listed in Table 5.1 by different mutation classes. The overall missing detection rate for SNP mutants is thus calculated to be 20.48%, indicating that one in every five of all the possible mutants with a SNP from the wild type is undetectable by HRM analysis.

Each of the mutant amplotypes listed in Table 5.1 differs from the wild-type by one SNP. However if the mutants are compared against each other, all pairwise comparisons involve double SNPs (DNPs), aside from the subset of cases where these SNPs occur at the same base position. Thus the missed detection rate for DNP mutants can also be estimated by counting those who share the same mutant Tm values, as shown in Table 5.1. However, to avoid counting different mutants occurring on the same nucleotide sites, the detection rate between DNP mutants is calculated separately within each mutation class and across each pair of mutation classes.

Class I mutations in this case can occur at 306 possible nucleotide sites within the amplified region of rps16*I*01, thus there can be 306×(306-1)/2 = 46,665 possible combinations of DNP mutation types when both SNPs are class I mutations. Within this class, in total 20 mutant Tm groups were identified (group A to T in Table 5.1), each comprising a number of different mutant amplotypes, as shown in column 4 of Table 5.1.

Table 5.1 Detection of SNP mutants within the four classes of nucleotide substitutions. Mutants with different Tm values are considered as detectable. The mutants are grouped by their Tm1 and Tm2 values, and the number of mutants is counted for each Tm group to facilitate the missing detection analysis for DN mutants (mutants varying by double nucleotide substitutions).

Tm group	Tm1 (°C)	Tm2 (°C)	Class I	Class II	Class III	Class IV	Total
As WT*	77.7	80.6	5	4	49	130	
A	77.7	81.1	8	15	0	0	
B	77.7	81.2	7	5	0	0	
C	77.7	80.1	11	10	0	0	
D	77.7	80.8	10	4	7	2	
E	77.7	80.2	9	8	0	0	
F	77.7	81.0	23	11	0	0	
G	77.7	80.3	8	16	0	0	
H	77.7	80.9	8	19	5	2	
I	77.7	80.4	6	2	8	5	
J	77.7	80.0	1	2	0	0	
K	77.7	80.5	6	3	4	8	
L	77.7	80.7	8	9	3	13	
M	77.8	80.7	4	4	0	0	
N	77.8	80.6	39	29	7	37	
O	77.6	80.6	11	9	12	13	
P	77.5	80.6	26	29	0	0	
Q	77.9	80.6	48	67	0	1	
R	78.0	80.6	51	44	0	0	
S	77.4	80.6	13	12	0	0	
T	78.1	80.6	4	3	0	0	
U	77.7	81.3	0	1	0	0	
Total			306	306	95	211	918
Detectable			301	302	46	81	730
Detection rate (%)			98.36	98.69	48.42	38.39	79.52
Missing detection rate (%)			1.64	1.31	51.58	61.61	20.48

* ‘as WT’ means the same Tm values as the wild type

Within each Tm group, taking group A as the example, 8 one-step mutants were found, thus 8×(8-1)/2 = 28 pairs of them are counted as indistinguishable DNP mutant pairs. The same counting was performed within each Tm group including the ‘As wild’ group, and finally added up as the total missed numbers for DNP mutants when both SNPS are class I mutations, which is 4,176 in this case. In the same way, the missed numbers are calculated for DNP mutants when both SNPS fall in class II, III and IV, respectively. The result is shown in column 4 of Table 5.2.

Besides the DNP mutants when both SNP mutations are in the same mutation class, those with the two mutations falling within different classes are also considered. Taking for example two mutants that differ by one class I and one class II mutations and are also indistinguishable by HRM analysis, one possibility is that they would share group B Tm values (Tm1=77.7°C, Tm2=81.2°C), in this case involving 7 class I mutants and 5 class II mutants (Table 5.1). So the example pair should be one of the 7×5 candidate combinations. However, the SNPs on one of the 7 class I candidates and one of the 5 class II candidates were found to both occur at nucleotide position 63, and the same co-occurrence was seen at position 70. Thus the possible combination number in this case is 7×5-2 = 33, i.e. in total there are 33 pairs of class I × class II DNP

mutants sharing the same group B Tm values. This number of non-distinguishable pairs was calculated for each Tm group including the wild-type Tm group, and finally totalled 8,530 pairs. This means that in total 8,530 pairs of class I × class II DNP mutants are indistinguishable by HRM analysis. In the same way the missed detection rate of DNP mutants was calculated for each combination of mutation classes. Finally the overall missed detection rate was calculated for DNP mutants. The results are listed in Table 5.2.

Overall, in 9.73% of the cases where two mutants vary by two nucleotide substitutions, the mutants cannot be distinguished from each other by HRM analysis. It is considered as an estimate of the missed detection rate for all the DNP mutant amplotypes from the same wild type of rps16*I*, i.e. 9.73% of the DNP mutant amplotypes cannot be distinguished from the same wild type rps16*I*01. The overall missing detection rate for double nucleotide mutations is approximately half of that seen for single point mutations.

Table 5.2 Missed detection rate for HRM analysis of DNP mutants, where the SNPs in the paired amplicons are in the same mutation class (rows 1 to 4) or different classes (rows 6 to 10).

Combination of DNP mutation	Possible number of DNP mutants		Number of missed detections	Missed detection rate (%)
Class I × class I	306×305/2	46665	4176	8.95
Class II × class II	306×305/2	46665	4667	10.00
Class III × class III	95×94/2	4465	1331	29.81
Class IV × class IV	211×210/2	22155	9247	41.74
Class I × class II	306×305	93330	8530	9.14
Class I × class III	305×95	28975	856	2.95
Class I × class IV	305×211	64355	2498	3.88
Class II × class III	305×95	28975	682	2.35
Class II × class IV	305×211	64355	1971	3.06
Class III × class IV	211×95	20045	6920	34.52
In total		419985	40878	9.73

5.3.2 The effect of amplicon size on occurrence of multiple melting domains

Within the 650 DNA sequences ranging between 50bp and 650bp in length, 307 rendered more than one melting peak, of which 249 rendered double melting peaks and 58 rendered triple peaks. The number of the sequences rendering multiple peaks was counted out of 50 sequences for each size and shown in Figure 5.1.

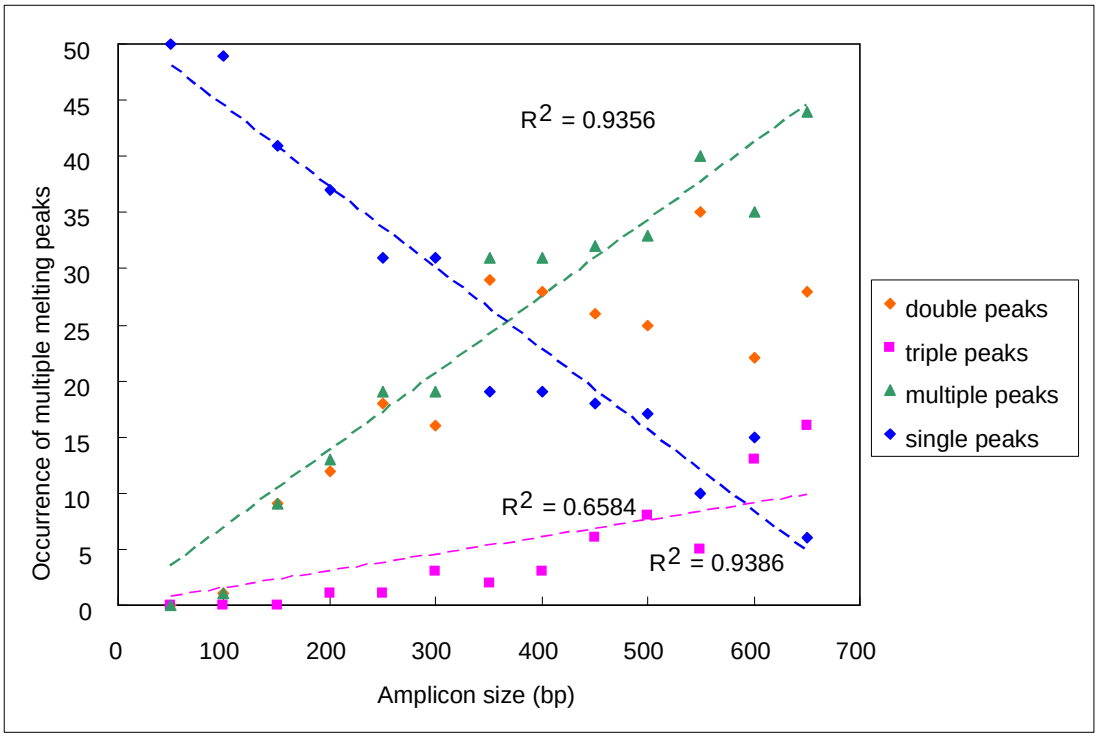


Figure 5.1 The number of sequences carrying multiple melting peaks out of 50 random sequences for each amplicon size. The occurrence of double peaks and triple peaks are shown separately for each group. The linear trend lines are made for the occurrence of single peaks (blue), triple peaks (pink) and multiple peaks (occurrence of triple peaks plus that of double peaks, shown in green) with the intercept set to zero.

It is seen from the permutation test results illustrated in Figure 5.1 that larger amplicons tend to generate multiple melting peaks with a higher frequency compared to smaller amplicons. Below 300bp most amplicons yielded single melt peaks, however from 300 to 350bp multiple peak amplicons become prevalent. While the correlation between amplicon size and number of double melting peaks is strong between 100 and 300bp, it becomes weaker between 350 and 650bp. Above 350bp the

correlation is more positive for triple peaks and where double and triple peaks are counted together as multiple peaks.

5.3.3 The effect of amplicon size on sensitivity of HRM analysis

Seven amplicons of 100bp, 150bp, 257bp, 354bp, 456bp, 533bp and 550bp were chosen as representatives of different size intervals, all of which show clear double melting peaks as shown in Figure 5.2. Two additional amplicons of 150bp and 250bp in length that showed a single melting peak were also included.

Class III (G/C) mutations were generated at all G and C nucleotide sites within the nine wild-type sequences except the first and last 20bp regarded as the primer-binding regions. The number of possible G/C mutations within each wild type is listed in Table 5.3. Both SNP and DNP mutant amplicons from the wild types were subject to *in silico* HRM analysis using uMelt^{SM}. The missed detection rates of both types of mutants were calculated separately for each of the nine wild types, which are listed in Table 5.3.

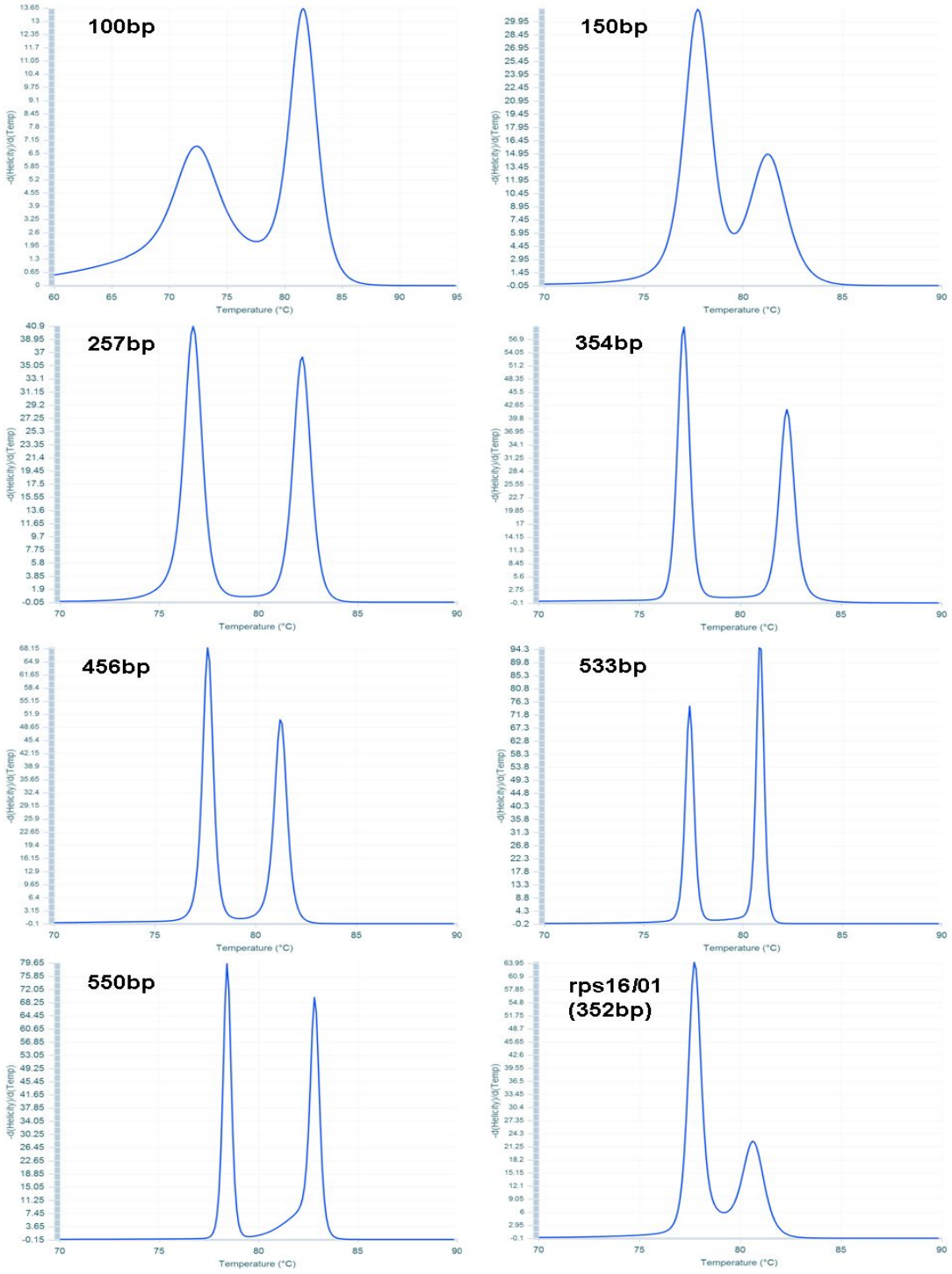


Figure 5.2 I*n silico* melting peaks of the eight representative amplicons with different sizes. The first seven amplicons are chosen as wild types to generate mutants to allow missing detection rate to be estimated. Each of the chosen amplicons shows distinct double peaks with the distance between Tm1 and Tm2 greater than 3°C (3.5-9.5°C), which is favourable in HRM analysis to bring more mutations to be visible. The *in silico* melting profile of the 352bp rps16*I*01 (see Chapter 2) is co-illustrated for comparison.

