Genetic diversity studies of *Trifolium* species from the extremes of the UK

by

SERENE HARGREAVES

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Biosciences
College of Life and Environmental Sciences
The University of Birmingham
September 2010
ABSTRACT

Crop wild relatives have been identified as ecologically and economically important plant genetic resources but are often a neglected resource. The recognition of the need for their specific conservation and their value for future use has been strengthened by the Convention on Biological Diversity and the International Treaty on Plant Genetic Resources for Food and Agriculture, both of which have been ratified by the UK.

This thesis provides a detailed view of the ecological, geographic and genetic background to three crop wild relative species, *Trifolium dubium*, *T. pratense* and *T. repens*, of which the latter two are amongst some of the most economically important legume species in the UK. Assessments of ecogeography, amplified fragment polymorphism and single nucleotide polymorphism markers were employed to investigate the distribution of variation in these species across the UK, including outlying island sites. Based on this information it was possible to look for isolation by distance in populations in UK; identify areas containing unique variation; assess the conservation importance of island sites surrounding the UK and speculate on the causes of the observed patterns of diversity.

Conservation recommendations were based on the cumulative data from this research to identify how the recommendations change with an increased focus on genetic diversity. These results provide insights into the use of different types of background information when setting conservation plans in widespread species, contributing to the development of conservation strategies for widespread species in general.
ACKNOWLEDGMENTS

I would like to start by expressing my gratitude to my supervisors Dr. Brian Ford-Lloyd and Dr. Nigel Maxted for their suggestions, support and for the opportunities they have given me throughout my project.

I would like to thank the staff at IGER who made me welcome during my stay and who helped me learn the ropes in the lab and in the field; Mike Abberton, Sue Dalton, Charlotte Deakin, Galina Latypova, Kirsten Skøt, Leif Skøt and Ian Thomas. In particular I would like to acknowledge Kirsten Skøt for kindly allowing me to learn in her spotless genomics lab, and to Leif Skøt for helpful discussions and for providing *T. pratense* EST sequences.

Many thanks are due to all my colleagues in S208 and W204 for their friendship, good advice and encouragement, especially Joana Magos Brehm, Cristina Sousa Correia, Marianne Mitchell and Fay Hughes for guiding me through the start of my PhD and Andrea Harper, Jennifer Preston and Claire Teeling for helpful discussions and support throughout the rest.

Most importantly I would like to thank my family for their encouragement and for their faith in me throughout this project and David Jacoby for his help with the collecting mission to the Isles of Scilly and with his patience and constant support.

Finally, I would also like to acknowledge the BBSRC and the BSBI who supplied the financial support for this project.
TABLE OF CONTENTS

Chapter 1. General Introduction ................................................................. 1
  1.1 What is diversity and why conserve it? .............................................. 1
  1.2 Biodiversity crisis and UK commitments .......................................... 1
  1.3 Genetic resources and crop wild relatives ......................................... 2
  1.4 Background to the current study ....................................................... 4
  1.5 A general introduction to genetic diversity studies .......................... 6
      1.5.1 SSRs ......................................................................................... 7
      1.5.2 Focus on Amplified Fragment Length Polymorphism (AFLP) ...... 8
      1.5.3 Focus on SNPs ......................................................................... 10
  1.6 Quantifying genetic differences ....................................................... 31
  1.7 Gene flow and genetic diversity ....................................................... 32
      1.7.1 The consequence of introgression in conservation .................. 32
      1.7.2 Measurements of gene flow .................................................... 33
      1.7.3 Landscape genetic analysis and its implications for population genetics 35
  1.8 Mainland versus island populations ................................................. 39
      1.8.1 Previous studies of diversity in mainland versus island populations 40
  1.9 Thesis outline ................................................................................... 41

Chapter 2. An ecogeographic background to three *Trifolium* species in the UK 43
  2.1 Ecogeography as a tool for conservation analyses .......................... 43
  2.2 Aims of ecogeographic survey ....................................................... 43
  2.3 Target taxa ....................................................................................... 44
  2.4 *Trifolium* genetics ......................................................................... 45
      2.4.1 *Trifolium repens* genetic diversity ........................................ 46
      2.4.2 *Trifolium pratense* genetic diversity ...................................... 48
  2.5 *Trifolium dubium* Sibth .................................................................. 51
  2.6 *Trifolium pratense* L. ..................................................................... 56
  2.7 *Trifolium repens* L. ........................................................................ 61
  2.8 Target Island Site - Isles of Scilly .................................................... 66
      2.8.1 Environmental history ............................................................ 67
      2.8.2 Conservation ........................................................................... 69
2.8.3 Presence of species ................................................................. 70
2.9 Target Island Site - Outer Hebrides ........................................... 71
  2.9.1 Environmental history .......................................................... 73
  2.9.2 Conservation ......................................................................... 74
  2.9.3 Presence of species ............................................................... 75
2.10 Target Island Site - Shetland Isles .......................................... 76
  2.10.1 Environmental history .......................................................... 78
  2.10.2 Conservation ......................................................................... 79
  2.10.3 Presence of target species ..................................................... 80
2.11 Target Island Site - Skye, Inner Hebrides ................................ 81
  2.11.1 Environmental history .......................................................... 83
  2.11.2 Conservation ......................................................................... 85
  2.11.3 Presence of species ............................................................... 85
2.12 Target Island Site - St Kilda, Outer Hebrides ......................... 86
  2.12.1 Environmental history .......................................................... 88
  2.12.2 Conservation ......................................................................... 90
  2.12.3 Presence of target species ..................................................... 91
2.13 Current threat status ............................................................... 92
2.14 Conservation status on the basis of ecogeographic information 95

Chapter 3. Genetic structure of three *Trifolium* species in the UK using AFLP 97

3.1 Introduction ................................................................................ 97
3.2 Methods .................................................................................... 100
  3.2.1 Population sampling ............................................................... 100
  3.2.2 Amplified Fragment Length Polymorphism ........................... 110
  3.2.3 Population genetic analysis .................................................... 112
  3.2.4 Isolation by distance .............................................................. 115
  3.2.5 Environmental impact on genetic diversity ......................... 116
  3.2.6 Island versus mainland heterozygosity ................................. 119
3.3 Results - *Trifolium dubium* ..................................................... 120
  3.3.1 Descriptive population genetics ............................................. 120
  3.3.2 Population subdivision .......................................................... 120
  3.3.3 Genetic relationships among populations ............................. 121
  3.3.4 Isolation by Distance ............................................................ 123
3.3.5 Environmental impacts on patterns of genetic diversity .................................. 124
3.3.6 Island versus mainland heterozygosity ............................................................. 126

3.4 Results - *Trifolium pratense* ................................................................................... 127
3.4.1 Descriptive population genetics ....................................................................... 127
3.4.2 Population subdivision ..................................................................................... 128
3.4.3 Genetic relationships among populations ......................................................... 128
3.4.4 Isolation by distance ......................................................................................... 130
3.4.5 Environmental impacts on patterns of genetic diversity .................................. 131
3.4.6 Island versus mainland heterozygosity ............................................................. 133

3.5 Results - *Trifolium repens* ...................................................................................... 134
3.5.1 Descriptive population genetics ....................................................................... 134
3.5.2 Population subdivision ..................................................................................... 134
3.5.3 Genetic relationships among populations ......................................................... 136
3.5.4 Isolation by distance ......................................................................................... 139
3.5.5 Environmental impacts on patterns of genetic diversity .................................. 140
3.5.6 Island versus mainland heterozygosity ............................................................. 142

3.6 Discussion ............................................................................................................... 143
3.6.1 The spatial pattern of diversity in *Trifolium* species across the UK................. 143
3.6.2 *T. repens* – genetic exchange in a crop-wild species complex ......................... 147
3.6.3 Priority areas for conservation .......................................................................... 149

Chapter 4. Applying SNPs to conservation questions; case study of *T. pratense* ........ 153
4.1 Introduction ............................................................................................................. 153

4.2 Methods .................................................................................................................. 156
4.2.1 Candidate gene selection .................................................................................. 156
4.2.2 Leaf material and DNA extraction ................................................................... 161
4.2.3 Primer design .................................................................................................... 161
4.2.4 Trial set ............................................................................................................. 161
4.2.5 Locus amplification ......................................................................................... 161
4.2.6 Sequencing ....................................................................................................... 162
4.2.7 Genetic diversity analysis – programs used for *T. pratense* analysis .............. 163

4.3 Results .................................................................................................................... 168
4.3.1 DNA extraction and sequencing....................................................................... 168
4.3.2 Intraspecific diversity and neutrality analysis .................................................. 171
4.3.3 Genetic differentiation ................................................................. 174
4.3.4 Isolation by distance ...................................................................... 174
4.3.5 Analysis of molecular variation ...................................................... 176
4.3.6 Geographic haplotype diversity ...................................................... 177
4.3.7 Linkage Disequilibrium ................................................................. 183
4.3.8 Analysis involving all loci ............................................................... 184
4.4 Discussion ...................................................................................... 187

Chapter 5. Are widespread *Trifolium* species really of least concern? ......... 193