Table 5.3 Missed detection of class III mutations with both SNP and DNP mutants of the wild type DNA sequences with different sizes. The chi-squared (χ^2) value indicates if the difference between the observed SNP and DNP detection rates for each template is significant. The Pearson's *r* correlation coefficient is calculated to test the dependence between each type of missed rate and amplicon size for the amplicons with double melting peaks.

Size (bp)	GC conc. (%)	Class III SNP mutants			Class III DNP mutants			χ^2
		Total occurrence	Missed detections	Missed rate (%)	Total occurrence	Missed detections	Missed rate (%)	
Amplicons with double melting peaks								
100	38	26	7	26.92	325	12	3.69	16.77**
150	42	43	14	32.56	903	133	14.73	4.16*
257	41	85	42	49.41	3570	1039	29.10	9.15**
354	39	125	70	56.00	7750	2845	36.71	11.81**
352[1)]	*32*	*95*	*49*	*51.58*	*4465*	*1331*	*29.81*	*12.87***
456	38	158	111	70.25	12403	6499	52.40	12.07**
533	38	188	106	56.38	17578	7046	40.08	12.58**
550	41	208	145	69.71	21528	11071	51.43	35.09**
Pearson's *r*				0.9160			0.9323	
Amplicons with a single melting peak								
150	43	50	28	56.00	1225	432	35.27	3.48
250	39	84	45	53.57	3986	1332	38.21	7.93*

1) The 352bp amplicon rps16*I*01 as wild type is also included in the table, but not included for correlation test.

* stands for significance at $p<0.05$; ** stands for significance at $p<0.01$.

The results show that the occurrence of multiple melting peaks is confirmed to provide better HRM sensitivity, which is supported by the lower missed detection rates given by the 150bp and 257bp amplicons with double melting peaks than those given by the 150bp and 250bp amplicons with single melting peak respectively.

The values of Pearson's *r* reveal positive correlation between the amplicon size and the missed detection rates in both SNPs and DNPs, while the amplicons all yield well distinguished double melting peaks. However, the amplicons of 533bp and 550bp have a lower probability of missed detection than the 456bp amplicon (Table 5.3 and Figure 5.3). This implies a non-linear relationship between amplicon size and the missed detection rate for larger amplicons, which may in the present case, becomes non-positive with amplicon sizes greater than 450bp.

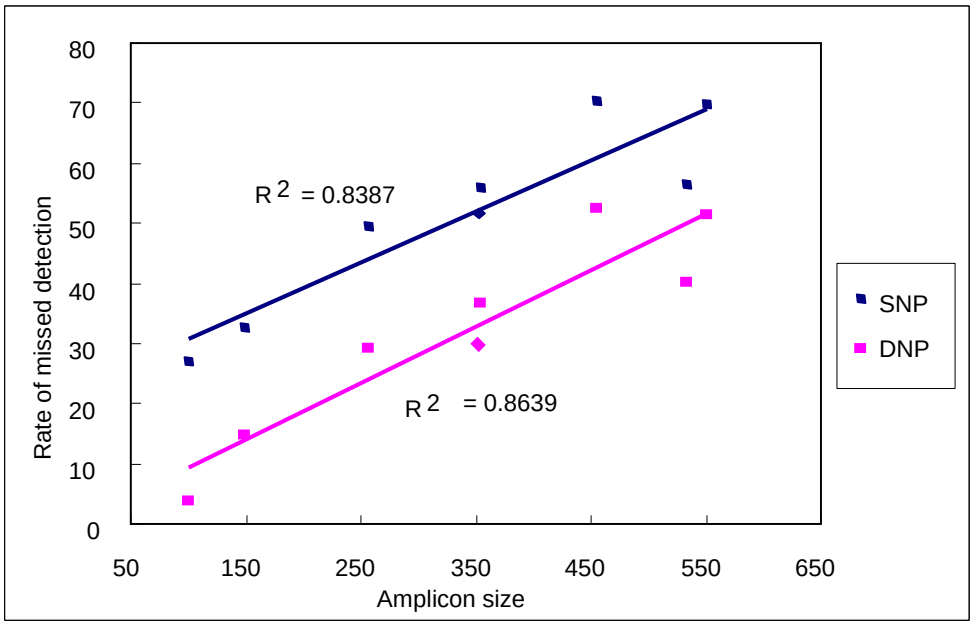


Figure 5.3 The rates of missed detection for both SNP and SNP mutations plotted against amplicon size. Amplicons are randomly generated DNA sequences with designated sizes from 100 to 550bp, with GC content set around 40%. Linear regression is performed across all the data points showing positive correlation between both missed detection rates and amplicon size. The missed detection rates for the 352bp rps16/01 are also plotted on the graph (in blue and pink diamonds for SNPs and DNPs) but not included for drawing the trend line.

Also it is clear that the missed detection rate of DNP mutation is lower than that of SNP mutation in every case except for the 150bp single-peak amplicon, which is supported by the result of χ^2 test as shown in Table 5.3. The result suggests that DNP mutants are more likely to be distinguished by HRM analysis than SNP mutants, which is in accordance with the result in section 5.3.1.

The G/C mutation test based on the realistic wild type, rps16*I*01 of 352bp with 32% GC content, is also listed in Table 5.3 to be compared with the artificially generated amplicons. While rps16*I*01 has a similar length to the 354bp artificial amplicon, its missed detection rates of both SNP and DNP G/C mutations are closer to those of the 257bp artificial amplicon. In Figure 5.3 the missed detection rates of rps16*I*01 are plotted onto the diagram, where its data points are close to the trend lines based on the data points from the artificial templates.

5.3.4 HRM performance with published data

5.3.4.1 HRM analysis with mitochondrial loci on *Hyla sarda*

In a study of the Tyrrhenian tree frog, *Hyla sarda* (Hylidae), distributed in

the Corsica–Sardinia islands, Bisconti and co-workers (2011) revealed 35 haplotypes of the Cyt b (cytochrome *b*) locus and 38 haplotypes of the ND1 (NADH dehydroge-nase subunit 1) locus in mitochondrial genomes by exhaustive DNA sequencing of all available individual samples with the two mitochondrial loci. *In silico* HRM analysis with uMelt^{SM} was conducted with the two mtDNA regions, by splitting Cyt *b* and ND1 into two and three amplicons respectively. It is noted that no indels were included in the DNA sequences from the original data.

HRM performance with Cyt *b*

The originally amplified locus of Cyt *b* was 531bp in size, and for the purpose of *in silico* HRM analysis was split into two amplicons, recorded as CytbI with 1-296bp and CytbII with 272-531bp. The two amplicons were designed to cover the full length of the original Cyt *b* locus while priming the conservative regions across all the detected haplotypes. The CytbI amplicon shows double melting peaks (indicated as Tm11 and Tm12) while the CytbII amplicon shows a single peak with a shoulder region on the left side (lower temperature) (indicated as Tm22 and Tm21). The entire Cyt *b* locus was also used as amplicon for the *in silico* HRM analysis, which also showed double melting peaks (Tm1 and Tm2). The

Tm values obtained from the *in silico* melting analysis are shown in Table 5.4, where the Tm21 value indicates the summit of the shoulder for CytbII where visible. This melting shoulder disappears with one of the amplicons of CytbII, which is also a discriminative feature between amplotypes.

From the results shown in Table 5.4 it is seen that when the entire 531bp Cyt *b* locus was amplified for HRM analysis, only seven of the haplotypes show distinct Tm1 and Tm2 values, while each of the other 28 haplotypes shares the same Tm values with at least one another haplotype. However, when the locus is split into two amplicons for HRM analysis, out of the 19 amplotypes of CytbI and 21amplotypes of CytbII revealed originally by exhaustive DNA sequencing, most could be distinguished by their Tm values via HRM analysis, with only a few indistinguishable from one another. CytbI01 and I05 shared the same Tm11 and Tm12 values and thus cannot be distinguished via HRM analysis, and the same situation is seen in the cases of CytbI06 vs. I15, CytbII02 vs. II07 and CytbII03 vs. II05.

Table 5.4 *In silico* Tm values of Cyt *b* amplicons from Bisconti *et al.* (2011), where Tm1 and Tm2 values are obtained by inputting the entire 531bp locus as the amplicon. Tm11 and Tm12 are obtained from CytbI, Tm21 and Tm22 are from CytbII. *Tm21 indicates the melting shoulder, which is absent in CytbII18.

Composite Cyt *b* haplotype	Tm1 (°C)	Tm2 (°C)	CytbI	Tm11 (°C)	Tm12 (°C)	CytbII	Tm21* (°C)	Tm22 (°C)
Cytb01	86.7	90.4	CytbI01	86.7	89.7	CytbII01	87.5	89.7
Cytb02	86.7	90.5	CytbI02	86.7	89.9	CytbII01	87.5	89.7
Cytb03	86.7	90.3	CytbI03	86.7	89.5	CytbII02	87.5	89.5
Cytb04	86.7	90.4	CytbI02	86.7	89.9	CytbII02	87.5	89.5
Cytb05	86.5	90.5	CytbI04	86.5	89.9	CytbII01	87.5	89.7
Cytb06	86.7	90.4	CytbI05	86.7	89.7	CytbII01	87.5	89.7
Cytb07	86.7	90.6	CytbI02	86.7	89.9	CytbII03	87.5	89.9
Cytb08	86.7	90.6	CytbI02	86.7	89.9	CytbII04	88.3	89.7
Cytb09	86.7	90.6	CytbI02	86.7	89.9	CytbII05	87.5	89.9
Cytb10	86.4	90.5	CytbI06	86.4	89.9	CytbII01	87.5	89.7
Cytb11	86.7	90.6	CytbI07	86.7	90.2	CytbII01	87.5	89.7
Cytb12	86.7	90.4	CytbI08	86.7	89.6	CytbII06	87.5	89.8
Cytb13	86.7	90.4	CytbI02	86.7	89.9	CytbII07	87.5	89.5
Cytb14	86.7	90.3	CytbI02	86.7	89.9	CytbII08	87.5	89.4
Cytb15	87.0	90.5	CytbI09	87.0	90.0	CytbII01	87.5	89.7
Cytb16	86.7	90.5	CytbI10	86.7	90.1	CytbII09	88.2	89.5
Cytb17	86.7	90.4	CytbI11	86.7	90.1	CytbII10	87.5	89.5
Cytb18	86.9	90.5	CytbI12	86.9	89.9	CytbII01	87.5	89.7
Cytb19	86.8	90.3	CytbI13	86.8	89.4	CytbII11	87.5	89.5
Cytb20	86.8	90.3	CytbI14	86.8	89.5	CytbII11	87.5	89.5
Cytb21	86.4	90.5	CytbI15	86.4	89.9	CytbII01	87.5	89.7
Cytb22	86.7	90.7	CytbI02	86.7	89.9	CytbII12	88.2	89.8
Cytb23	86.7	90.4	CytbI02	86.7	89.9	CytbII13	87.6	89.5
Cytb24	86.7	90.6	CytbI02	86.7	89.9	CytbII14	88.2	89.7
Cytb25	86.7	90.7	CytbI16	86.7	90.2	CytbII04	88.3	89.7
Cytb26	86.7	90.5	CytbI02	86.7	89.9	CytbII15	87.5	89.8
Cytb27	86.7	90.7	CytbI10	86.7	90.1	CytbII04	88.3	89.7
Cytb28	87.0	90.7	CytbI17	87.0	89.9	CytbII16	88.1	89.8
Cytb29	86.7	90.3	CytbI02	86.7	89.9	CytbII17	86.7	89.7
Cytb30	86.7	90.7	CytbI02	86.7	89.9	CytbII18	none*	89.7
Cytb31	86.8	90.5	CytbI14	86.8	89.5	CytbII19	87.5	89.9
Cytb32	86.7	90.6	CytbI02	86.7	89.9	CytbII20	88.6	89.7
Cytb33	86.8	90.2	CytbI18	86.8	89.1	CytbII11	87.5	89.5
Cytb34	87.2	90.3	CytbI19	87.2	89.5	CytbII11	87.5	89.5
Cytb35	86.7	90.7	CytbI02	86.7	89.9	CytbII21	88.5	89.7

As a result, in four cases two composite haplotypes are indistinguishable from each other based on HRM analysis (Cytb01 and Cytb06, Cytb04 and Cytb13, Cytb07 and Cytb09 and Cytb10 and Cytb21). Thus in total 27 out of the 35 haplotypes could be well distinguished by HRM analysis and the other eight haplotypes would be identified as four, with four of them masked by the shared DNA melting profiles.

None of the indistinguishable pairs of haplotypes was found to co-occur within the same population in this study based on Bisconti and co-workers' (2011) data. This again indicates that the strategy of performing HRM assays in batches of populations may reduce the possibility of missing detection in a realistic study. If the strategy described in Chapter 2 had been used with the Cyt b locus in Bisconti and co-workers' (2011) study, all the 35 haplotypes identified would have been revealed by the above protocol of HRM analysis.

HRM performance with ND1

The originally amplified locus of ND1 was 698bp in size, and was split into three amplicons, recorded as NDI (1-287bp), NDII (260-421bp) and NDIII (404-698bp).

Table 5.5 *In silico* Tm values of ND1 amplicons from Bisconti *et al.* (2011), where Tm11, Tm21 and Tm31 values are obtained by splitting the entire 698bp locus into three amplicons

ND1 haplotype	NDI	Tm11 (°C)	NDII	Tm21 (°C)	NDIII	Tm31 (°C)
ND01	NDI01	87.5	NDII01	85.1	NDIII01	87.1
ND02	NDI02	87.8	NDII01	85.1	NDIII02	87.2
ND03	NDI03	87.6	NDII01	85.1	NDIII02	87.2
ND04	NDI01	87.5	NDII01	85.1	NDIII02	87.2
ND05	NDI04	87.3	NDII01	85.1	NDIII02	87.2
ND06	NDI01	87.5	NDII01	85.1	NDIII03	87.0
ND07	NDI05	87.3	NDII01	85.1	NDIII02	87.2
ND08	NDI06	87.4	NDII01	85.1	NDIII01	87.1
ND09	NDI01	87.5	NDII02	85.2	NDIII02	87.2
ND10	NDI07	87.4	NDII01	85.1	NDIII04	86.8
ND11	NDI08	87.4	NDII01	85.1	NDIII05	86.9
ND12	NDI08	87.4	NDII01	85.1	NDIII06	87.1
ND13	NDI09	87.6	NDII01	85.1	NDIII07	87.3
ND14	NDI07	87.4	NDII01	85.1	NDIII05	86.9
ND15	NDI10	87.7	NDII01	85.1	NDIII02	87.2
ND16	NDI11	87.5	NDII01	85.1	NDIII02	87.2
ND17	NDI12	87.4	NDII01	85.1	NDIII08	87.0
ND18	NDI01	87.5	NDII03	84.8	NDIII01	87.1
ND19	NDI13	87.7	NDII01	85.1	NDIII01	87.1
ND20	NDI01	87.5	NDII04	85.4	NDIII01	87.1
ND21	NDI07	87.4	NDII01	85.1	NDIII09	86.8
ND22	NDI14	87.8	NDII05	84.5	NDIII10	87.0
ND23	NDI21	87.8	NDII06	84.8	NDIII11	86.8
ND24	NDI21	87.8	NDII06	84.8	NDIII12	86.9
ND25	NDI07	87.4	NDII01	85.1	NDIII13	86.8
ND26	NDI14	87.8	NDII05	84.5	NDIII14	86.8
ND27	NDI15	87.4	NDII07	84.7	NDIII01	87.1
ND28	NDI16	87.4	NDII01	85.1	NDIII01	87.1
ND29	NDI17	88.0	NDII06	84.8	NDIII15	86.8
ND30	NDI17	88.0	NDII08	84.4	NDIII14	86.8
ND31	NDI18	87.9	NDII04	85.4	NDIII01	87.1
ND32	NDI19	87.3	NDII01	85.1	NDIII01	87.1
ND33	NDI01	87.5	NDII01	85.1	NDIII16	87.0
ND34	NDI01	87.5	NDII01	85.1	NDIII17	87.2
ND35	NDI01	87.5	NDII01	85.1	NDIII18	87.2
ND36	NDI01	87.5	NDII01	85.1	NDIII19	87.0
ND37	NDI17	88.0	NDII06	84.8	NDIII20	86.6
ND38	NDI20	87.9	NDII06	84.8	NDIII21	87.0

All the three amplicons showed single melting peaks, indicated by the Tm11, Tm21 and Tm31 values shown in Table 5.5. It is seen that in six cases different ND1 haplotypes cannot be distinguished by HRM analysis even if cross-check between the three amplicons was performed during posterior DNA sequencing. In one case ND04, ND16, ND34 and ND35 shared the same Tm values. Other cases where two or three haplotypes cannot be distinguished include ND05 vs. ND07, ND06 vs. ND33 vs. ND36, ND08 vs. ND12 vs. ND28, ND10 vs. ND21 vs. ND25 and ND11 vs. ND14. In total 17 of the 38 haplotypes are masked by 6 different Tm profiles while the other 21 haplotypes show unique Tm profiles. Thus 11 out of the 38 haplotypes could have been missed if HRM analysis had been solely used for haplotype detection.

Also in terms of distribution of the haplotypes, only in one of the six cases the haplotypes sharing the same Tm values are found to co-occur within the same population, i.e. ND34 and ND35 are both found in the Monte Arcosu population from the island of Sardinia. This indicates that if the haplotypes are distinguished by HRM analysis within each population, only in one case a haplotype would be missed.

5.3.4.2 HRM analysis with a mitochondrial locus on *Littorina* sp.

In a phylogeographic study of the North Atlantic rough periwinkle, *Littorina saxatilis* (Littorinidae) and its sister species, Doellman and co-workers (2011) found 69 haplotypes of the mitochondrial ND1 (NADH dehydroge-nase subunit 1) locus via exhaustive DNA sequencing. Here the 633bp locus was divided into three amplicons that can be amplified by conservative primers, which are recorded as NDI (1-241bp), NDII (222-421bp) and NDIII (395-633bp) for in silico HRM analysis.

The three amplicons of the ND1 locus showed four melting peaks within the in silico HRM analysis, with the corresponding Tm values shown in Table 5.6.

Table 5.6 *In silico* Tm values of ND1 amplicons from Doellman *et al.* (2011), where Tm11, Tm21 and Tm31 values are obtained by splitting the entire 633bp locus into three amplicons.

ND1 haplotype	NDI	Tm11 (°C)	Tm12 (°C)	NDII	Tm21 (°C)	NDIII	Tm31 (°C)
ND01	NDI01	83.0	85.7	NDII01	83.7	NDIII01	83.0
ND02	NDI02	83.0	85.4	NDII01	83.7	NDIII01	83.0
ND03	NDI03	82.1	84.9	NDII02	83.1	NDIII02	82.3
ND04	NDI04	83.3	85.4	NDII03	83.2	NDIII03	82.8
ND05	NDI04	83.3	85.4	NDII04	83.3	NDIII01	83.0
ND06	NDI04	83.3	85.4	NDII05	82.9	NDIII04	82.7
ND07	NDI05	83.1	85.4	NDII03	83.2	NDIII04	82.7
ND08	NDI01	83.0	85.7	NDII01	83.7	NDIII05	82.8
ND09	NDI06	83.2	85.2	NDII06	83.3	NDIII06	82.7
ND10	NDI07	83.7	85.3	NDII07	83.5	NDIII07	82.9
ND11	NDI08	83.8	85.4	NDII07	83.5	NDIII07	82.9
ND12	NDI09	83.2	85.2	NDII06	83.3	NDIII08	82.8
ND13	NDI09	83.2	85.2	NDII02	83.1	NDIII02	82.3
ND14	NDI09	83.2	85.2	NDII02	83.1	NDIII09	82.5
ND15	NDI10	82.3	85.4	NDII07	83.5	NDIII10	83.3
ND16	NDI11	83.5	85.3-4*	NDII01	83.7	NDIII07	82.9
ND17	NDI04	83.3	85.4	NDII07	83.5	NDIII05	82.8
ND18	NDI04	83.3	85.4	NDII07	83.5	NDIII01	83.0
ND19	NDI11	83.5	85.3-4*	NDII08	83.4	NDIII07	82.9
ND20	NDI12	83.0	85.3	NDII07	83.5	NDIII01	83.0
ND21	NDI02	83.0	85.4	NDII01	83.7	NDIII11	83.1
ND22	NDI04	83.3	85.4	NDII07	83.5	NDIII12	82.8
ND23	NDI01	83.0	85.7	NDII09	84.0	NDIII01	83.0
ND24	NDI13	82.7	85.7	NDII01	83.7	NDIII01	83.0
ND25	NDI04	83.3	85.4	NDII10	83.5	NDIII01	83.0
ND26	NDI10	82.3	85.4	NDII07	83.5	NDIII01	83.0
ND27	NDI14	83.3	85.4	NDII07	83.5	NDIII01	83.0
ND28	NDI04	83.3	85.4	NDII07	83.5	NDIII07	82.9
ND29	NDI04	83.3	85.4	NDII11	83.5	NDIII01	83.0
ND30	NDI15	83.0	85.7	NDII11	83.5	NDIII01	83.0
ND31	NDI04	83.3	85.4	NDII12	83.7	NDIII01	83.0
ND32	NDI11	83.5	85.3-4*	NDII07	83.5	NDIII13	82.9
ND33	NDI16	83.8	85.4	NDII13	83.3	NDIII14	82.8
ND34	NDI04	83.3	85.4	NDII14	83.8	NDIII01	83.0
ND35	NDI17	83.2	85.5	NDII11	83.5	NDIII01	83.0
ND36	NDI18	none	85.4	NDII12	83.7	NDIII01	83.0

Table 5.6 (continued) *In silico* Tm values of ND1 amplicons from Doellman *et al.* (2011), where Tm11, Tm21 and Tm31 values are obtained by splitting the entire 633bp locus into three amplicons.

ND1 haplotype	NDI (1-241bp)	Tm11 (°C)	Tm12 (°C)	NDII (222-421bp)	Tm21 (°C)	NDIII (395-633bp)	Tm31 (°C)
ND37	NDI19	82.7	85.5	NDII15	83.3	NDIII01	83.0
ND38	NDI04	83.3	85.4	NDII11	83.5	NDIII05	82.8
ND39	NDI04	83.3	85.4	NDII07	83.5	NDIII15	83.3
ND40	NDI20	none	85.1	NDII01	83.7	NDIII07	82.9
ND41	NDI04	83.3	85.4	NDII16	83.1	NDIII01	83.0
ND42	NDI21	83.2	85.3-4*	NDII01	83.7	NDIII01	83.0
ND43	NDI22	83.1	85.8	NDII07	83.5	NDIII01	83.0
ND44	NDI21	83.2	85.3-4*	NDII17	83.9	NDIII01	83.0
ND45	NDI21	83.2	85.3-4*	NDII01	83.7	NDIII16	83.2
ND46	NDI21	83.2	85.3-4*	NDII18	83.5	NDIII01	83.0
ND47	NDI21	83.2	85.3-4*	NDII13	83.3	NDIII01	83.0
ND48	NDI23	82.9	84.9	NDII07	83.5	NDIII01	83.0
ND49	NDI23	82.9	84.9	NDII19	83.9	NDIII01	83.0
ND50	NDI23	82.9	84.9	NDII07	83.5	NDIII17	83.1
ND51	NDI23	82.9	84.9	NDII20	83.9	NDIII01	83.0
ND52	NDI23	82.9	84.9	NDII07	83.5	NDIII18	82.8
ND53	NDI23	82.9	84.9	NDII21	83.7	NDIII01	83.0
ND54	NDI24	82.8	84.7	NDII07	83.5	NDIII01	83.0
ND55	NDI25	82.7	85.2	NDII07	83.5	NDIII01	83.0
ND56	NDI26	82.5	85.5	NDII22	83.6-7*	NDIII01	83.0
ND57	NDI23	82.9	84.9	NDII23	83.7	NDIII01	83.0
ND58	NDI21	83.2	85.3-4*	NDII24	83.4	NDIII01	83.0
ND59	NDI27	82.7	85.2	NDII07	83.5	NDIII01	83.0
ND60	NDI04	83.3	85.4	NDII01	83.7	NDIII01	83.0
ND61	NDI27	82.7	85.2	NDII25	83.8	NDIII01	83.0
ND62	NDI23	82.9	84.9	NDII26	83.6	NDIII01	83.0
ND63	NDI23	82.9	84.9	NDII07	83.5	NDIII07	82.9
ND64	NDI28	none	85.0	NDII07	83.5	NDIII01	83.0
ND65	NDI27	82.7	85.2	NDII27	83.7	NDIII01	83.0
ND66	NDI23	82.9	84.9	NDII28	83.1	NDIII01	83.0
ND67	NDI29	83.0	85.3	NDII07	83.5	NDIII01	83.0
ND68	NDI27	82.7	85.2	NDII29	83.4	NDIII01	83.0
ND69	NDI23	82.9	84.9	NDII07	83.5	NDIII19	83.3

* These Tm values indicate the corresponding melting peak shows a square tip which is distinct from those with sharp tips of adjacent Tm values.

Based on the combination of the Tm values in Table 5.6, 52 out of the 69 haplotypes have their unique melting profiles, while 17 of them fell within seven Tm profiles so that in total 10 haplotypes of the entire ND1 locus could be missed by HRM analysis. The 17 haplotypes with shared Tm values included ND17 vs. ND22 vs. ND38, DN15 vs. ND18 vs. DN27 vs. ND29, ND20 vs. ND67, ND31 vs. ND60, ND49 vs. ND51, ND53 vs. ND57 and ND55 vs. ND59.

Within this case it is not possible to see if those indistinguishable haplotypes co-occur in some of the populations due to lack of information, so it cannot be estimated how they are likely to be distinguished within each population. However to gain additional insight into the likely mistake caused by missed detection in HRM analysis, the phylogenetic relationship among the ND1 haplotypes has been inferred and is shown in the maximum-likelihood tree in Figure 5.4. It is seen that in three cases the undistinguished haplotypes fall into different clades, which may lead to significant mistakes in phylogeographic analyses, while in the other four cases, the haplotypes sharing the same Tm values are in the same clade, which may only lead to minor mistakes.