5.1 Wider implications for widespread species conservation .................... 194
5.2 A comparison of the efficiency of AFLP and SNP markers in *Trifolium*
    pratense ............................................................................................... 195
5.3 Limitations ....................................................................................... 199
5.4 Future work ..................................................................................... 200

References ............................................................................................... 203

Appendix 1. CTAB DNA extraction protocol .............................................. 241
Appendix 2: Composition of stock solutions .............................................. 242
Appendix 3. Distance matrices .................................................................. 243
Appendix 4: TIGR transcript assemblies database ...................................... 246
    A4.1 TA1010_57577 transcript assembly ................................................ 246
    A4.2 TA1548_57577 transcript assembly .................................................. 247
    A4.3 TA3078_57577 transcript assembly .................................................. 247
    A4.4 TA3695_57577 transcript assembly .................................................. 247
    A4.5 TA555_57577 transcript assembly .................................................... 248
    A4.6 TA989_57577 transcript assembly .................................................... 248
    References ......................................................................................... 249

Appendix 5: Primer choice ....................................................................... 250
    BB910055 *T. pratense* primer design .................................................... 250
    BB910055 *T. dubium* primer design ...................................................... 250
    BB925852 *T. pratense* primer design .................................................... 253
    BB925852 *T. dubium* primer design ...................................................... 253
    TA1548_57577 *T. pratense* primer design (using BB915621) ................. 254
    TA1548_57577 *T. dubium* primer design (using BB915621) ................. 254
    TA3695_57577 *T. pratense* primer design (using BB916074) ................. 255
LIST OF FIGURES

Figure 1.1. Principal component analysis of white clover populations and landraces ............... 5
Figure 1.2. SNP basics ............................................................................................................. 12
Figure 1.3. Decision making flow chart for SNP detection using candidate genes. .............. 26
Figure 1.4. The increase in the availability of sequence data on the World Wide Web .......... 30
Figure 2.1. 10km grid square distribution map of target species across the UK ...................... 50
Figure 2.2: *Trifolium dubium*. ............................................................................................. 54
Figure 2.3 World distribution of *Trifolium dubium* .............................................................. 55
Figure 2.4: *Trifolium pratense* ............................................................................................ 59
Figure 2.5 World distribution of *Trifolium pratense* .......................................................... 60
Figure 2.6: *Trifolium repens* ................................................................................................ 64
Figure 2.7 World distribution of *Trifolium repens* ............................................................... 65
Figure 2.8. Climate of St. Mawgans, average 1971-2000 ....................................................... 67
Figure 2.9. Climate of Stornoway, average 1971-2000 ........................................................... 72
Figure 2.10. Climate of Lerwick, average 1971-2000 ............................................................. 77
Figure 2.11. Climate of Tirree, average 1971-2000. .............................................................. 82
Figure 3.1 Locations of the 11 *T. dubium* collection sites across the UK ......................... 107
Figure 3.2 Locations of the 16 *T. pratense* collection sites across the UK ....................... 108
Figure 3.3 Locations of the 27 *T. repens* collection sites across the UK ....................... 109
Figure 3.4 *T. dubium* UPGMA dendrogram ........................................................................ 122
Figure 3.5 Mantel tests of isolation by distance ........................................................................ 123
Figure 3.6 Pairwise average kinship coefficient of 165 *T. dubium* individuals ................. 124
Figure 3.7 *T. pratense* UPGMA dendrogram ..................................................................... 129
Figure 3.8 Mantel tests of isolation by distance ........................................................................ 130
Figure 3.9 Pairwise average kinship coefficient of 231 *T. pratense* individuals ............... 131
Figure 3.10 *T. repens* UPGMA dendrogram .......................................................................... 137
Figure 3.11 STRUCTURE analysis of AFLP data for *T. repens* across the UK ............... 138
Figure 3.12 Mantel tests of isolation by distance ..................................................................... 139
Figure 3.13 Pairwise average kinship coefficient of 602 *T. repens* individuals ............... 140
Figure 4.1 Pictorial representations of loci amplified from *Trifolium pratense* ............... 170
Figure 4.2 $F_{ST}$ values for five loci assessed in the study .................................................. 174
Figure 4.3 Mantel tests of isolation by distance plotted using transformed $F_{ST}$ .................. 175
Figure 4.4 T21 haplotype analysis ................................................................. 178
Figure 4.5 T52 haplotype analysis ................................................................. 179
Figure 4.6 T55 haplotype analysis ................................................................. 180
Figure 4.7 T74 haplotype analysis ................................................................. 181
Figure 4.8 T96 haplotype analysis ................................................................. 182
Figure 4.9 Linkage disequilibrium heatmaps for five loci assessed in *T. pratense* .......... 183
Figure 4.10 Population structure of *T. pratense* populations ......................... 184
Figure 4.11 STRUCTURE results imaged in DISTRUCT for *K*=3 ...................... 185
Figure 5.1. Pearson correlation coefficient (*r*) estimates ................................. 196
Figure 5.2. Distribution of polymorphism information content (PIC) data .............. 198
## LIST OF TABLES

Table 2.1 IUCN categories system, version 3.1 (IUCN, 2001) ................................................................. 93
Table 2.2 IUCN summary of five criteria (A-E) ....................................................................................... 94
Table 3.1 Collection site data for all populations ..................................................................................... 104
Table 3.2 Primer optimization ................................................................................................................. 111
Table 3.3: AFLP error rate ....................................................................................................................... 112
Table 3.4 Variable sets used in dbRDA analyses ...................................................................................... 117
Table 3.5 Correlation ($r^2$) between Stornoway/St Kilda weather station data ........................................ 118
Table 3.6 Population genetic analysis of *T. dubium* populations ............................................................ 120
Table 3.7 AMOVA of variation across 165 *T. dubium* individuals ........................................................... 121
Table 3.8 *T. dubium* redundancy based analysis ..................................................................................... 125
Table 3.9 *T. dubium* island versus mainland heterozygosity ............................................................... 126
Table 3.10 Population genetic analysis of *T. pratense* populations ....................................................... 127
Table 3.11 *T. pratense* redundancy based analysis ............................................................................... 128
Table 3.12 *T. pratense* island versus mainland heterozygosity ............................................................ 132
Table 3.13 Population genetic analysis of *T. repens* ............................................................................. 133
Table 3.14 AMOVA of variation across 602 *T. repens* individuals ........................................................ 136
Table 3.15 *T. repens* redundancy based analysis .................................................................................... 138
Table 3.16 Estimated membership coefficients for each *T. repens* population ........................................ 141
Table 3.17 *T. repens* island versus mainland heterozygosity ............................................................... 142
Table 4.1 Blast search for putative *T. pratense* drought stress associated genes .................................. 157
Table 4.2 Components of PCR mix to a final volume of 25µl per individual reaction .............................. 162
Table 4.3 Annealing temperature for each locus after PCR optimisation .............................................. 162
Table 4.4 Polymorphisms found in the five loci for *T. pratense* individuals ........................................ 168
Table 4.5 Summary polymorphism statistics for all five *T. pratense* loci .............................................. 171
Table 4.6 Neutrality tests in 5 *Trifolium pratense* loci ............................................................................ 173
Table 4.7 AMOVA for the five *T. pratense* loci assessed in the study ................................................ 176
Chapter 1. **General Introduction**

1.1 **What is diversity and why conserve it?**

The definition of biodiversity put forward during the CBD is the most generally accepted, defining biodiversity as the “variability among living organisms from all sources…and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems” (UNCED, 1992). The final part of this definition, “within species diversity” or genetic diversity will be the main focus of this assessment, and one that is generally less well studied than the ecosystem and species level of diversity.

Many authors underline the link between the level of genetic diversity and the persistence of populations (Frankel & Soulé, 1981; Gilpin & Soulé, 1986). Low variation can lead to increased extinction risk and lower fitness through the effects of genetic drift and inbreeding depression in the short term and the inability to adapt to a changing environment in the longer term (Höglund, 2009). Whilst genetic variation is not the only cause of population fluctuations, extinctions or species persistence, when in conjunction with demographic factors, genetic diversity studies provide an important insight into species persistence and viability. These insights are thus imperative to the conservation of species in light of the increasing risks to species from threats such as habitat loss and fragmentation, and in particular due to the potential effects of a changing climate. The loss of diversity, and thus the loss of the genes, species and ecosystems that provide the basis for future adaptation, will have major economic and social costs (Heywood, 1995).

1.2 **Biodiversity crisis and UK commitments**

The loss of the world’s diversity is occurring on a vast scale, however the rate of loss is ultimately pitted against the few resources available for conservation. Only by assessing
where, why and how species survive can we start to assess what conservation action is required, in order to create more efficient and effective conservation plans.

The UN Convention on Biological Diversity (CBD) held in 1992 represented a significant landmark in the field of conservation biology. The CBD brought the importance of biodiversity loss to the attention of governments, research communities and the public by defining the challenge and highlighting the cooperation necessary to reverse the loss of the world’s resources. By ratifying the CBD the parties, including the UK, committed themselves to specific targets in biodiversity conservation, in particular to “achieve by 2010 a significant reduction of the current rate of biodiversity loss” (CBD, 2002). Within the overall 2010 framework, target 3.1, to “promote the conservation of genetic diversity” and target 8.2 to “maintain biological resources that support sustainable livelihoods, local food security and health care” will be addressed in this study.