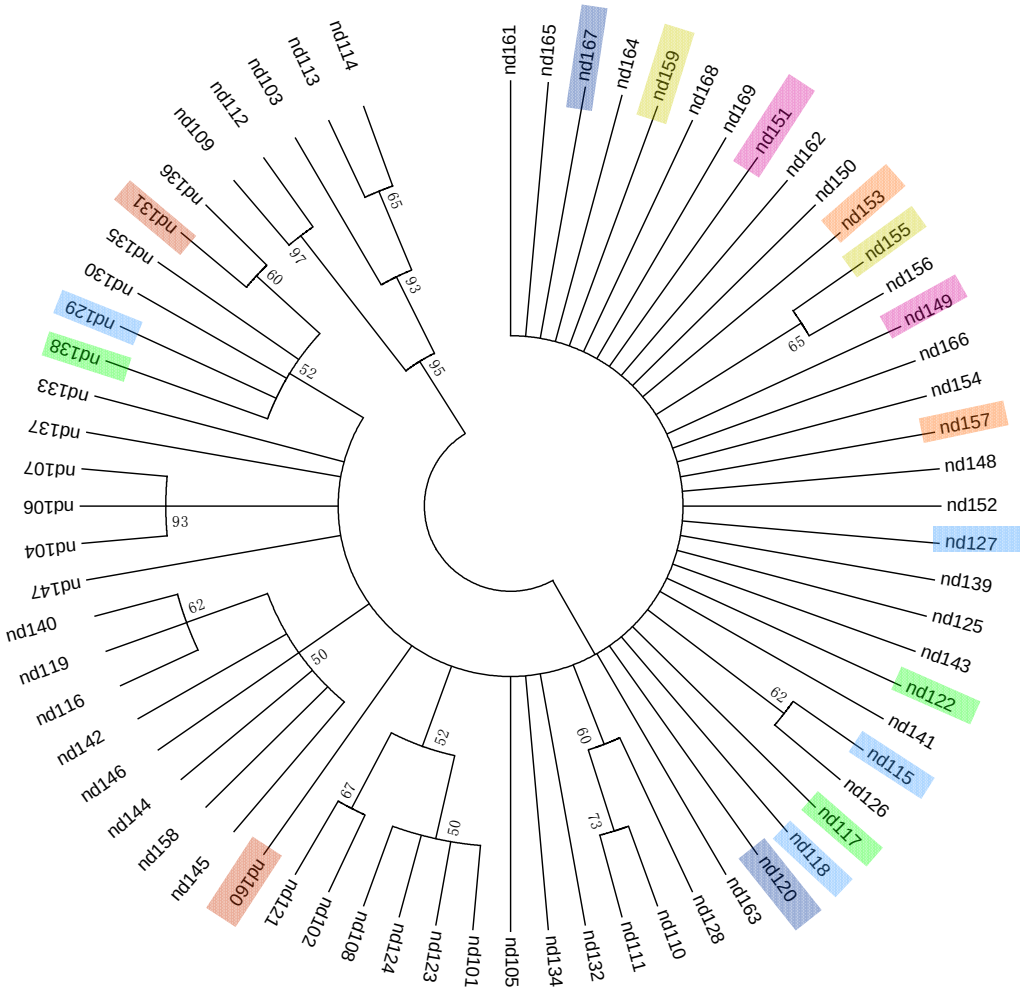


Figure 5.4 The inferred phylogenetic relationship among the ND1 haplotypes identified from *Littorina sp.* by Doellman *et al.* (2011). The maximum likelihood tree is obtained using MEGA 5.05 (Tamura *et al.* 2011) with support values using bootstrap (Felsenstein 1985) analysis with 100 replicates. The branches with support values less than 50% are collapsed. The haplotypes sharing the same Tm values in the *in silico* HRM analysis are labelled with the same colour.

Another mitochondrial locus, ND6, was also analyzed by Doellman and co-workers (2011) in their original study. The locus is not ideally suited

for HRM analysis because of its size (521bp) and because there is only a single peak. The template locus cannot be split into smaller amplicons because there is no conservative region across the detected haplotypes within the locus for internal primers to bind with.

5.3.4.3 HRM analysis with chloroplast loci on *Cedrela fissilis*

In a recent study of the Neotropical seasonal forest tree, *Cedrela fissilis* (Meliaceae), Garcia and co-workers (2011) analyzed 169 individual samples of the species from Brazil and Bolivia using the sequences of three chloroplast loci, trnS-trnG (716bp), trnT-trnL (1498bp) and psbB-psbT-psbN (689bp). In total five haplotypes were found with trnS-trnG, 16 haplotypes were found with trnT-trnL and two haplotypes were found with psbB-psbT-psbN. Indels were taken out of the analysis as Garcia and co-workers did in their original study.

HRM performance with trnS-trnG

The trnS-trnG locus was divided into two amplicons of 356bp and 382bp respectively with 22bp overlap, for the *in silico* HRM analysis. Both the amplicons rendered double melting peaks, the Tm values of which have

helped identifying all the five trnS-trnG haplotypes (Table 5.7).

Table 5.7 *In silico* Tm values of the trnS-trnG amplicons from Garcia *et al.* (2011), SGI (1-356bp) and SGII (335-716bp). Tm11 and Tm12 are obtained from SGI, Tm21 and Tm22 are from SGII.

Composite trnS-trnG haplotype	SGI	Tm11 (°C)	Tm12 (°C)	SGII	Tm21 (°C)	Tm22 (°C)
SG01	I01	77.5	80.3	II01	83.4	88.9
SG02	I02	77.5	79.9	II01	83.4	88.9
SG03	I03	77.3	80.3	II01	83.4	88.9
SG04	I01	77.5	80.3	II02	83.6	88.9
SG05	I04	77.3	80.3	II02	83.6	88.9

Although two of the SGI amplotypes, SGI03 and SGI04, cannot be distinguished by their Tm values, the corresponding composite haplotypes, SG03 and SG05 could be distinguished by their Tm difference in their SGII amplotypes. The detection rate is 100% in this case with trnS-trnG.

HRM performance with trnT-trnL

Due to the large size of trnT-trnL used in the study, the locus is split into four different amplicons. Because of the small number and the sparse distribution of the variable nucleotide sites therein (and large invariant sequence sections), the four amplicons have been designed to be small

and distinct from each other as possible without overlaps to increase HRM sensitivity. The four amplicons are recorded as TLI (1-352bp), TLII (601-840bp), TLIII (1005-1211) and TLIV (1230-1487) in Table 5.8, with the *in silico* Tm values for the corresponding amplotypes.

Table 5.8 *In silico* Tm values of trnT-trnL amplicons from Garcia *et al.* (2011), where the entire 1498bp locus was split into four discrete amplicons. TLIII shows a single melting peak while each of the other amplicons shows double peaks or a single peak with a shoulder.

trnT-trnL haplotype	TLI	Tm11 (°C)	Tm12 (°C)	TLII	Tm21 (°C)*	Tm22 (°C)	TLIII	Tm31 (°C)	TLIV	Tm41 (°C)	Tm42 (°C)
TL01	I01	74.6	78.2	II01	--	84.9	III01	85.8	IV01	81.6	85.0
TL02	I01	74.6	78.2	II01	--	84.9	III02	85.9	IV01	81.6	85.0
TL03	I01	74.6	78.2	II02	--	84.9	III01	85.8	IV01	81.6	85.0
TL04	I01	74.6	78.2	II01	--	84.9	III02	85.9	IV02	81.9	84.9
TL05	I02	74.6	78.2	II01	--	84.9	III02	85.9	IV01	81.6	85.0
TL06	I03	74.6	78.4	II01	--	84.9	III03	86.1	IV01	81.6	85.0
TL07	I04	74.2	77.8	II01	--	84.9	III02	85.9	IV01	81.6	85.0
TL08	I05	74.6	77.8	II01	--	84.9	III02	85.9	IV01	81.6	85.0
TL09	I06	74.6	78.0	II01	--	84.9	III02	85.9	IV01	81.6	85.0
TL10	I07	74.6	78.4	II01	--	84.9	III02	85.9	IV01	81.6	85.0
TL11	I08	74.6	78.4	II01	--	84.9	III02	85.9	IV01	81.6	85.0
TL12	I08	74.6	78.4	II03	83.5	85.0	III02	85.9	IV01	81.6	85.0
TL13	I08	74.6	78.4	II01	--	84.9	III02	85.9	IV03	80.5	85.0
TL14	I08	74.6	78.4	II01	--	84.9	III02	85.9	IV04	81.6	84.8
TL15	I08	74.6	78.4	II01	--	84.9	III02	85.9	IV05	81.6	84.7
TL16	I01	74.6	78.2	II01	--	84.9	III03	86.1	IV01	81.6	85.0

* The Tm21 corresponds to a melting shoulder in the melting profile of one of the TLII amplicons.

As shown in Table 5.8, the 16 trnT-trnL haplotypes are made of eight TLI amplotypes, three TLII amplotypes, three TLIII amplotypes and five TLIV amplotypes. Finally ten of the haplotypes have a unique combination of Tm values, while six of them share three distinct melting

profiles, i.e. TL01 vs. TL03, TL02 vs. TL05 and TL10 vs. TL11. In total three out of the 16 haplotypes would be missed if HRM analysis was used for haplotype detection with trnT-trnL in this case.

HRM performance with psbB-psbT-psbN

In total two haplotypes of the psbB-psbT-psbN locus were revealed by DNA sequencing in the original study. The 689bp locus turned out not required to be split into sub-regions as amplicons for HRM analysis, because the result showed three melting peaks and the Tm values of one of the peaks are sufficient to distinguish the two haplotypes. While one of the haplotypes showed Tm1=84.6°C, Tm2=86.7°C and Tm3=88.3°C, the other showed Tm1=84.6°C, Tm2=86.6°C and Tm3=88.3°C, with 0.1°C difference between them in Tm2.

As the geographic information is unavailable for us to assign all the haplotypes to the exact original populations, a phylogenetic analysis has been made for the trnT-trnL haplotypes (Figure 5.5), with which missing detection has happened.

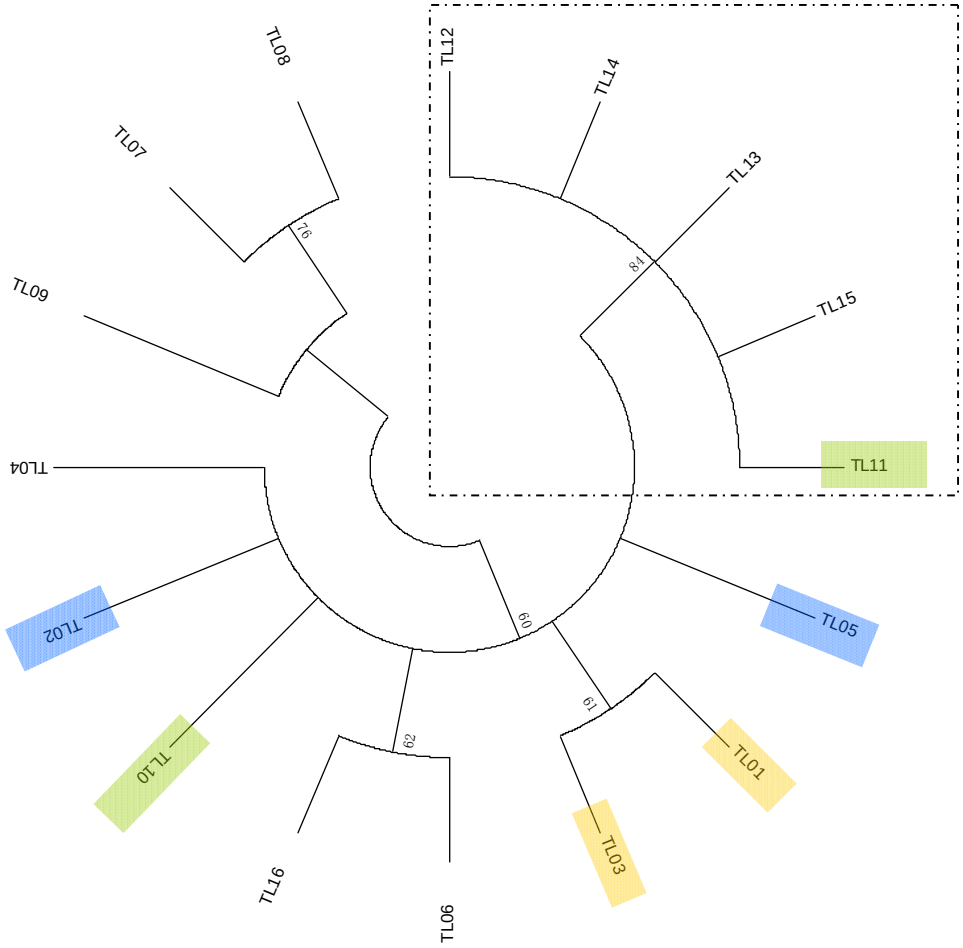


Figure 5.5 The inferred phylogenetic relationship among trnT-trnL haplotypes of *Cedrela fissilis* identified by Garcia el al. (2011). The phylogeny is constructed using the maximum likelihood method with the software MEGA5.05 (Tamura *et al.* 2011) with support values using bootstrap (Felsenstein 2005). Branches with support values less than 50% are collapsed. The clade group including TL11-15 and the clade group including all other haplotypes are considered to be the two clade groups as described by Garcia and co-workers in their original research, which do not co-occur in any single population.

It is seen that out of the three pairs of indistinguishable haplotypes, TL01 and TL03 are close to each other in genealogy and so are TL02 and TL05. Based on Garcia and co-workers' result, these haplotypes are within the

same clade group and could possibly co-occur in the same places. The missed detection of these haplotypes appeared to be unavailable; however the actual impact on phylogeographic analysis may be limited. The haplotypes TL10 and TL11 are in different clade groups, thus a case of missed detection between these would account for a potentially more significant lose phylogeographic information. However, the missing detection between TL10 and TL11 could be avoided if the HRM assay is performed in batches by populations, as they do not co-occur in the same place based on Garcia and co-workers' result.