1.3 GENETIC RESOURCES AND CROP WILD RELATIVES

In addition to the CBD, the UK has ratified the International Treaty on Plant Genetic Resources for Food and Agriculture (FAO, 2001), which outlines, amongst others, the objective to “survey and inventory plant genetic resources for food and agriculture, taking into account the status and degree of variation in existing populations, including those that are of potential use.”

The term “plant genetic resources” encompasses all plant genetic material, which are often described in terms of their actual or potential use for agriculture. Interbreeding with wild relatives provides a new source of genetic variation both for the improvement of crops and the development of new varieties (Jain, 1975; Schoen & Brown, 1993; Tanksley & McCouch, 1997). IPGRI (1993) defines plant genetic resources as “genetic material of plants which is of value as a resource for the present and future generation of people”. Therefore, in addition to
crops themselves, related crop or wild species that can be interbred with crop species are of particular value (Maxted *et al.*, 2006; Heywood *et al.*, 2007).

Harlan and de Wet (1971) attempted to quantify the degree of relatedness between wild relatives and their associated crop with the gene pool concept, which enabled priorities for conservation to be inferred from the proximity to socio-economically important species. Harlan and de Wet (1971) propose three gene pools, from a primary gene pool containing the cultivated and wild forms of the crop species to the tertiary gene pool where gene transfer is very difficult or impossible. With the increasing use of biotechnological techniques in gene transfer these groupings have become less distinct, with Maxted and Hawkes (1997) proposing the ‘gene sea’ concept, visualising the gene sea as a network of interrelational gene pools. This gene sea concept is important for use in determining conservation priorities, and highlights the significance of conserving total genetic diversity, including wild species, for future potential use ( Tanksley & McCouch, 1997).

Crop wild relative (CWR) conservation has been the subject of many reviews (e.g. Maxted, 2003; Meilleur & Hodgkin, 2004), with their prioritisation in conservation policies increasing significantly in recent years. Indeed the FAO’s Global Plan of Action for the Conservation and Sustainable Utilisation of Plant Genetic Resources for Food and Agriculture (1996) highlights CWRs in priority 4: ‘Promoting in situ conservation of wild crop relatives and wild plants for food production.’ The genetic material present in CWRs, and even in less closely related taxa, can contribute to both the long term persistence of domesticated species through crop improvement and, through natural genetic exchange, contribute to the productivity of agro-ecosystems (Meilleur & Hodgkin, 2004). Wild relatives deriving from different environmental conditions provide a wide pool of resources, including pest and virus resistance, resistance to abiotic stresses, increased yield and improved quality (Hajjar & Hogkin, 2007; Maxted & Kell, 2008). The values of such wild relatives can be high with
Prescott-Allen and Prescott-Allen (1986) calculating their value to the North American economy as $340 million a year. Pimentel et al. (1997) in a more recent assessment of the input of genetic resources to the North American economy estimated their value to be $20 billion per year based on increased crop yields.

The species that will be assessed in this study include two crop species *Trifolium repens* L. and *T. pratense* L., as well as a wild relative of *T. repens*, *T. dubium* Sibth. (Bulińska-Radomska, 2000), which has not, until the acknowledgement of CWR importance, been considered of high conservation value.

### 1.4 Background to the Current Study

In light of the UKs objectives for genetic conservation, finding more effective sampling strategies have become of high importance. A collaboration between researchers at the Millennium Seed Bank (the Royal Botanic Gardens, Kew), Horticultural Research International (Wellesbourne), the Institute of Grassland and Environmental Research (Aberystwyth) and the University of Birmingham assessed the relationship between ecogeographic (ecological and genetic) and genetic diversity data in various UK socio-economically important species, attempting to determine the importance of ecogeography in defining the patterns of genetic diversity (Maxted et al., unpublished results). Of the eight species assessed, three species; *Beta vulgaris* subsp. *maritima* (L.) Arcang., *Lolium perenne* L. and *T. repens*, were found to have no correlation between genetic diversity and ecogeographic factors. Therefore, for these taxa, it was concluded that sampling based on ecogeographic factors would not necessarily capture maximum genetic diversity; instead prior genetic diversity assessments would be required.

However, a recent assessment of *T. repens* populations on St Kilda, an island group found 64km west of the Outer Hebrides, found that the populations were highly genetically
distinct from one reference mainland population (Hirano, 2005). Using AFLPs, five populations from St Kilda were compared to two landraces and one ‘wild’ population. In accordance with other published genetic diversity studies in this species, _T. repens_ showed high levels of within population variation, with little difference in the levels of within population variation between mainland and island populations. Principle component analysis indicated the uniqueness of the island group in terms of genetic diversity (see Figure 1.1).

![Principal component analysis of white clover populations and landraces based on 351 AFLP markers.](image)

The study highlighted populations that have avoided the introgression that has affected much of the UK’s _T. repens_ population, whether through geographic or agricultural isolation or both. This undoubtedly provides a valuable insight in terms of defining a reservoir for future breeding programs and conservation activities in the UK. It was proposed to expand this study to include species with a different cultivation history and to more island groups.
surrounding the UK, to assess both the distribution of variation and the impact of genetic pollution on the inherent patterns of genetic diversity in these species and in related taxa.

All three target taxa used in this study, *T. dubium*, *T. pratense* and *T. repens*, are widespread across the UK and as such often occur within protected area limits; however due to their ubiquitous nature do not often specifically feature in management plans. The assessment of the distribution of diversity will identify areas of distinct variation, allowing the prioritisation of areas for *in situ* conservation and *ex situ* collections, as well as highlighting the need for specific management in selected sites of these otherwise common species.

1.5 A GENERAL INTRODUCTION TO GENETIC DIVERSITY STUDIES

The use of molecular markers to assess and quantify genetic variation is one of the core themes in molecular ecology and conservation genetics, with the continuing development of molecular markers helping to increase the impact of molecular genetics on ecology. The continuous adoption of new technologies, from isozymes, RAPDs and RFLPs through to AFLPs and microsatellites, has generated thousands of genetic diversity studies focusing on population genetics and conservation (for reviews see for example Parker *et al.*, 1998; Ouborg *et al.*, 1999; Sunnucks, 2000; Vignal *et al.*, 2002; Avise, 2004). Appropriate measurements of diversity can elucidate the type, extent and partitioning of variation in a species, which provides valuable information on species identity, diagnostics and relationships for conservation and breeding strategies (Westman & Kresovich, 1997).

Traditionally, analyses of diversity centred on quantifying the expression of phenotypic traits (famously studied by Mendel and his peas) and have expanded to include biochemical, and now DNA-based methodologies (Ouborg *et al.*, 1999). The ever increasing number of DNA-based techniques have overtaken other methodologies due to limitations centred around the environmental influence in phenotypic analyses and a need for higher levels of resolution
than that determined by biochemical analyses (Höglund, 2009). Although biochemical analyses marked a large improvement in diversity studies, there is the assumption that protein and isozyme variation directly reflect heritable changes, whereas DNA-based methodologies allow a direct analysis of the genome (Ouborg et al., 1999).

DNA-based methodologies are based on sequence variation and often use molecular markers, with some techniques able to detect differences as small as one base pair between two genotypes. Among the many molecular marker techniques; amplified fragment length polymorphism (AFLP) and microsatellites stand out as the most commonly used today, with single nucleotide polymorphisms (SNPs) identified as a potential molecular marker for the future (Vignal et al., 2002). Of the many different markers and techniques for the analysis of genetic diversity, no one technique is universally correct, with usage dependant on the research question, data required, and resource availability (Ouborg et al., 1999).

1.5.1 Microsatellites

Microsatellite markers have remained one of the most popular methods in conservation genetic methods since their adoption (e.g. Jarne & Lagoda, 1996; Ellegren, 2004). They consist of tandem repeats of nucleotides, which are found in a relatively high frequency in most taxa, and can be known as simple sequence repeats (SSRs), variable number tandem repeats (VNTRs) and short tandem repeats (STRs) (Selkoe & Toonen, 2006). The flanking regions of microsatellites are used as primer binding sites to amplify the region. It is the variation in the number of repeats of these microsatellites that provides the variation between individuals required for analysis, with di-, tri- and tetra-nucleotide repeats the most frequently used for analysis (Li et al. 2002). With the high level of mutation in these regions (between $10^{-2}$ and $10^{-6}$) these regions provide a high number of allelic differences between individuals, which is highly beneficial to genetic studies (Ellegren, 2004; Selkoe & Toonen, 2006).
Microsatellite markers provide a very useful tool for conservation genetic analysis, due to both a high information content of each locus – perhaps consisting of up to 20 alleles (Jarne & Lagoda, 1996) and their codominant nature. The power of a large multi-locus microsatellite study is perhaps unrivalled, compared to the most commonly used genetic marker methods (Selkoe & Toonen, 2006), with Gerber (2000) showing that 159 AFLP markers had the equivalent power as just under 6 multi-locus microsatellites in parentage analysis.

However, while powerful, one of the major downfalls of microsatellites is in their detection, finding microsatellites is both time-consuming and expensive. This method is also complicated by the possibility of homoplasy (identically sized but differing alleles), which can overinflate estimates of gene flow and lower estimates of allelic diversity (Selkoe & Toonin, 2006). In addition, due to the specificity or the PCR reaction, once found these primers are almost entirely species specific (Sunnocks, 2000). Thus in nonmodel organisms, and where studies are resource-constrained, this method is less often used.

1.5.2 **FOCUS ON AMPLIFIED FRAGMENT LENGTH POLYMORPHISM (AFLP)**

AFLP is a more recent approach to restriction fragment analysis incorporating the power of PCR (Vos et al., 1995). DNA is digested by two restriction enzymes before the addition of adapters that incorporate the primer sites for PCR. The fragments can be selectively amplified by extending the primers into the original fragment, and performing multiple PCRs.

The large number of polymorphisms that are generated by AFLP analysis is one of the major advantages of this method with Geleta et al. (2006) finding AFLP analysis 14 times more efficient at detecting polymorphisms than microsatellites. In addition, AFLP assays are reportedly robust, with a study across eight laboratories finding highly comparable results for
AFLP assays, in comparison to the low reproducibility of RAPD assays (Jones et al., 1997). No prior sequence information is required before using this technique, as opposed to microsatellite and RFLPs where prior characterisation of the genome is required. With increasing sequencing studies this advantage will diminish, however, in terms of less well studied species, the minimal preliminary work offers a major advantage.

Due to the dominant nature of AFLPs it has been suggested that 2-10 times more individuals need to be sampled than when using co-dominant markers (Lynch & Milligan, 1994). However Krauss and Peakall (1998) suggest that the large number of polymorphisms generated in AFLP analysis may overcome the problem of using a dominant marker.

Many of the disadvantages of AFLPs are similar to those of other types of molecular marker assays and are reviewed in detail by Robinson and Harris (1999). One of the greatest problems in AFLP analysis is homology. Scoring co-migrating non-homologous bands can lead to an over-estimate of similarity, although many researchers argue that the probability of co-migrating bands being dissimilar is very small. However, Robinson and Harris (1999) point out that in a mapping study of Solanum, when sequencing 20 putatively homologous bands, 19 were found to be nearly identical, leaving 5% that are dissimilar. This problem with homology could point to problems with its use above the species level in determining phylogenies.

High correlations have been found between AFLP assays and the more resource costly RFLP and microsatellite assays (Powell et al., 1996; Geleta et al., 2006). With the reproducibility, low preliminary work and minimal time needed to generate large number of polymorphic markers AFLPs will be one of the methods adopted in this study.
1.5.3 Focus on SNPs

At present, the potential for single nucleotide polymorphisms (SNPs) to be used as molecular markers in molecular ecology and conservation genetic studies is gaining momentum (van Tienderen et al., 2002; Morin et al., 2004; Ouborg et al., 2010a). While the assessment of DNA sequence variation to provide molecular markers is not entirely novel as previous marker methods have been underpinned by sequence variation (restriction fragment length polymorphism, RFLP and amplified fragment length polymorphism, AFLP), technological progress in DNA sequencing including next generation sequencing is enabling researchers to analyse the underlying sequence variation hinted at by earlier molecular marker approaches (Brookes, 1999; Vignal et al., 2002; Schlötterer, 2004).

The interest in using SNPs as molecular markers for ecology and conservation studies derives from such polymorphisms overcoming many of the limitations of conventional marker systems, as well as their increasing impact and use in human disease and other model species studies (e.g. Cho et al., 1999; Gabriel et al., 2002; Zhu et al., 2003; Samani et al., 2007). In comparison to other markers, SNPs are known to be highly abundant and widespread throughout many species genomes, whereas microsatellites, the previous marker of choice for population genetic studies, are thought to occur once in every 6-7kb, as well as being difficult to isolate in some species (Cardle et al., 2000; Morin et al., 2004). Indeed, where other molecular marker systems have shown little diversity in a particular species, the potential abundance of SNPs renders them highly attractive, as seen in *Coffea arabica* where SNP identification was found to be the most appropriate marker to detect polymorphism in this low diversity species (Zarate et al., 2010). The higher mutation rate of microsatellites also requires caution, particularly when comparing distant lineages; those loci identical in size can be different in descent (homoplasy), while the lower mutation rate for SNPs and their simpler mutation model makes analysis easier (Vignal, 2002; Morin et al., 2004; Payseur & Cutter,
Finally, high-throughput, cost effective methods are increasingly available for SNP
detection following SNP identification and validation (Gupta et al., 2001; Syvänen, 2001).

The choice of marker to use in a particular study is most often a question of precision
against convenience (Sunnocks, 2000). Nevertheless with the potential advantages of using
SNPs as molecular markers, and with the increasing availability of high-throughput methods
for sequencing and genotyping assays, SNPs provide an exciting prospect for answering
ecological and conservation questions. Particularly in plants, the application of SNPs to these
types of questions is still in its infancy (Morin et al., 2004; Seddon et al. 2005; Ganal et al.,
2009). Here, I review research on SNPs in conservation and ecology, particularly focusing on
what has been learnt from SNP discovery in model organisms and animal species and how
such knowledge can be applied to SNP discovery in non-model plant species.

SNP basics

Single nucleotide polymorphisms (SNPs) refer to single base pair differences in DNA
between normal individual members of a given population(s). Both Brookes (1999) and
Vignal et al. (2002) suggest that for such single base pair positions to be considered SNPs the
least frequent allele should have a frequency of 1% or greater. However many authors
describe any differing locus as a SNP, most often due to a lack of frequency information or a
small sample size (Brookes, 1999). Sequence polymorphisms are frequent in the genome,
with SNP frequency in humans found to be 1/300-1,000 base pairs when comparing two
human chromosomes (Aitken et al., 2004). Frequencies identified in some plant species can
be higher, with reports of 1/23bp on average in Vitis vinifera (16 cultivars, Dong et al., 2010),
1/31 in non-coding regions and 1/124 in coding regions of the highly diverse Zea mays (36
inbred lines, Ching et al., 2002) and 1/130 on average across coding and non-coding regions
in *Beta vulgaris* (2 inbred lines, Schneider *et al*., 2001), with lower levels of diversity, 1/504 on average, in *Glycine max* (9 genotypes, Van *et al*., 2004).

Although in principle all base variants should be possible at one position in a DNA sequence (Figure 1.2a), SNPs are generally known as biallelic markers because tri- and tetra-allelic variants are rare, particularly in humans (Brookes, 1999). This prevalence in bi-allelic SNP types results in part from a bias in mutation types, as well as local and regional base pair composition (Morton, 1995; Vignal *et al*., 2002). Mutation types consist of transitions and transversions (Figure 1.2b), and as there are twice as many possible transversions to transitions the expected ratio of transitions to transversions should be 0.5, assuming all substitutions are random (Vignal *et al*., 2002).

![Figure 1.2. SNP basics; a) Image depicting a tetra-allelic SNP position b) Types of nucleotide substitutions between the four possible bases. Transitions only occur between the two purines (A & G) and between the two pyrimidines (C & T).](image)

In practice however there is often a bias towards transitions in eukaryote genomes (but see Keller *et al*., 2007). While this transition/transversion ratio holds in humans, in plants more variability is reported, with ratios from 0.23 in *Coffea arabica* (Zarate *et al*., 2010) and 1.2 found in genes of wild *Lycopersicon* (Frankel *et al*., 2003), to around 1.7 in genes of *Arabidopsis thaliana* (Martínez-Castilla & Alvarez-Buylla, 2003) and 1.82 in *Populus*
tremula (Ingvarsson, 2008). A ratio as high as 3.6 has been reported in *Vitis vinifera* (Dong *et al.*, 2010). Transition/transversion ratios vary between species and within species with different study designs finding both a transition bias, and equal transition/transversion rates in different gene regions of the same plant species (Zhu *et al.*, 2003; Van *et al.*, 2005).

In addition to the unequal transition/transversion ratio, the low rate of mutation may contribute to the bi-allelic nature of SNPs, as the possibility of two independent changes at the same position will be rare (Vignal *et al.*, 2002). The majority of published nuclear substitution rates concern mammalian genomes; most likely due to difficulties arising from complex orthologous relationships in species where there is high gene duplication such as that seen in plant species (DeRose-Wilson & Gaut, 2007). Commonly quoted nucleotide substitution rates in mammals are around $1-2 \times 10^{-9}$ substitutions per nucleotide per generation in humans (Crow, 1994), and between $1 \times 10^{-9}$ and $5 \times 10^{-9}$ per nucleotides per year at neutral positions in mammals generally (Vignal *et al.*, 2002). Comparing nucleotide substitution rates among plant nuclear genes, Wolfe *et al.* (1987) determined rates of around $5-30 \times 10^{-9}$ for synonymous substitutions per site per year, but suggest that the most accurate estimations are nearer the lower bound, giving similar rates to that determined in mammals. Gaut (1998) in a comparison of genes between rice and maize found rates of $6.03 \times 10^{-9}$ and $9.43 \times 10^{-10}$ for synonymous and non-synonymous substitutions respectively, similar to that found by Wolfe *et al.* (1987). Previously, considerable attention has been paid to chloroplast genes when comparing plant lineages as they are not limited by multiple copies inherent in the nuclear genome or by the slow evolution of the mitochondrial genome. However both nuclear and mitochondrial genomes are now becoming of more interest to evolutionary biologists, as questions of population biology and molecular genetics can be answered by comparing the evolution across the three genomes (Muse, 2000).
Considerations for SNP genetic variation studies

Number of loci and DNA polymorphism

Although abundant, SNPs, because of their bi-allelic nature yield a lower information content per locus than microsatellites, where each locus can number over 20 alleles (e.g. Poteaux et al., 1999). For this reason a larger number of SNPs is required when measuring population genetic parameters to compensate for their lower information content (Brumfield et al., 2003; Aitken et al., 2004; Morin et al., 2004). Indeed, Varshney et al. (2007), when comparing marker systems in Hordeum determined that SNP markers are less suitable for diversity studies when compared to AFLP or microsatellite markers when comparing equal numbers of SNP and microsatellite markers.