5.3.4.4 HRM analysis with chloroplast loci on *Palicourea padifolia*

In a phylogeographic study of the distylous shrub, *Palicourea padifolia* (Rubiaceae) in Mexican cloud forests, Gutiérrez-Rodríguez and co-workers (2011) investigated 122 individuals across 22 populations using the chloroplast loci rpl32-trnL and trnS-trnG. In total 29 concatenated haplotypes were revealed by DNA sequencing, which are composed of 17 haplotypes of rpl32-trnL and 13 haplotypes of trnS-trnG. Based on our initial test, the trnS-trnG in this case is not suitable for HRM analysis because many of the polymorphic nucleotides are too close to one end of the DNA sequence and thus the primers for amplification

are hard to design. However the 657bp rpl32-trnL locus can be divided into three amplicons to facilitate HRM analysis, including RLI (1-201bp), RLII (201-490bp) and RLIII (490-657bp). The result of the in silico HRM analysis based on the three amplicons are shown in Table 5.9.

In total 13 out of the 17 rpl32-trnL haplotypes show their unique combinations of Tm values, while four of them share two melting profiles, i.e. RL05 and RL17 share the same Tm values and so do RL11 and RL15.Thus two out of the 17 haplotypes would be missed if HRM was used for haplotype detection in this case. Considering the pairwise distinction between the existent haplotypes, two indistinguishable haplotype pairs are seen out of the 136 (17×16/2) pairs of haplotypes, indicating the missing detection rate of 1.47%.

Table 5.9 *In silico* Tm values of rpl32-trnL amplicons from Gutiérrez-Rodríguez *et al.* (2011). Tm11 is obtained from RLI, Tm21 and Tm22 are from RLII, Tm31 and Tm32 are from RLIII. Especially, the double peaks for RLIII appear only with RLIII02 and only one peak appears for RLIII01 and RLIII03.

Composite rpl32-trnL haplotype	RLI	Tm11 (°C)	RLII	Tm21 (°C)	Tm22 (°C)	RLIII	Tm31 (°C)	Tm32 (°C)
RL01	RLI01	77.4	RLII01	74.3	76.9	RLIII01	--	73.1
RL02	RLI01	77.4	RLII02	74.4	77.3	RLIII01	--	73.1
RL03	RLI02	77.1	RLII01	74.3	76.9	RLIII01	--	73.1
RL04	RLI01	77.4	RLII03	74.3	77.5	RLIII01	--	73.1
RL05	RLI01	77.4	RLII04	74.4	77.0	RLIII01	--	73.1
RL06	RLI01	77.4	RLII05	74.7	76.9	RLIII01	--	73.1
RL07	RLI01	77.4	RLII06	73.9	76.9	RLIII01	--	73.1
RL08	RLI01	77.4	RLII07	74.3	77.1	RLIII01	--	73.1
RL09	RLI03	77.4*	RLII01	74.3	76.9	RLIII01	--	73.1
RL10	RLI04	77.8	RLII01	74.3	76.9	RLIII01	--	73.1
RL11	RLI01	77.4	RLII08	74.4	76.9	RLIII01	--	73.1
RL12	RLI01	77.4	RLII08	74.4	76.9	RLIII02	72.2	74.0
RL13	RLI01	77.4	RLII09	74.4	76.4	RLIII01	--	73.1
RL14	RLI01	77.4	RLII10	74.4	76.3	RLIII01	--	73.1
RL15	RLI01	77.4	RLII11	74.4	76.9	RLIII01	--	73.1
RL16	RLI05	77.6	RLII12	74.8	76.9	RLIII03	72.3	--
RL17	RLI01	77.4	RLII13	74.4	77.0	RLIII01	--	73.1

* The RLI amplicon RLI03 (contained in RL09) shows an extra melting shoulder, which distinguish it from other RLI amplotypes.

In the original work by Gutiérrez-Rodríguez and co-workers, the concatenated rpl32-trnL+trnS-trnG haplotypes corresponding to RL05 and RL17 were recorded as G and AC and those corresponding to RL11 and RL15 were recorded as V and Z. The haplotypes G and AC did not coexist in any of the studied populations, while the haplotypes V and Z co-occurred in one of the 22 populations.

In the population from Montebello (recorded as MB by the authors), five individuals were found to be haplotype V and one of haplotype Z. The haplotype Z is a rare haplotype and only found in this population, while haplotype V is a common haplotype found in two other populations. Based on phylogenetic analysis, haplotype V is one of the central haplotypes while haplotype Z is terminal, only two mutational steps from haplotype V. The two haplotypes would not be distinguished by HRM analysis based on the protocols described in Chapter 2. However, this would not lead to significant loss of information in phylogeographic analysis based on the phylogenetic relationship among the haplotypes (provided in the original literature).

5.3.4.5 HRM analysis with chloroplast loci on *Arenaria* species

Four species in the genus *Arenaria* as well as the other species, *Sagina caespitosa* in the Family of Caryophyllaceae were investigated by Westergaard and co-workers (Westergaard *et al.* 2011). Three individuals of A. humifusa and seven individuals of *A. longipedunculata* were sequenced with two chloroplast loci, rps16 and rps32-trnL.

HRM analysis with rps16

With the locus rps16, in total 568bp was aligned among all the individuals of the two *Arenaria* species. Three haplotypes were found in *A. humifusa*, varying in the combination of a single G/T mutation and a 4-bp indel, and two haplotypes were found in *A. longipedunculata*, varying by two SNPs (A/C and A/T) from each other. When the entire c.568 region was used as the amplicon for *in silico* HRM analysis, three melting peaks were generated for *A. humifusa* and four melting peaks were shown for *A. longipedunculata*.

The three haplotypes of *A. humifusa* are successfully distinguished by the *in silico* HRM analysis, where the Tm1 values for the three haplotypes are 79.7°C, 79.6°C and 79.4°C respectively while Tm2=82.2°C and Tm3=85.5°C for all the three haplotypes.

The two haplotypes of *A. longipedunculata* are distinguished by the second out of the four melting peaks, one with Tm2=80.3°C and the other haplotype with Tm2 unreadable because the peak nearly disappears. The Tm values of the other three peaks are the same for the two haplotypes (Tm1=79.5°C, Tm3=82.1°C, Tm4=85.1°C).

HRM analysis with rps32-trnL

With small single indels included for genotyping, two haplotypes and four haplotypes of rps32-trnL were found respectively from the two species via DNA sequencing by the original authors. Within the 975bp DNA region of the loci aligned among all the individuals of the two species, the variable nucleotides between 1 and 459bp are sufficient to identify all the haplotypes. Thus the c.459bp amplicon has been put into HRM analysis to investigate whether the haplotypes could be distinguished from one another.

The two haplotypes of *A. humifusa* vary in a single T/G mutation and two single nucleotide indels, which show triple melting peaks and are distinguished by the Tm values of the third peak (Tm3=82.8°C for haplotype 1 and Tm3=82.9°C for haplotype 2, while Tm1=78.8°C and Tm2=80.3°C for both haplotypes).

The four haplotypes of *A. longipedunculata* vary in different combinations of one A/G substitution and two indels. They all show double melting peaks corresponding to Tm1 and Tm3 found with *A. humifusa*, which distinguish them from the other species. The Tm1 values

could distinguish one of the four haplotypes from the others, while the other three share the same Tm values (Tm1=79.2°C for one haplotype and Tm1=79.0°C for the other three haplotypes, while Tm3=83.0°C for all the four haplotypes). The well distinguished haplotype vary from others by an A/G mutation, while the other three vary from each other by one or two indels.

5.3.4.6 Overall evaluation on the *in silico* HRM analysis with published data

Based on the above analyses with published data, the amplicon size, occurrence of multiple peaks, number of haplotypes that cannot be distinguished by HRM analysis and missing detection rate are summarized in Table 5.10 (next page).

Table 5.10 Summary of *in silico* HRM analyses with published sequence data. For each DNA locus, the amplicons used for HRM are listed. The size, number of melting peaks, haplotype number and missing detection rate are provided of each amplicon. The lines in bold are the haplotype detection information for the entire locus.

Locus (size in bp)	Amplicon size (bp)	No. melting peaks	No. haps	No. haps sharing Tms	No. missed haps	Frequency of missed haps
Bisconti *et al.* 2011	*Hyla sarda* (Hylidae)					
Cyt *b* (531bp)	296	2	19	4	2	0.11
	260	2	21	4	2	0.10
			35	**8**	**4**	**0.11**
ND1 (698bp)	287	1	21	18	12	0.57
	161	1	8	0	0	0.00
	294	1	21	17	12	0.57
			38	**17**	**11**	**0.29**
Doellman *et al.* 2011	*Littorina saxatilis* (Littorinidae)					
ND1 (633bp)	241	2	29	8	3	0.10
	199	1	29	24	17	0.59
	239	1	19	15	10	0.53
			69	**17**	**10**	**0.14**
Garcia *et al.* 2011	*Cedrela fissilis* (Meliaceae)					
trnS-trnG (716bp)	356	2	4	2	1	0.25
	381	2	2	0	0	0.00
			5	**0**	**0**	**0.00**
trnT-trnL (1498bp)	352	2	8	5	3	0.38
	240	*1	3	2	1	0.33
	207	1	3	0	0	0.00
	258	2	5	0	0	0.00
			16	**6**	**3**	**0.19**
Gutiérrez-Rodríguez *et al.* 2011	*Palicourea padifolia* (Rubiaceae)					
rpl32-trnL (657bp)	201	1	5	0	0	0.00
	290	2	13	5	3	0.23
	168	*1	3	0	0	0.00
			17	**4**	**2**	**0.12**
Westergaard *et al.* 2011						
Arenaria humifusa (Caryophyllaceae)						
rps16	568	3	**3**	**0**	**0**	**0.00**
rpl32-trnL	459	3	**2**	**0**	**0**	**0.00**
A. longipedunculata						
rps16	568	4	**2**	**0**	**0**	**0.00**
rpl32-trnL	459	2	**4**	**3**	**2**	**0.50**

* The amplicon shows a single melting peak for most alleles but shows double peaks for rare alleles.

It is seen from Table 5.10 overall missed detection rates per locus for *in silico* HRM analysis varied from zero to 50% of the existent haplotypes. However, in most cases this missed proportion was between 11 and 19%.

5.4 Discussion

5.4.1 HRM sensitivity to different mutation types

The *in silico* modeling of HRM sensitivity to different SNP mutation classes on the wild-type rps16*I* amplicon estimated the likely incidence of SNPs in rps16*I* to be 51.58% and 61.61% for class III and IV mutations respectively, compared to <2% for Class I and II mutations. This has confirmed that class I and class II SNPs are easier to detect than Class III and IV SNPs, which was proposed and partially verified by Liew and co-workers (2004). The same pattern is also shown by our *in vitro* results where the classes I and II SNPs can be detected by HRM analysis when occurring in an amplicon up to c.360bp (rps16*I*, 1 case of class II SNP) or c.390bp (rps16*II*, 1 case of class I SNP) at or above the 0.2°C threshold selected for identifying between-amplicon deviation (refer to Table 2.10-12 in Chapter 2 and Dang *et al.* 2012). In our haplotype dataset four Class III and two Class IV SNPs were observed, however only in one

case (a Class III G/C mutation between *II*05 and *II*11 of rps16*II*) did this mutation constitute the sole difference between composite haplotypes – and this was the only instance where *in vitro* HRM failed to distinguish between haplotypes of rps16.

The HRM sensitivity showed similar pattern across different classes of DNP mutations. When the two SNPs in the same mutant fall within the same SNP class, the same pattern was discovered that class I DNPs rendered the lowest missed detection rate (8.95%), while class II, III and IV provided higher missed detection rates (10.00%, 29.81% and 41.74% respectively). Most inter-class DNP mutations rendered significantly lower missing detection rates (2.35-9.14%), with only class III × class IV mutations showing an elevated missing detection rate (34.52%), which is in line with the high missed detection rate of class III and IV SNPs when considered separately (51.58% and 61.61% respectively). In general, DNPs appear to be more easily detected by HRM analysis than SNPs.

The present study has not investigated multiple site mutations (e.g. more than two nucleotides) within an amplicon. One reason is that the computational work would become time prohibitive as there would be exponentially increased number of mutants with three or more

substitutions. The second reason is that in many cases of phylogeographic studies, the haplotypes co-occurring within the same population or within the set of target populations vary from each other by only one or two substitutions per ~700bp (e.g. Gutiérrez-Rodríguez et al. 2011; Bisconti et al. 2011; Garcia et al. 2011), because usually the divergence between lineages is shallow over intra-specific divergence time scales. However, the studied populations can contain divergent lineages varying by as many as 5-8 substitutions within an amplicon of 350-700bp (e.g. in a case of the Arenaria species described in Chapter 2 and in the sea snail, *Littorina saxatilis* reported by Doellman et al. 2011), where the deep divergence may have been formed by pre-glacial maximum (i.e. Pleistocene) survival and separation.

The efficacy of *in vitro* HRM analysis in detecting multiple site mutations has already been verified in our case study of the *Arenaria* species, where up to 5-8 substitutions within an amplicon of 350-700bp were readily revealed by HRM assays. However, there is a need that future works of theoretical evaluation of missed detection rate should be done to understand the HRM sensitivity with more polymorphic nucleotide sites and all possible combinations of mutation types. Ideally multiple site mutations on up to 5 substitutions within an amplicon of 200-400bp

should be included for evaluation of HRM sensitivity, and this will be carried out in future works.

Also as introduced in section 5.1.2.1, indels are not examined in the present study. Future works could examine HRM sensitivity to indels of different length from 1bp up to 20bp of the kind in the *Arenaria* species (Chapter 2 and Dang *et al.* 2012) in order to evaluate the missed detection rate quantitatively. However, as with *in-silico* evaluation of multiple mutations, there is an exponential increase in the number of permutations and combinations that need to be considered when evaluating HRM sensitivity to different indel categories.

5.4.2 The effect of amplicon size on HRM efficacy

Traditionally it was considered preferable to use smaller amplicons in HRM analysis to achieve higher sensitivity. However the results in Table 5.3 showed that a 354bp amplicon with double melting peaks can perform as well as a 150bp amplicon does, as they both give the missed detection rate of 56% for SNPs and 35-36% for DNPs. This suggests the possibility of using larger amplicons with multiple melting peaks to achieve similar levels of screening sensitivity as small amplicons with single melting

peaks.