Studies using DNA sequences to determine optimal sampling strategies for coalescent based estimates of population genetic parameters such as θ, the proportion of polymorphic sites in a population (see Carling & Brumfield, 2007) found that, as one would expect, the accuracy of estimates was improved with increasing numbers of independently segregating SNP loci (Pluzhnikov & Donnelly, 1996; Felsenstein, 2006). However, Carling and Brumfield (2007) found that increasing the number of loci to above 25 had little effect on overall accuracy of the estimate of θ, as 81% of the total improvement was explained by increasing the number of loci from 1 to 5, and 98% when increasing to 25 loci. When determining optimum sampling strategies under a cost-per-base or cost-per-read scenario, both Pluzhnikov & Donnelly (1996) and Felsenstein (2006) determined that the addition of independent loci will increase accuracy over extending the individual locus sequence length. Carling and Brumfield (2007) indicated that while the more-loci-shorter-sequence strategy holds in most cases, where θ is low, longer sequences are required to increase the accuracy of
the estimate over increasing the number of loci. In any case Kuhner et al. (2000) point out that in cases of very low θ, extremely large numbers of SNPs will be required.

Using simulations Mariette et al. (2002) found that at least four times as many bi-allelic markers were as efficient at elucidating gene diversity in natural populations as one co-dominant, multi-allelic marker, but at least 10 were required when migration is high and heterogeneity within populations is low. However, the bi-allelic markers compared in this study were AFLPs, dominant markers in contrast to co-dominant SNPs; hence fewer SNPs than AFLPs per study would be required to reveal the same level of population differentiation (Morin et al., 2004). The differences in the perceived number of marker loci required can be large, with Ryynänen et al. (2007) indicating that up to 100 SNPs have been quoted as the number that may be required for pedigree reconstruction (Anderson & Garza, 2006), while just 3 microsatellites have proven sufficient for parentage assignment studies. Glaubitz et al. (2003) in a simulation study found that around five times as many SNPs as microsatellites are required to reliably determine population genetic relationships, while in the case of linkage map construction, around three moderately polymorphic SNP loci are equivalent to every microsatellite (Kruglyak, 1997), which Brumfield et al. (2003) propose may be similar to the relative number required for population genetic studies. More recently a comparison of microsatellite and SNP markers has corroborated the suggestion of Brumfield et al. (2003) with three times as many SNPs required for the same information content in poultry and cattle (Schopen et al., 2008). It is important to note that, while these studies often reflect an optimum number of SNPs, the number of loci and sample size will necessarily be both a function of the trade-off between cost, information content and the evolutionary context of the question asked. Narum et al. (2008), in a study in Chinook salmon, Oncorhynchus tshawytscha, found that random microsatellites were more informative markers for assignment tests than random SNPs overall, corroborating similar findings in studies in
humans (Rosenberg et al., 2003; Liu et al., 2005). However, in this study 7 out of the top 10 most informative markers were SNPs, and SNPs had 20% higher assignment accuracy than microsatellites in 4 of the 29 populations studied. Thus Narum et al. (2008) advocate a more complementary strategy of using a suite of marker types for population genetic studies. Indeed, while emphasizing that a large number loci should be preferable, Ryynänen et al. (2007) found that in some cases a relatively small number of SNPs could provide concordant results with that of microsatellite markers in *Salmo salar*.

**Ascertainment bias**

Along with numbers of samples and loci to be used, a critical concern when designing SNP surveys is their ascertainment (Kuhner et al., 2000; Wakeley et al., 2001; Brumfield et al., 2003; Nielsen & Signorovitch, 2003; Clark et al., 2005). Ascertainment bias can be particularly prevalent in SNP studies, arising from researcher influenced criteria for SNP determination and selection of the individuals and/or loci sampled, culminating in results that are not representative of the total population. Any statistical analyses that depend upon the determination of accurate allele frequencies will be affected. The criteria for a polymorphism being designated a SNP can lead to an ascertainment bias. For example using a low-frequency cut-off to avoid sequencing errors will lose resolution of the pattern of diversity (Brumfield et al., 2003). Using a SNP discovery panel of a small subset of individuals will lead to an ascertainment bias if only those SNPs defined by the panel are subsequently genotyped, with SNPs present in the rest of the population remaining undetected (Brumfield et al., 2003). As another example, the preferential choice of high frequency alleles would allow the phenomenon of population expansion to go undetected, as the large number of low frequency alleles resulting from the expansion may not be assessed (Nielsen, 2000). The statistical impact of the problems associated with ascertainment bias is well documented (e.g. Kuhner et
Kuhner et al. (2000) determine that the impact of misrepresenting panel designated SNPs as sample SNPs can result in a 10-fold difference in the results, with the magnitude of the effect dependent on the ascertainment method and model chosen. Panel based ascertainment is however often necessary, but the ascertainment bias must be reduced through the use of a geographically widespread panel (Morin et al., 2004). In the case of ecology and conservation based studies ascertainment bias may be exacerbated where samples come from complex (or unknown) population structures, making the designation of the panel more difficult, and thus more variation is likely to go undetected (Rosenblum & Novembre, 2007). However, while ascertainment bias is undoubtedly a problem in analyses that require a random genetic sample, using an ascertainment set from a wide variety of populations can actually provide additional power to population assignment and genetic differentiation analyses (Morin et al., 2004). Beyond the ascertainment set, locus selection automatically introduces bias by the non-random coverage of the genome. Therefore, while searches for SNPs in candidate genes give some idea of functionality and enables adaptive diversity to be considered, studies requiring unbiased estimates of genomic variation would benefit from methods that sequence random sections of the genome.

Ascertainment bias can be corrected for in some analyses if information about the panel and criteria for designating sites as SNPs is retained (e.g. Kuhner et al., 2000, Nielsen & Signorovitch, 2003). As such it is vital that researchers retain and publish information attaining to SNP discovery so that results can be correctly interpreted.

**Heterozygote ascertainment**

Another consideration of particular importance for conservation and ecology based studies arises from haplotype and heterozygote ascertainment, with wild natural populations
more likely to contain heterozygotes than crop plants derived from inbreeding. For example, Hyten et al. (2006) found the wild progenitor of cultivated Glycine max, G. soja, contains twice as much nucleotide diversity per base pair as elite cultivars.

Two of the most common methods to determine heterozygotes are cloning PCR products and the comparison of signal intensities. Cloning PCR products will produce unambiguous haplotypes from obligate outcrossing species (Edwards et al., 2007). However the high cost of cloning and large amount of sequencing required can render this method less attractive. An alternative method uses differences in sequence traces to infer heterozygotes from signal intensity, due to a reduction in intensity at around ~50% of the height of a heterozygous peak compared to surrounding homozygous peaks (Brumfield et al., 2003). However Zhang & Hewitt (2003) indicate that two or more heterozygous positions in one sequence read can make this method unreliable. Other methods include likelihood approaches to statistically phase DNA sequences, and computer programs have been developed for this process (Stephens et al., 2001; Stephens & Donelly, 2003; Scheet & Stephens, 2006; Templeton, 2006; Wang & Xu, 2003). As no statistical approach can work accurately when the original dataset consists of ambiguous genotypes both Zhang and Hewitt (2003) and Templeton (2006) discuss the need for experimental methods to be used in conjunction with the latest statistical approaches to determine haplotypes with the greatest accuracy.

The problem of ploidy

Polyploidy, the result of more than two complete genomes per cell, occurs in many taxonomic groups (Soltis & Soltis, 2000; Otto & Whitton, 2000; Legatt & Iwama, 2003), but is particularly prevalent in plant species, with most estimates suggesting that 60-70% of angiosperms have a polyploid ancestry (Blanc & Wolfe 2004; Tang et al., 2004; Cui et al. 2006). Polyploids can be divided into two groups on the basis of their origin, with those
polyploids derived from the multiplication of a single genome termed autopolyploids, or if
derived from combination of the genomes of two distinct species, allopolyploids. A further
classification in polyploids can be made on the basis of their mode of inheritance. In recent
autopolyploids the homologous and/or homeologous chromosomes pair at random and can
also form multivalents at meiosis, (Bever & Felber, 1992; Obbard et al. 2006), a process
termed polysomic inheritance. For allopolyploids the differentiated sets of chromosomes may
pair separately as in their diploid progenitors, a process known as disomic inheritance.
However, while allopolyploids can be generally thought of as disomic and autopolyploids as
polysomic, both types of inheritance actually represent two extremes at the ends of a
continuum. For allopolyploids, intermediate modes of inheritance can occur where meiotic
pairing occurs between closely related parental genotypes; hence crossing over occurs
between homeologous chromosomes, a process that can homogenise the genome over time
(Sybenga, 1996). In autopolyploids the four homologous chromosomes can differentiate to
form two pairs of chromosomes (diploidisation), eventually leading to disomic inheritance
(Wolfe, 2001; Ramsey & Schemske 2002). In fact even in recent allo- and autopolyploids
intermediate modes of inheritance are seen, with either closely related parental genotypes
allowing multivalent formation, or small genomes and low chiasma frequencies in
autopolyploids leading to disomic inheritance (Sybenga, 1999; Otto & Whitton, 2000; De
Silva et al., 2005; Obbard et al., 2006).