Nevertheless, and DNP mutations respectively (while all the amplicons of different sizes included in this analysis have well identified double melting peaks, there is a significant decrease in sensitivity when comparing the 100bp amplicon group to the 550bp amplicon group (Missed detection rates for SNPs increase from 26.92 to 69.71% over this range, while DNP sensitivity increase from 3.69 to 51.43%.) As a result, in terms of a general HRM protocol for identifying haplotypes at population level, it is desirable that selected DNA loci should be put into an *in silico* HRM analysis to establish the number of melting peaks that HRM will yield, and the theoretical rate of missed detection.

5.4.3 P*ost-hoc* HRM performance with the published data

In the previous sections the focus is on the probability of distinguishing two different allele amplicons by HRM analysis. However in realistic cases the most important concern is how many out of all the existent haplotypes can be detected by HRM analysis.

In the case where ND1 was used to genotype *Hyla sarda* by Bisconti *et al.*

(2011), 29% of the haplotypes would be missed by HRM analysis. The high missed detection is probably due to the fact that three of the amplicons all provide single melting peaks while two of them are nearly 300bp in length, which is too large in HRM analysis. It is clear in this case that when at least one of the amplicons generates multiple melting peaks, more haplotypes would be distinguished.

In terms of missed detection with each amplicon, most amplicons between 240bp and 300bp with double melting peaks achieved small missed detection frequency between 10% and 20%. It is improbable to make a comprehensive analysis of the size effect on detection rate with these realistic DNA sequences because they vary in melting peak numbers and the number of extant haplotypes. What we can conclude is that in the five published datasets analyzed in section 5.3.4, HRM could be applied for haplotype detection with an expected rate of c.10-20% haplotypes missed. However this can be improved by doing the assays in batches of single populations and by cross-checking between amplicons and between loci during posterior DNA sequencing.

Within a few of the cases in section 5.3.4 the *in silico* HRM analysis was able to achieve 100% detection rate, however the number of total existent

haplotypes is small (2 to 5), indicating that the high success rate may be only a sampling effect. In the cases above where more (16 to 69) haplotypes are present, the 10-20% missed detection frequency is more likely.

In the case where rpl32-trnL was used to genotype *A. longipedunculata* by Westergaard *et al.* (2011), two out of the four haplotypes (50%) would be missed by HRM analysis because they share the same Tm profile with one another haplotype. The three haplotypes are different only in minor simple sequence repeats, and the fact that HRM analysis is not sensitive to small variation in SSRs may explain the failure (see Chapter 2). This high failure frequency, however, may also be due in some part to the small sample effect, as only four haplotypes are present.

In section 5.3.4.2, the mitochondrial locus ND6 was given as an example to show that not all DNA loci widely used in phylogeographic studies are suitable for HRM analysis. However researchers can choose a locus that is favourable for HRM analysis with the help of the in silico HRM simulation in uMelt^{SM}, based on which primer/amplicon combinations can be evaluated as to their performance in HRM analysis.

5.4.4 Reducing missed detection by the cross-check strategy

For haplotype detection of each DNA locus in the present study, the posterior confirmation by DNA sequencing is performed under the cross-check strategy among the amplicons, i.e. whenever an individual sample shows a distinct Tm profile at one of the amplicons, the sample should be sequenced with the entire locus containing other amplicons. The strategy helps reduce the possibility of missed detection and also saves the workload of DNA sequencing. The cross-check strategy can also be conducted between DNA loci, i.e. in the case of Bisconti *et al.* (2011), whenever a sample is identified to carry a different haplotype of Cyt *b* by HRM analysis in a population, it should be sequenced with both Cyt *b* and ND1 although it may carry the same melting profile with other samples in the HRM analysis of ND1, and vice versa from ND1 to Cyt *b*. If this strategy had been used with HRM analysis in this study, no missed detection would have arisen and all the existent haplotypes from the mitochondrial genomes would have been revealed by HRM analysis. Compared to performing exhaustive DNA sequencing of 169 individuals with the two mtDNA loci, as done by Bisconti and co-workers' (2011), only 84 individuals need to be sequenced to reveal all of the 68 Cyt b-ND1 concatenated haplotypes.

Another post-HRM haplotype grouping strategy has been mentioned in both Chapter 2 and the present chapter, which is to designate the samples into hypothetical haplotypes after HRM assays within each population, rather than across all the populations before sending representative samples to be sequenced. The rationale behind the strategy lies in that the individuals are more likely to share haplotypes with others within the same population than those from different populations, which is usually true due to phylogeographic history. Considering the posterior probability, the individuals from the same population sharing the same melting profile in HRM assays are more likely to belong to the same haplotype than those from different populations separately. A similar rationale lies behind the strategy of performing cross-check between different amplicons and between different loci for sequencing validation, in that individuals with different haplotypes in one DNA region are less likely to share the same haplotype in the other DNA region, even if they share the same melting profile in HRM assays. The success of the above strategies has been demonstrated by our work on *A. ciliata* and *A. norvegica* (Chapter 2) and by the *in silico* HRM analysis based on the published data (section 5.3.4).

5.4.5 Missed detection rate: from theoretical assessment to realistic evaluation

In sections 5.3.1 and 5.3.3 the pairwise missed detection rate was evaluated to assess the effect of mutation classes and amplicon size. It is noted that the pairwise missed detection rate is different from the proportion of haplotypes that are missed in a given case study. There is not a linear relationship between the pairwise missed detection rate and the missed detection proportion, even the probability distribution of the latter cannot be easily determined when the pairwise missed detection rate is given. As a result, the latter missed detection proportion can only be evaluated in case studies, either using published data based on exhaustive sequencing, as we did in the present study, or using simulated population genetic data in computer, which could be done in future works. In the five case studies above, the proportion of missed haplotypes is mostly between 10% and 20%, and reduced to nearly zero when the HRM assay is performed in batches of single populations and cross-check is carried out during posterior sequencing. The theoretical pairwise missing detection rate is c.20% for SNP mutations and c.10% for DNP mutations, similar to the missed detection proportion in the above case studies, indicating the two parameters may be at the same order of magnitude, and

the theoretical pairwise missing detection rate could serve as an index of efficacy of HRM analysis in haplotype detection.

The pairwise missed detection rate of single class III (C/G) mutations ranges from 26.92% to 70.25% with the sequences of 100-550bp in the size effect evaluation. Considering the single class III mutations with the missed detection rate of 51.58% based on the *A. ciliata* sequences in the first part (see section 5.3.1 and Table 5.3 in section 5.3.3), the overall missed detection rate amplicon sizes between 100bp and 350bp is considered to be lower or similar to that in the *A. ciliata* case, where the HRM analysis has been demonstrated successful in haplotype detection. As the occurrence of multiple melting peaks is positively correlated with the amplicon size while the missed detection rate is acceptable, larger amplicons are allowed to be used for this aim. However, *in silico* simulation should be carried out before the DNA locus and amplicons are determined for haplotype detection, in order to choose those amplicons rendering multiple peaks to enhance the detection rate.

The successful detection rates evaluated in the present work are conservative estimates, because the performance of *in vitro* HRM analysis can be more sensitive than predicted by the *in silico* simulation by

uMelt[SM], which has been demonstrated by our work in Chapter 2, where the *in vitro* ΔTm values are usually greater than the in silico ones. Furthermore, the entire shape of the melting curve can now be considered aside from merely Tm values in new models of real-time PCR instruments, e.g. Rotor Gene from Qiagen, which provides more information of DNA polymorphism and higher sensitivity of haplotype screening. Our work on the *Arenaria* species with trnT-trnL also demonstrated that the DNA melting shape can help distinguish haplotypes with identical Tm values (Chapter 2). As a result, it is likely that *in vitro* HRM assays can now provide better performance than predicted in the present work.

5.5 Conclusion

Based on the evaluation with both simulated and published realistic DNA sequences, HRM analysis is demonstrated as a useful tool of haplotype detection in phylogeographic research. The 10-20% missing detection probability of HRM analysis can partly be overcome by the three strategies during posterior sequencing validation. While not all DNA loci are suitable for HRM analysis, *in silico* simulation can be used before the *in vitro* work to minimize the missing detection rate, by choosing the

amplicons with multiple melting peaks and lowest theoretical missing detection rate. This new technology thus presents many opportunities for detection of unknown haplotypes in future phylogeographic studies.

Chapter 6

General discussion

6.1 The prospect of applying HRM analysis in phylogeographic research

6.1.1 Empirical evaluation of HRM sensitivity

In Chapter 2 of the present study, High-Resolution Melting (HRM) analysis based on real-time PCR as a method for haplotype detection has been applied in phylogeographic studies of three species in the family of Caryophyllaceae, including *Arenaria ciliata*, *A. norvegica* and *Minuartia recurva*. Two chloroplast DNA loci, rps16 intron and trnT-trnL intergenic spacer (which are c.750bp and c.640bp respectively for the *Arenaria* species and c.690bp and c.540bp respectively for *M. recurva*), have been tested with HRM analysis. Because the optimal amplicon size for informative HRM analysis should be less than c.350bp, as explained in Chapter 2 and 5, each locus was split into two amplicons for HRM analysis. Based on the protocol used in the case of *Arenaria* species, HRM analysis has been demonstrated to be able to reveal 19 out of 20 haplotypes of rps16 and all of the 24 haplotypes of trnT-trnL. In the case of *M. recurva*, HRM analysis was able to distinguish all of the four haplotypes of rps16 and three out of eight (or five if variation in SSRs is not considered) haplotypes of trnT-trnL. Posterior DNA sequencing of

two random samples in each putative haplotype revealed only one new haplotype of rps16 with the species *A. ciliata* and one new haplotype of trnT-trnL with *M. recurva*. In summary, HRM analysis has been demonstrated as a useful tool for haplotype detection in these case studies, where nearly all the sampled individuals can be designated to corresponding haplotypes, without need of being exhaustively sequenced, an as such the efficiency of haplotype screening at the PCR stage is significantly improved.

However, small (1-2bp) DNA variation in SSRs has emerged not to be as amenable to HRM analysis as substitutions of the same sizes, as demonstrated both in the case of trnT-trnL with *Minuartia recurva* in Chapter 2 and in the case of rpl32-trnL with *Arenaria longipedunculata* based on the *in silico* HRM analysis with Westergaard and co-workers' data (2011) in Chapter 5. Although polymorphisms in SSRs are usually removed in phylogenetic analysis because they are considered to be evolutionarily labile and homoplasious characters, potentially providing misleading information (Small *et al.* 1998), they are sometimes used as informative signatures showing generic variation at or under the population level (Provan *et al.* 2001; Borsch & Quandt 2009). The lower sensitivity of HRM to variation in SSRs may thus prove as a limitation to

its use in a broader scope.

Also it should be kept in mind that compared to DNA sequencing, HRM analysis provides an indirect method of genotype identification, which inherently involves a small but consistent possibility of missed detections in each case. For further use and possible improvement of the method, it is demanded that the possibility of missed detection should be quantitatively assessed.

6.1.2 Quantitative evaluation of HRM sensitivity

The theoretical possibility of missed detection in HRM analysis has been quantitatively evaluated in Chapter 5. With help of the software uMelt^{SM} developed by Dwight *et al.* (2011), it is possible to put both actually recorded and computer-generated DNA sequences into *in silico* HRM analysis and to examine how they are likely to be discriminated by their melting curves on computer. In the present study described in Chapter 5, the differentiated amenability of four classes of single nucleotide substitutions to HRM analysis, and the effects of both amplicon size and number of melting domains on HRM sensitivity have been quantitatively examined, to understand the general likelihood that two alleles of the

same DNA locus cannot be distinguished by HRM analysis.

From the *in silico* assessment based on the 352bp rps16*I* amplicon, the missed detection rates for class I (A/G or C/T) and II (A/C or G/T) substitutions were significantly lower than class III (C/G) and IV (A/T) substitutions, and while the missed detection rates for class III (C/G) mutations were 51.58% for SNPs and 12.87% for DNPs, the overall missed detection rates were 20.48% for SNPs and 9.73% for DNPs.

It has also been shown that amplicon size (100-550bp) is positively correlated to missed detection rate (26.92-69.71% for C/G SNPs and 16.77-35.09% for C/G DNPs) when all the amplicons carry double melting peaks. Considering that with the 352bp rps16*I* template a missed detection rate of 51.58% for C/G SNPs turned out to correspond to around a 10-20% total missed rate (based on the GC content in the template), the amplicons smaller than 350bp in the above analysis are all expected to yield a lower missed detection rate, which may also be acceptable in a case study comparable to the *Arenaria* case.

It should be noted that the present *in silico* evaluation of missed detection is a preliminary work, and further analysis is needed to make a

comprehensive evaluation considering all possible types of mutations occurring within different amplicons. First, all the four classes of substitutions should be considered for evaluation of HRM sensitivity to a given amplicon. Second, indels (including those in SSRs) with different lengths should be taken into account. Third, both substitutions and indels at multiple (possibly 2-5 as usually seen at intra-specific level) nucleotide sites should be included as possible mutations within a given amplicon. Finally, with the aim of comparing amplicons of different sizes in their amenability to HRM analysis, ideally 100-1000 random sequences should be generated for each amplicon size to facilitate a test that provides more accurate estimation. A significantly higher computational capacity, as well as considerable research time, is needed to achieve the above goals.

Also based on available DNA sequences from published phylogeographic studies, where exhaustive sequencing was carried out to reveal the haplotype identity of every individual sample, *in silico* HRM analysis was performed to demonstrate how this new technique may help distinguish the extant haplotypes and to estimate how many of the haplotypes would have been missed by HRM analysis. Based on the result in Chapter 5, typically 10-20% of all the extant haplotypes may be missed in HRM analysis. However, this figure can range from zero to

50% in each specific case, depending on the specific amplicons used in the original study and the actual types of variation occurring in the populations.