The presence of polyploidy and the complexities described above can cause difficulties
when quantifying genetic variation and population differentiation, unlike in the more simple
cases of diploid genomes. Instead of the simple presence/absence associated with SNP
detection in diploids, polyploids would require a frequency measure of each base in each
genotype. For example a diploid individual heterozygous at a particular allele will have the
allelic phenotype AB. In the case of a heterozygous tetraploid, the genotype may be AAAB,
AABB, ABBB, depending on the copy number of each allele, known as allele dosage. In polyploids with low numbers of genomes, particularly tetraploids, allele dosage can be estimated when using certain types of markers by looking at band intensity on gels or peak heights/areas (e.g. Prober et al., 1998; Hardy & Vekmans, 2001; García-Verdugo et al., 2009) on electropherograms. Using this method in autopolyploids it is then possible to estimate the genotype from the allelic phenotype, and if the inheritance is known to be polysomic general population genetic statistics can be adapted to enable analysis (Obbard et al., 2006). A more complex issue arises when the species has disomic inheritance, in allopolyploids and in diploidised autotetraploids. In these cases it is not clear which alleles are associated with which of the duplicate loci, or isoloci. A tetraploid that is heterozygous at a particular locus may be homozygous at both isoloci or heterozygous at both, or just one locus. Where allele dosage can be scored it is possible to calculate gene frequencies at isoloci and then use this information to calculate genetic diversity statistics (Obbard et al., 2006). Both Waples (1988) and more recently De Silva et al. (2005) propose ways to estimate allele dosage information from microsatellite markers in allotetraploid species to enable further analysis, although both methods use fully disomic species, which as outlined above is often an extreme, with natural polyploids often showing an intermediate form of inheritance.

Due to the relatively recent use of SNPs for generating molecular marker data there have been fewer attempts to produce statistical programs that can analyse SNP data from polyploids, particularly as cloning can directly determine haplotypes and allele dosages without the ambiguities associated with estimating allele dosage. While cloning can produce unambiguous results for SNP discovery and genotyping, a large number of clones need to be sequenced from a PCR pool to be sure of obtaining each homoeologous variant at an allotetraploid locus with a high probability of success (Tiffin & Gaut, 2001; Caldwell et al., 2004; Simko, 2004). Hand et al. (2008) for example sampled 24 individual clones from each
T. repens (allotetraploid) individual, thus for the large numbers of individuals required for population genetic and conservation analyses this method will be unrealistically limiting.

SNP studies in allopolyploid plant species have mostly used the progenitor comparison approach, where data are separated into isoloci using pedigree analysis or parental genotypes (although see also Ching et al., 2002). As an example of this approach, Hand et al. (2008) used information from Trifolium occidentale and T. pallescens as the most related species to the diploid progenitors of T. repens to separate the allotetraploid genome for further analysis. Cotton species, Gossypium hirsutum and G. barbadense, are split into their constituent A and D genomes for further analysis (Small et al., 1999). This method enables researchers to calculate diversity statistics directly from the two separate progenitor genomes.

It is clear that SNP detection and subsequent analysis in polyploid plant species, particularly allopolyploid species, is both more resource intensive, and requires more detailed information on the progenitors of the species than will be available for most conservation studies.

SNP identification methods in non-model organisms

SNP identification is possible by comparing the sequence information for a locus from several individuals, with many methods having been described (Gupta et al., 2001; Rafalski, 2002; Edward et al., 2008; Ganal et al., 2009). Whilst a thorough assessment of all SNP discovery methods is beyond the scope of this review, direct SNP discovery and validation in non-model organisms generally falls into 2 themes dependent on prior information availability.

No prior sequence information - Whole genomic shotgun sequencing

This approach uses a genomic library derived from a mixture of DNA from a number of individuals, enabling the sequencing of random segments of the genome (Gupta et al., 2001).
The method is disadvantaged because of the expense involved and the probability of over-representation of repetitive sequences (Primmer et al., 2002). Nonetheless, in the complete absence of sequencing information large amounts of high quality sequence data can be produced for non-model species. Further, unlike locus specific amplification, the method sequences random segments of the genome and produces haplotype information, removing some of the problems of ascertainment bias associated with locus selection as well as heterozygote assignment.

**No prior sequence information - Reduced representation shotgun (RRS)**

While this method is similar to the whole genome shotgun sequencing approach above, it relies on using a subset of each individual’s genome to reduce the amount of re-sequencing required to discover SNPs (Altshuler et al., 2000). It consists of mixing the DNA of several individuals, and selecting from those a subset of the genome to create a library for further sequencing, with the subset criteria based on for example restriction fragments size, methylation, or copy number (Barbazuk et al., 2005). This selection ensures that a similar subset of fragments from different individuals is selected for sequencing, increasing the efficiency of SNP identification. Hyten et al. (2010) used this method in conjunction with next generation sequencing to identify 7,108 to 25,047 putative SNPs in Glycine max.

**No prior sequence information - SNPs by AFLP (SBA)**

This approach is particularly applicable to non-model organisms, with AFLP often the marker of choice for genetic diversity studies in species with no prior genetic information. This method makes further use of the ligation of the primer site following AFLP amplification, to sequence random homologous bands from different individuals (Nicod & Largiader, 2003; Roden et al., 2009). Further, highly informative markers or potentially adaptive markers can be selected by choosing markers that show specific differences among
populations or groups in the AFLP study (Bensch et al., 2002). However, Edward et al. (2008) note that sequencing fragments derived from AFLP analysis is complicated where multiple copies of the target sequence are present in a genome, through either polyploidy or where the AFLP fragments represent a member of a repetitive element family.

No prior sequence information - Next generation sequencing

It is highly likely that the sequencing of non-model organisms will greatly benefit from the large advances being made in sequencing technology. Next-generation sequencing technologies can generate high quality sequence information of 20-30 mbp in one run relatively cheaply, compared to around 67,000 bp per hour using traditional Sanger sequencing methods (Margulies et al., 2005; Vera et al., 2008). While this new technology can greatly increase the speed at which sequences are generated, the average length of sequences may only be around 100bp, compared to the average of 700bp generated from traditional sequencing methods (Margulies et al., 2005). However, the higher level of redundant coverage of genes can specifically lend this method to SNP discovery when sequencing a diverse array of individuals in non-model species. A next generation sequencing run made up of sequencing a pool of diverse individuals is likely to generate thousands of SNPs, in addition to providing thousands of contigs that can be compared to available sequencing information for annotation (Vera et al., 2008; Gompert et al., 2010; Ouborg et al., 2010). Thus, this technology is likely to provide one of the most feasible methods for SNP discovery in non-model plant species (Ouborg et al., 2010). Recently published studies show the potential of this method, with Vera et al. (2008), Gompert et al. (2010) and Van Bers et al. (2010) discovering a large number of novel SNPs in species where little prior genomic information was available. The potential to assemble the short sequence reads into contigs is noted to be difficult in the absence of a published assembly reference genome (Trombetti et
al., 2007; Vera et al., 2008). Therefore the ability to determine the potential adaptive significance of these SNPs by identifying coding and non-coding regions of the genomes can be made more complex in non-model species where only distantly related species data are available for use as a reference guide to assembly. Further, Gompert et al. (2010) note that the large amount of unassembled reads generated provide no information for genetic variation studies, and thus sequence assembly into contigs is vital to the use of such data for addressing population and conservation questions.

A further limitation may be the higher error rate in next generation sequencing compared to Sanger sequencing, with the potential to overestimate the level of polymorphism in the sample (Margulies et al., 2005; Harismendy et al., 2009). High error rate in these sequences will overinflate nucleotide diversity and this will be evident in an excess of rare alleles (Pool et al., 2010). Validation of 454 Life Sciences (http://www.454.com) identified SNPs, using cloning and traditional Sanger sequencing has been shown in some cases to be relatively high, with Barbazuk et al. (2007) reporting >88% SNP detection accuracy in Zea mays.

**Prior sequence information available- In silico SNP discovery**

Nucleotide databases can be used to identify potential primers by aligning sequences derived from different individuals (Picoult-Newberg et al., 1999). In particular large numbers of ESTs have now been generated for various crop plants and it is possible to mine this information using various tools to discover SNPs (e.g. Buetow et al., 1999; Gorbach et al., 2009). However, large-scale *in silico* SNP discovery efforts require a large number of sequences from a diverse set of individuals, and thus for the majority of non-model species this will not be the case. A small amount of sequence data in the target species or any closely related species can be used to aid SNP discovery in targeted areas, as detailed below.
Prior sequence information available- Locus specific amplification

With the presence of sequence data, specific genic regions can be selected and amplified by PCR (e.g. *Arabidopsis thaliana*, Brock *et al.*, 2007; *Hordeum vulgare* ssp. *spontaneum*, Lin *et al.*, 2001; *Zea mays* ssp. *parviglumis*, Moeller & Tiffin, 2008). The sequences of several individuals can be compared to discover novel SNP sites (see Figure 1.3 for a flow diagram of decision-making for SNP detection using a candidate gene approach). Even in the absence of annotated gene fragments for the species in question, this method is increasingly being used to compare nucleotide databases of the target species with that of species where more genetic information is available to discover candidate genes of interest (Eveno *et al.*, 2008; Hand *et al.*, 2008).
In the presence of comparative sequences of the target locus in other species, alignments can highlight areas of conserved nucleotides in which to site primers. Within this overall approach several different methods have been proposed. Amongst the most common is
comparative anchor tagged sequences (CATS) or exon priming intron crossing (EPIC) where highly conserved exon regions are used to design primers to amplify unconserved intron sequences (Palumbi & Baker, 1994; Lyons et al., 1997). Aitken et al. (2004) used this approach to sequence targeted geneic regions and from there to discover SNPs across a selection of mammal taxa, with 50% of primers successfully amplifying putative homologues in around half of the species analysed. When using methods such as this, there needs to be some caution - members of gene families and genes with recognised pseudogenes need to be excluded to reduce the extent to which multiple genome fragments are amplified (Primmer et al., 2002). Thus, using these methods in plant species may prove difficult due to the high levels of gene duplication and polyploidy. Ryynänen & Primmer (2006) expand on this idea to produce intron-primed exon crossing (IPEC) to reduce the probability of amplifying duplicated genes, siting primers in more variable genomic regions. From this method, 95% of loci were successfully amplified, providing a clean single PCR product for sequencing in *Salmo salar*, compared to 30% success rate using the EPIC strategy. Clearly in the absence of any sequencing information for closely related taxa this method becomes limited, as design of primers based on more closely-related rather than more distantly related species are more likely to provide a higher rate of success (Housley et al., 2006). Indeed SNP discovery studies in non-sequenced species have proven highly successful when using sequence information from closely related taxa (Adams et al., 2006; Sacks & Louie, 2008).

In the absence of sequence from even distantly related taxa, regions of conserved nucleotides can become patchy, requiring the need for degenerate primer design, and using a pool of primers to amplify a selected locus (Kwok et al., 1994). Specifically, degenerate oligonucleotide primed (DOP)-PCR uses primers with 5’ G/C rich anchors and unique 3’ nucleotides, confining degenerate nucleotides (approximately one third of the primer length)
to the centre of the primer, which has increased the efficiency of the degenerate PCR reaction (Jordan et al., 2002; Janiak et al., 2008).

However multiple cloning and sequencing of the same PCR product remains as a major potential problem in some cases, with Janiak et al. (2008) indicating that ploidy and genome duplication in plant genomes can render degenerate PCR unsuitable for SNP discovery. Whilst more technically complex than other methods mentioned above, some authors have successfully managed to use degenerate PCR to discover SNPs in plant species, particularly in studies of genic regions that are highly conserved between species. For example, degenerate primers designed to amplify resistance genes in *Glycine max* (Kanazin, 1996), have been used successfully to amplify the same region in *Zingiber officianalis* (Nair & Thomas, 2007) and degenerate primers designed from conserved resistance regions of *Oryza sativa* and *Lycopersicon* were used to design primers for amplification in citrus species (Deng & Gmitter, 2003).

* A promising future for SNPs in conservation studies?

The likely effect of climate change, coupled with the expanding ecological footprint of an ever increasing human population portends significant future environmental change. Species persistence in a changing environment is dependent on its vulnerability to these changes and its ability to adapt. While most studies of genetic variation to date have focused on neutral diversity, this is by no means always an efficient surrogate for adaptive diversity (Reed & Frankham, 2001). While neutral diversity has provided a large amount of information for conservation genetics, the relationship between variation and ecologically important traits may be untested (Ouborg et al., 2010). Whilst adaptive or environmental change can be shown by monitoring the changes in variation indicated by neutral markers,
variation is affected only after a population size decrease or a change to gene flow (Hoffmann & Willi, 2008).

SNPs provide an exciting prospect for studies of ecologically important traits, providing markers that can be directly tested for selection, and thus increasing the knowledge of the effects of selection on functional traits in wild populations and the ability of these species to adapt to future changes. In terms of conservation, the potential to identify both those populations most threatened by environmental changes and those containing the highest diversity in functionally important genes is a clear advantage for future conservation efforts (van Tienderen, 2002; Hoffmann & Willi, 2008).

It is possible that, due to data quality, wide genome coverage, high variability and the potential to identify genes under selection, SNP markers may become the most popular marker in ecological and conservation genetics studies. However, for non-model species the methods described above indicate the high level of resources and/or prior information required to identify enough SNPs to successfully carry out a conservation genetic study (see also Morin et al., 2004). Despite the interest in SNPs as molecular markers there have been few studies assessing ecological and conservation based questions in non-model species. The majority of SNP studies focus on humans and other model organisms, with only six species accounting for over 90% of all the submissions to dbSNP as at August 2010 (Sherry et al., 2010). With the availability of large sequencing datasets with information from different individuals, much of the SNP detection in model organisms can be completed in silico. However, despite the growing number of complete or near complete genomes and large numbers of sequences currently available (see Figure 1.4), for the vast majority of species sequence data is limited or nonexistent.
For non-model organisms, this lack of available sequencing information provides a major technical and economic obstacle to SNP discovery. While the high potential for SNPs use is confirmed by crop plant studies in adaptation and diversity, resource constraints mean that studies in non-model plant species are lagging behind (but see Novaes et al., 2008, Beatty et al., 2010 and Friesen et al., 2010 for examples). Indeed Schlötterer (2004) questioned whether, in small population sizes of the type typically used for conservation studies, microsatellites may provide a more cost effective marker. Until recently the cost of SNP development, in addition to the lower information content per locus, suggested that resource constraints in non-model species will restrict marker choice to the more traditional methods, as the traditional markers such as AFLPs and microsatellites still provide excellent tools to assess demographic processes. The potential shown in their use in model species, and the likely increase in discovery as more sequences are made available still suggests that SNPs will become the marker of choice for future studies, particularly in light of next generation sequencing technology. However, until larger numbers of sequences are made available for a
more diverse set of organisms, the choice of marker in conservation studies for non-model species is still very much a question of finding the most cost-effective marker for both the study design and question being asked. Consequently in the meantime more traditional methods will provide the most informative markers where resources are limiting.

1.6 QUANTIFYING GENETIC DIFFERENCES

A basic concept underlying most neutral genetic theories, where the marker measured is neutral to selection, is that genetic diversity is positively correlated to effective population size ($N_e$) at equilibrium. Hence, it follows that smaller population sizes will be more prone to loss of rare alleles and have lower average expected heterozygosity ($H_e$) than those with larger $N_e$ values (Beebee & Rowe, 2004). In a review of plant and animal taxa, Frankham (1996) found that this relationship held in the vast majority of cases. This relationship is useful in terms of a single random breeding population; however population subdivision is a feature of most naturally occurring populations.

Measuring and partitioning genetic diversity between subdivided populations, and using this information to infer population structure, has had a long history, from Wright’s $F$ statistics (1951) to high powered statistical computer programs used to produce multivariate analyses of genetic diversity assessments (e.g. Labate, 2000). There have been several addendums to the initial $F$ statistics, with Nei (1973) addressing the complication of multiple alleles and Weir and Cockerham (1984) addressing unequal sample sizes. The most frequently used genetic distance measure, a measure of dissimilarity between populations, was developed by Nei (1987) and can be used to infer phylogenetic relationships between species.
1.7 Gene flow and genetic diversity

In any discussion on gene flow it is important to establish certain definitions. “Gene flow” describes the movement of alleles between distinct populations due to the dispersal of gametes and zygotes. However, it has become increasingly clear that the emphasis should not be just on gene flow, rather on the incorporation and stabilisation of alleles into the recipient genepool. This “incorporation and stabilisation” of alleles is a factor often missed in discussions on gene flow (Arriola, 2005), and has led to the term “introgression”; a consequence of gene flow where introgressed alleles have become a permanent and stable part of the recipient populations genepool (Stewart et al., 2003; Arriola, 2005).

The term “genetic pollution” came into use with the increasing concern over GM crops, and as such is a rather controversial term, emoting unfavourable consequences. With the current discussion concerning its use in science, this will not be used further.

1.7.1 The consequence of introgression in conservation

One of the main causes of debate in the dispute over genetically modified crops concerns the introgression of transgenes from genetically modified (GM) crops into their wild relatives, which has stimulated much research in this area. This work has not only highlighted the potential introgression between GM crops and their wild relatives, but also into natural introgression between traditionally bred crops and their wild relatives.