6.1.3 Possible improvements of HRM sensitivity

The strategies implemented in the current protocols of HRM analysis has already reduced possibility of missed detection, mainly by cross-checking between amplicons of the same DNA locus and between two or three loci used in the study, which has been discussed in Chapter 5 (section 5.5.4). Also *in silico* evaluation of a candidate DNA locus is recommended to be performed before it is used in HRM analysis, which will provide information on how many melting peaks may be generated, how amenable the locus is to being split into smaller amplicons and ultimately what the likely rate of missed detection might be. Such *a-priori* assessments will certainly help researchers speed up and improve their success rate in HRM analysis.

Also with recent advances in HRM analysis, the overall shape of melting curves aside from merely Tm values can also help distinguish different amplicons. As described in Chapter 5 (section 5.5.5), this new function is

implemented in a variety of new models of real-time PCR instruments, including Rotor Gene from Qiagen and LightCycler 480 II from Roche. With help from this new function in HRM analysis, the success rate in haplotype detection will increase (bearing in mind the significant additional discriminatory value that emerged in this study from manual observation of melt curve variation as shown in Chapter 2, section 2.4.1.4), making the technique more promising to be widely applied in phylogeographic research.

Aside from the above, there is a possibility to make innovative improvements in technology, by combining HRM analysis with multiplex PCR or techniques of DNA fragmentation, where a higher number of shorter PCR products may be obtained in each assay and subjected to high resolution melting analysis simultaneously. The higher number of PCR products enable researchers to conduct cross-checks and thus to reduce missed detections, while the smaller sizes can inherently improve HRM sensitivity.

6.2 Phylogeographic history of the studied species in Europe and in Ireland

Based on the new haplotype detection protocol with HRM analysis developed in the present study, unexpected high level of polymorphisms of chloroplast DNA have been found from the two *Arenaria* species (Chapter 3). Meanwhile, the results suggested that the two previously recognized sister species do not fall into distinct monophyletic lineages, but rather, *A. norvegica* have been proved to construct two distinct sub-clades of *A. ciliata*, while the latter species consist of five deeply divergent haplotype clades, which are estimated to have diverged one to four million years ago. Even within each clade, more than ten haplotypes were identified which are estimated to have diverged 90-250 thousand years ago (in clade I) or 250-520 thousand years ago (in clade II and III). The fact that in many populations both clade I and clade II (or III) haplotypes are found to coexist indicates that the establishment of the populations may involve multiple colonization events, while the harbouring of unique haplotypes *e.g.* in Ireland and in Shetland implies their survival in the localities for a longer time than thought before. The result suggests that the Irish populations of *A. ciliata* may have survived the last ice age *in situ* rather than having immigrated after the end of

Pleistocene (c. 12, 000 years ago), implying an ice age refugium on or near Ben Bulben in northwest Ireland.

However, the study described in Chapter 4 has shown a different phylogeographic pattern in *M. recurva*. The genetic diversity of the species is significantly lower than that in the *Arenaria* species, as in total nine (or six if variation in SSRs is not considered) haplotypes were revealed from the populations sampled across its European distribution, which fall within two distinct clades. The divergence among haplotypes within each clade has been estimated to be 240-400 thousand years ago in clade I and 170-280 thousand years ago in clade II, which is comparable to the divergence time within each clade of the *Arenaria* species. However, across the distribution of *M. recurva* from Ireland, the Iberian Peninsula, the Pyrenees and the Alps, each population was found to contain only one haplotype (or at most two if variation in SSRs is taken into account), except in the Balkans each of the two populations was found to harbour two distinct haplotypes. The result suggests a glacial refugium in the Balkans region and a recent dispersal of the species across its majority distribution in continental Europe and Ireland. However, there is a disagreement between the phylogeny and the geographic distribution of the identified haplotypes. While the Waterford

population in South Ireland shares the same haplotype with those in the north side of Spain and in the Pyrenees, the Kerry population in southwest Ireland shared the haplotype with those in the north side of Spain and in the Alps. Nevertheless, a closer relationship in indicated between the Irish populations taken as a whole and the Iberian populations.

The results based on both *Arenaria* species and *M. recurva* have suggested a closer relationship between the Irish populations and the Iberian populations. While the *Arenaria* data have provided evidence that the populations on Ben Bulben may have survived the last ice age after they were established as early as 150-250 thousand years ago, the *Minuartia* data was insufficient to date the establishment of the populations in Waterford and in Kerry, although a similar time period of establishment is also possible.

Chapter 7

Bibliography

von Ahsen N, Wittwer CT, Schütz E (2001) Oligonucleotide melting temperatures under PCR conditions: nearest-neighbor corrections for Mg(2+), deoxynucleotide triphosphate, and dimethyl sulfoxide concentrations with comparison to alternative empirical formulas. *Clinical Chemistry*, **47**, 1956–1961.

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: *Second International Symposium on Information Theory* (eds Petrov BN, Csàki F), pp. 267–281. Akademiai Kiado, Budapest.

Angiosperm Phylogeny Group III (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society*, **161**, 105–121.

Avise JC (2000) *Phylogeography: the history and formation of species*. Harvard University Press, Cambridge, MA.

Avise JC, Arnold J, Ball RM, *et al.* (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, **18**, 489–522.

Bermingham E, Moritz C (1998) Comparative phylogeography: concepts and applications. *Molecular Ecology*, **7**, 367–369.

Bettin O, Cornejo C, Edwards PJ, Holderegger R (2007) Phylogeography of the high alpine plant *Senecio halleri* (Asteraceae) in the European Alps: in situ glacial survival with postglacial stepwise dispersal into peripheral areas. *Molecular Ecology*, **16**, 2517–2524.

Bisconti R, Canestrelli D, Colangelo P, Nascetti G (2011) Multiple lines of evidence for demographic and range expansion of a temperate species (*Hyla sarda*) during the last glaciation. *Molecular Ecology*, **20**, 5313–5327.

Blake RD, Bizzaro JW, Blake JD, *et al.* (1999) Statistical mechanical simulation of polymeric DNA melting with MELTSIM. *Bioinformatics*, **15**, 370–375.

Borsch T, Quandt D (2009) Mutational dynamics and phylogenetic utility of noncoding chloroplast DNA. *Plant Systematics and Evolution*, **282**, 169–199.

Clement M, Posada D, Crandall KA (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, **9**, 1657–1659.

Colhoun EA, Dickson JH, McCabe AM, Shotton FW (1972) A middle midlandian freshwater series at Derryvree, Maguiresbridge, County Fermanagh, Northern Ireland. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character. Royal Society (Great Britain)*, **180**, 273–292.

Dang X-D, Kelleher CT, Howard-Williams E, Meade CV (2012) Rapid identification of chloroplast haplotypes using High Resolution Melting analysis. *Molecular Ecology Resources*, **12**, 894–908.

Dixon CJ, Schönswetter P, Vargas P, Ertl S, Schneeweiss GM (2009) Bayesian hypothesis testing supports long-distance Pleistocene migrations in a European high mountain plant (*Androsace vitaliana*, Primulaceae). *Molecular Phylogenetics and Evolution*, **53**, 580–591.

Doellman MM, Trussell GC, Grahame JW, Vollmer SV (2011) Phylogeographic analysis reveals a deep lineage split within North Atlantic *Littorina saxatilis*. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character. Royal Society (Great Britain)*, **278**, 3175–3183.

Dong C, Vincent K, Sharp P (2009) Simultaneous mutation detection of three homoeologous genes in wheat by High Resolution Melting analysis and mutation surveyor. *BMC Plant Biology*, **9**, 143.

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, **19**, 11–15.

Dupanloup I, Schneider S, Excoffier L (2002) A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology*, **11**, 2571–2581.

Dwight Z, Palais R, Wittwer CT (2011) uMELT: prediction of high-resolution melting curves and dynamic melting profiles of PCR products in a rich web application. *Bioinformatics*, **27**, 1019–1020.

Excoffier L (2004) Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Molecular Ecology*, **13**, 853–864.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Flanders J, Jones G, Benda P, *et al.* (2009) Phylogeography of the greater horseshoe bat, *Rhinolophus ferrumequinum*: contrasting results from mitochondrial and microsatellite data. *Molecular Ecology*, **18**, 306–318.

Fu Y-X (1996) New statistical tests of neutrality for DNA samples from a population. *Genetics*, **143**, 557–570.

Garcia MG, Silva RS, Carniello MA, *et al.* (2011) Molecular evidence of cryptic speciation, historical range expansion, and recent intraspecific hybridization in the Neotropical seasonal forest tree *Cedrela fissilis* (Meliaceae). *Molecular Phylogenetics and Evolution*, **61**, 639–649.

Green PR (2007) *Minuartia recurva* found in Co. Waterford. *BSBI News*, **104**, 4.

Guggisberg A, Mansion G, Kelso S, Conti E (2006) Evolution of biogeographic patterns, ploidy levels, and breeding systems in a diploid-polyploid species complex of Primula. *The New Phytologist*, **171**, 617–632.

Guindon S, Delsuc F, Dufayard J-F, Gascuel O (2009) Estimating maximum likelihood phylogenies with PhyML. In: *Bioinformatics for DNA Sequence Analysis* (ed Posada D), pp. 113–137. Humana Press, Totowa, NJ.

Gutiérrez-Rodríguez C, Ornelas JF, Rodríguez-Gómez F (2011) Chloroplast DNA phylogeography of a distylous shrub (*Palicourea padifolia*, Rubiaceae) reveals past fragmentation and demographic expansion in Mexican cloud forests. *Molecular Phylogenetics and Evolution*, **61**, 603–615.

Hall TA (1999) BioEdit:a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, **41**, 95–98.

Harpending HC (1994) Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Human Biology*, **66**, 591–600.

Herrmann MG, Durtschi JD, Bromley LK, Wittwer CT, Voelkerding KV (2006) Amplicon DNA melting analysis for mutation scanning and genotyping: cross-platform comparison of instruments and dyes. *Clinical Chemistry*, **52**, 494–503.

Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, **111**, 147–164.

Huguet JM, Bizarro CV, Forns N, *et al.* (2010) Single-molecule derivation of salt dependent base-pair free energies in DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 15431–15436.

Jakob SS, Blattner FR (2006) A chloroplast genealogy of hordeum (poaceae): Long-term persisting haplotypes, incomplete lineage sorting, regional extinction, and the consequences for phylogenetic inference. *Molecular Biology and Evolution*, **23**, 1602–1612.

Jalas J, Suominen J (1983) *Atlas Florae Europaeae: Distribution of Vascular Plants in Europe, Volume 6: Caryophyllaceae (Alsinoideae and Paronychioideae)*. Committee for Mapping the Flora of Europe, Helsinki.

Jalas J, Suominen J (1986) *Atlas Florae Europaeae: Distribution of Vascular Plants in Europe, Volume 7: Caryophyllaceae (Silenoideae)*. Committee for Mapping the Flora of Europe, Helsinki.

Kelleher CT, Hodkinson TR, Kelly DL, Douglas GC (2004) Characterisation of chloroplast DNA haplotypes to reveal the provenance and genetic structure of oaks in Ireland. *Forest Ecology and Management*, **189**, 123–131.

Kingman JFC (1982) On the genealogy of large populations. *Journal of Applied Probability*, **19**, 27–43.

Liew M, Pryor R, Palais R, *et al.* (2004) Genotyping of single-nucleotide polymorphisms by high-resolution melting of small amplicons. *Clinical Chemistry*, **50**, 1156–1164.

Lyon E, Wittwer CT (2009) LightCycler technology in molecular diagnostics. *The Journal of Molecular Diagnostics*, **11**, 93–101.

Mader E, Lohwasser U, Börner A, Novak J (2010) Population structures of genebank accessions of *Salvia officinalis* L. (Lamiaceae) revealed by high resolution melting analysis. *Biochemical Systematics and Ecology*, **38**, 178–186.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

Mast AR, Kelso S, Richards AJ, *et al.* (2001) Phylogenetic relationships in *Primula* L. and related genera (Primulaceae) based on noncoding chloroplast DNA. *International Journal of Plant Sciences*, **162**, 1381–1400.

Mitchell FJG (2006) Where Did Ireland's Trees Come From? *Biology & Environment: Proceedings of the Royal Irish Academy*, **106**, 251–259.

Mitchell F, Ryan M (1997) *Reading the Irish Landscape*. Town House, Dublin.

Monis PT, Giglio S, Saint CP (2005) Comparison of SYTO9 and SYBR Green I for real-time polymerase chain reaction and investigation of the effect of dye concentration on amplification and DNA melting curve analysis. *Analytical Biochemistry*, **340**, 24–34.

Pauchard A, Kueffer C, Dietz H, *et al.* (2009) Ain't no mountain high enough: plant invasions reaching new elevations. *Frontiers in Ecology and the Environment*, **7**, 479–486.

Pauls SU, Lumbsch HT, Haase P (2006) Phylogeography of the montane caddisfly Drusus discolor: evidence for multiple refugia and periglacial survival. *Molecular Ecology*, **15**, 2153–2169.

Petit RJ, Aguinagalde I, de Beaulieu J-L, *et al.* (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*, **300**, 1563–1565.

Poland D (1974) Recursion relation generation of probability profiles for specific-sequence macromolecules with long-range correlations. *Biopolymers*, **13**, 1859–1871.

Posada D (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253–1256.

Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends in Ecology & Evolution*, **16**, 142–147.

R Development Core Team (2011) R: A Language and Environment for Statistical Computing.

Raymond M, Rousset F (1995) An exact test for population differentiation. *Evolution*, **49**, 1280–1283.

Reed GH, Wittwer CT (2004) Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. *Clinical Chemistry*, **50**, 1748–1754.

Ririe KM, Rasmussen RP, Wittwer CT (1997) Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. *Analytical Biochemistry*, **245**, 154–160.

Rogers AR (2004) Lecture Notes on Gene Genealogies: the Theoretical Mismatch Distribution. University of Utah, Salt Lake City.

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, **9**, 552–569.

Rohlf FJ (1973) Algorithm 76. Hierarchical clustering using the minimum spanning tree. *The Computer Journal*, **16**, 93–95.

Schneider S, Excoffier L (1999) Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: Application to human mitochondrial DNA. *Genetics*, **152**, 1079–1089.

Shaw J, Lickey EB, Beck JT, *et al.* (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany*, **92**, 142–166.

Shaw J, Lickey EB, Schilling EE, Small RL (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany*, **94**, 275–288.

Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF (1998) The tortoise and the hare: choosing between noncoding plastome and nuclear adh sequences for phylogeny reconstruction in a recently diverged plant group. *American Journal of Botany*, **85**, 1301–1315.

Smith BL, Lu C-P, Alvarado Bremer JR (2010) High-resolution melting analysis (HRMA): a highly sensitive inexpensive genotyping alternative for population studies. *Molecular Ecology Resources*, **10**, 193–196.

Sosa V, Ruiz-Sanchez E, Rodriguez-Gomez FC (2009) Hidden phylogeographic complexity in the Sierra Madre Oriental: the case of the Mexican tulip poppy *Hunnemannia fumariifolia* (Papaveraceae). *Journal of Biogeography*, **36**, 18–27.

Studer B, Jensen LB, Fiil A, Asp T (2009) "Blind" mapping of genic DNA sequence polymorphisms in *Lolium perenne* L. by high resolution melting curve analysis. *Molecular Breeding*, **24**, 191–199.

Synge FM, Wright HEJ (1969) The Würm Ice Limit in the West of Ireland. In: *Quaternary Geology and Climate* (ed Wright HE). National Academy of Sciences, Washington, D. C.

Taberlet P, Fumagalli L, Wust-Saucy A-G, Cosson J-F (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology*, **7**, 453–464.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tajima F, Nei M (1984) Estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution*, **1**, 269–285.

Tamura K, Peterson D, Peterson N, *et al.* (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution*, **28**, 2731–2739.

Teacher AGF, Garner TWJ, Nichols RA (2009) European phylogeography of the common frog (*Rana temporaria*): routes of postglacial colonization into the British Isles, and evidence for an Irish glacial refugium. *Heredity*, **102**, 490–496.

Templeton AR, Crandall KA, Sing CF (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. cladogram estimation. *Genetics*, **132**, 619–633.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Tindall EA, Petersen DC, Woodbridge P, Schipany K, Hayes VM (2009) Assessing high-resolution melt curve analysis for accurate detection of gene variants in complex DNA fragments. *Human Mutation*, **30**, 876–883.

Tutin TG, Burges NA, Chater AO, *et al.* (1993) *Flora Europaea, Volume 1: Psilotaceae to Platanaceae*. Cambridge University Press, Cambridge.

Tøstesen E, Liu F, Jenssen T-K, Hovig E (2003) Speed-up of DNA melting algorithm with complete nearest neighbor properties. *Biopolymers*, **70**, 364–376.

Valcárcel V, Vargas P, Feliner GN (2006) Phylogenetic and phylogeographic analysis of the western Mediterranean *Arenaria* section Plinthine (Caryophyllaceae) based on nuclear, plastid, and morphological markers. *Taxon*, **55**, 297–312.

Valente LM, Savolainen V, Vargas P (2010) Unparalleled rates of species diversification in Europe. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character. Royal Society (Great Britain)*, **277**, 1489–1496.

Vamosi SM, Heard SB, Vamosi JC, Webb CO (2009) Emerging patterns in the comparative analysis of phylogenetic community structure. *Molecular Ecology*, **18**, 572–592.

Vossen RH a M, Aten E, Roos A, den Dunnen JT (2009) High-resolution melting analysis (HRMA): more than just sequence variant screening. *Human Mutation*, **30**, 860–866.

Walker K, Howard-Williams E, Meade CV (2010) The distribution and ecology of *Arenaria norvegica* Gunn. in Ireland. *Irish Naturalists' Journal*, In press.

Webb DA (1983) The flora of Ireland in its European context. *Journal of Life Sciences of the Royal Dublin Society*, **4**, 143–160.

Wei X-X, Yang Z-Y, Li Y, Wang X-Q (2010) Molecular phylogeny and biogeography of *Pseudotsuga* (Pinaceae): insights into the floristic relationship between Taiwan and its adjacent areas. *Molecular Phylogenetics and Evolution*, **55**, 776–785.

Westergaard KB, Alsos IG, Popp M, *et al.* (2011) Glacial survival may matter after all: nunatak signatures in the rare European populations of two west-arctic species. *Molecular Ecology*, **20**, 376–393.

Wingfield RTR (1995) A model of sea-levels in the Irish and Celtic seas during the end-Pleistocene to Holocene transition. *Geological Society, London, Special Publications*, **96**, 209–242.

Wittwer CT (2009) High-resolution DNA melting analysis: advancements and limitations. *Human Mutation*, **30**, 857–859.

Wittwer CT, Herrmann MG, Moss AA, Rasmussen RP (1997) Continuous fluorescence monitoring of rapid cycle DNA amplification. *BioTechniques*, **22**, 130–131, 134–138.

Wittwer CT, Reed GH, Gundry CN, Vandersteen JG, Pryor RJ (2003) High-resolution genotyping by amplicon melting analysis using LCGreen. *Clinical Chemistry*, **49**, 853–860.

Wolfe KH, Li W-H, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 9054–9058.

Woolley SM, Posada D, Crandall K a (2008) A comparison of phylogenetic network methods using computer simulation. *PloS One*, **3**, e1913.

Wu S-B, Franks TK, Hunt P, *et al.* (2009) Discrimination of SNP genotypes associated with complex haplotypes by high resolution melting analysis in almond: implications for improved marker efficiencies. *Molecular Breeding*, **25**, 351–357.

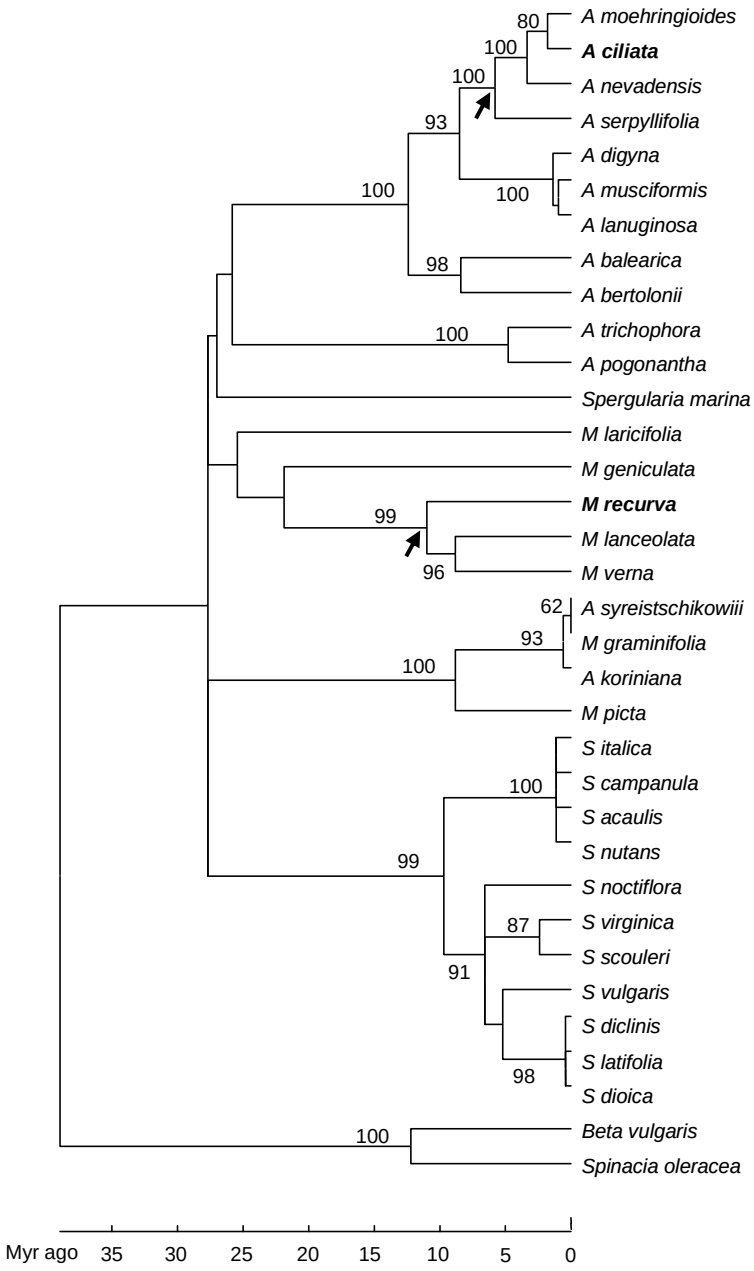
Wyse Jackson MB, Parnell JAN (1987) A biometric study of the *Arenaria ciliata* L. complex Caryophyllaceae. *Watsonia*, **16**, 373–382.

Yeramian E, Schaeffer F, Caudron B, Claverie P, Buc H (1990) An optimal formulation of the matrix method in statistical mechanics of one-dimensional interacting units: Efficient iterative algorithmic procedures. *Biopolymers*, **30**, 481–497.

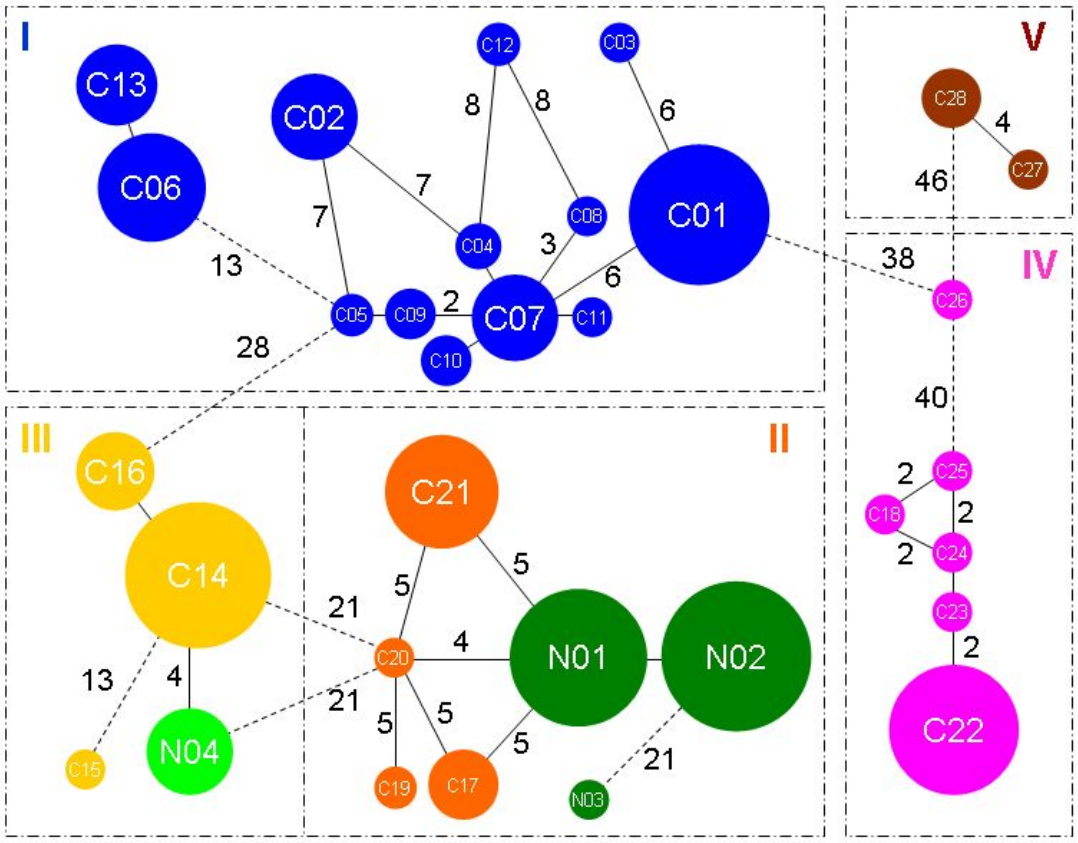
Appendix

Appendix Figure 1

Appendix Figure 2



Appendix Figure 1 Linearized maximum-likelihood phylogenetic tree of species mostly in the genera *Arenaria*, *Minuartia* and *Silene* in the family of Caryophyllaceae based on chloroplast matK sequences, with *Beta vulgaris* and *Spinacia oleracea* as outgroups. The sequences of *M. recurva* and *M. verna* were obtained from our own works, while the other sequences are from Valente *et al.* (2010). Based on the calibration by Valente and co-workers, the divergence time between *A. serpyllifolia* and *A. ciliata* was dated to 6-10 million years ago, while the divergence between *M. recurva* and *M. verna* was 11-18 million years ago. The two arrows indicate the above mentioned divergence events. The tree was regenerated using MEGA 5.05 (Tamura *et al.* 2011) with 100-replicate bootstrap test. Support values <50 are not shown. The scale at the bottom shows the lower estimation of divergence times before the present.



Appendix Figure 2 Minimum spanning network showing the genealogy among the haplotypes found in *Arenaria ciliata* (C01-C28) and *A. norvegica* (N01-N04). The network was based on the analysis within the software Arlequin 3.5.1.3 (Excoffier & Lischer 2010) and then drawn by hand. Variation in SSRs amidst the DNA sequences was included in the analysis, which is different from the construction of statistical parsimony networks as described in Chapter 3. In accordance with the results in Chapter 3, the network here shows that all the haplotypes are divided into five clades and the *A. norvegica* haplotypes are included in clades II and III.