Levels of natural introgression are higher than historically thought, with Ellstrand et al. (1999) in a review finding that 12 of the 13 most economically important crop species hybridise with wild relatives. Stewart et al. (2003) point out that the current estimates of confirmed introgressions are likely to be the very minimum, with less recent hybridisations and those between closely related species more likely to go undiscovered.
However, the extent and therefore the impact of introgression varies among species, populations and even years (Ellstrand et al., 1999). The level of importance in evolution is species and situation dependant; for example Ehrlich and Raven (1969) point out that sufficiently strong selection pressures could overcome the uniformity associated with introgression. It is nonetheless a potent force in the homogenisation of a species, the consequences for genetic diversity and evolution lying in its ability to sustain the evenness of species geno- and phenotypes, counteracting the factors leading to divergence in species (e.g. Mayr, 1970; Slatkin, 1985). Moreover, introgression contributes to genetic swamping, where a rapid increase in one genotype (or allele) replacing local ecotypes, leading to the decrease in genetic diversity of the species. The implication for conservationists being that the effects of introgression will serve to eradicate the genetic diversity that they are attempting to conserve.

In terms of this study the rates of gene flow and introgression are a factor in the availability of local adaptation and genetic variability between populations. Greene et al. (2004) found a relationship between environmental variation and genetic diversity in *Trifolium pratense*, but only for populations that were geographically isolated. Detecting population subdivision, with an idea of the rates and barriers of gene flow is the basis for defining units of evolutionary and conservational importance, a must for the effective management of genetic diversity (Manel et al., 2003). Furthermore, attempting to estimate levels of gene flow in species of varying levels of cultivation history will provide a valuable insight into the influence of agriculture on patterns of gene flow.

1.7.2 Measurements of gene flow

The importance of introgression is clear; however measurements are extremely difficult. Direct measurements of gene flow and dispersal are resource expensive and as a consequence limited in time and space, not reflecting stochastic events and often biased against long-
distance dispersal events (Whitlock & McCauley, 1999). In addition introgression itself is
often not measured, as direct measures of organism movement are not synonymous with the
incorporation of alleles in the recipient population.

In light of this, indirect measures are most often used to infer rates of gene flow and
consequently introgression, using comparisons of genetic data between populations to model
rates of gene flow. The derivation of many of these models and their use in estimating gene
flow has been the subject of many reviews (e.g. Slatkin, 1985; Neigel, 1997). Wright (1931)
put forward the island model to transform his measure of population subdivision, $F_{ST}$, into $N_m$,
a measure of the number of migrants coming into a population, hence a quantification of gene
flow. The concept of an ‘island’ model, with each separate component of a metapopulation
perceived as an ‘island’, has been the basis of many measurements of gene flow (Neigel,
1997). Of particular interest to this study are the isolation-by-distance models (Wright, 1943)
and stepping-stone-models (Kimura & Weiss, 1964) which correlate genetic variation
between populations with geographical distance.

Although it will be possible to infer levels of gene flow from $F_{ST}$, it is difficult to test
these methods and there is considerable controversy about the usefulness of these indirect
measures of gene flow and indeed whether these methods can be misleading (Neigel, 1997;
Bohonak et al. 1998; Bossart & Powell; 1998a; Whitlock & McCauley, 1999). The
assumptions of the population model on which these indirect methods base their estimations
of gene flow are the basis for much of the controversy. For example the most well known of
the models, the Wright (1931) island model, assumes no selection, no mutation, equal
population size, equal immigration/emigration rates and random migration (Hutchinson &
Templeton, 1999; Whitlock & McCauley, 1999), which is clearly a too simplified version of
the intricacies inherent in biological systems. It is likely that $F_{ST}$ will respond to changes in
migration rate, but may take many generations to do so, hence the value obtained may reflect past levels of migration but is unlikely to reflect current migration rates (Latta, 2006).

However, although caution must be taken in analysing these results, it is important to note that until the much called for advances occur, these analyses do provide an insight into the gene flow between populations, albeit with large degrees of error. The prevailing conclusion in the literature is that, as long as the limitations of the underlying population models are understood and the results critically evaluated, they provide useful inferences on gene flow and genetic boundaries (Neigel, 1997; Bossart & Prowell; 1998b; Whitlock & McCauley, 1999).

One of the points highlighted is a need for more realistic models in order to advance, which will be possible with the provision of more detailed information. Whitlock and McCauley (1999) note that by comparing contrasting elements of a metapopulation e.g. “mainland vs. island” populations, a comparison that will be addressed in this study, theoreticians will gain more insight into the importance of these factors in creating genetic patterns. Recent literature has given an increasing emphasis on landscape genetics, the explanation of genetic variation by comparison with landscape variables (Sork et al., 1999; Manel et al., 2003).

1.7.3 Landscape genetic analysis and its implications for population genetics

Quantification of the level of genetic diversity goes only part way to answering the question of how to assess the status, viability and threats for species in terms of conservation biology (Escudero et al., 2003). Landscape genetics, a discipline that “endorses those studies that combine population genetic data, adaptive or neutral, with data on landscape composition and configuration” (Holderegger & Wagner, 2006), attempts to bridge the gap between population genetic assessments and landscape ecology. Recent developments in the field of
landscape genetics have integrated the spatial and genetic patterns of this diversity, defining how the configuration of the landscape interacts with population genetics. Landscape genetics incorporate the fields of population genetics and landscape ecology allowing population geneticists to determine processes based on landscape composition, not just on spatial distance (Holderegger & Wagner, 2006; Storfer et al., 2007).

Conservation biologists face many difficulties in correlating the complexity of the real landscape to the genetics of the species contained within it. Barriers to gene flow, the number of distinct populations in a defined area, the underlying cryptic patterns of genetic diversity and the effects of habitat heterogeneity can all be identified by landscape genetic tools (Manel et al., 2003). The Mantel (1967) test, a test of the correlation between two matrices, is a cornerstone of landscape genetic analyses and has been widely applied in population genetics studies. Classically Mantel tests are used to test for isolation by distance (IBD), by comparing a matrix of genetic distance with one of geographic distance i.e. are populations in close proximity more genetically related than those that are geographically further apart? While the Mantel test remains a robust technique for population geneticists its main limitation lies in its ability to detect only linear patterns of spatial correlation (Escudero et al., 2003). To compensate for this limitation, if there is an expectation of non-linear population structure, non-linear distance metrics can be used; however this method requires a priori knowledge of population substructure (Heywood, 1991).

An extension to the Mantel test is the partial Mantel test, a comparison of three matrices, in essence comparing a response matrix to a predictor matrix while controlling for the effect of a third variable (Smouse et al., 1986). In this way the relative importance of environmental factors can be tested, while controlling for patterns resulting from spatial distance (Still et al., 2005; Storfer et al., 2007). Whilst this test has been used in population genetic studies and remains a useful indicator of environmental impacts on genetic diversity
patterns there has been some controversy over the statistical validity of this test (Raufaste & Rousset, 2001; Castellano & Balletto, 2002; Rousset, 2002).

Aside from matrix correlations the majority of landscape genetic literature now focuses on several statistical techniques; assignment tests, autocorrelation/correlograms, dispersal route analysis, combined GIS approaches and ordination (Manel et al., 2003; Storfer et al., 2007). Dispersal route analysis tests the relationship between genetic distance and alternate paths through habitats, for example; along a river or shortest straight line paths through suitable habitat (Michels et al., 2001; Coulon et al., 2004; Poissant et al., 2005). For more complex landscapes, these types of analyses can explain more variation in gene flow than traditional straight line Euclidean distance measures (Storfer et al., 2007).

Traditional population genetics often uses a priori population delimitation; however population assignment methods allow populations to be defined within a continuous habitat. Bayesian clustering approaches allow random mating individuals to be assigned to sets (populations) within the total dataset, based on the grouping of individuals that minimize Hardy Weinberg and gametic disequilibrium (Manel et al., 2003). Using this method authors have been able to ascertain genetic discontinuities in populations that are spatially continuous, exposing the dispersal patterns and fragmented landscapes more difficult to detect using traditional methods (He et al., 2004; Prentice et al., 2006; Grivet et al., 2008). Allocation procedures can be extended to determine the population of origin, i.e. the source population to which the genotype is most likely to belong (Duchesne & Bernatchez, 2002). Although these types of analyses have huge potential for population genetic research, they are limited in that, by assuming random mating, they preclude the use of selfing and partially selfing species.

Spatial autocorrelograms quantify the genetic relatedness between pairs of individuals and geographical distance, with no prior knowledge of spatial structure assumed (Sokal & Oden, 1978; Heywood, 1991; Manel et al., 2003). These techniques assess whether the
genotype of one individual at a known location is independent of that from another at a neighbouring locality (Escudero et al., 2003; Manel et al., 2003; Storfer et al., 2007). The results provide a graphical depiction of the association over geographical distance classes defining the spatial scale of the genetic pattern of diversity (Heuertz et al., 2003; Ishihama et al., 2005; Tero et al., 2005). Although spatial autocorrelation can define the scale of the pattern in a continuous population, this method needs to be coupled with other landscape genetic methods to determine which landscape boundaries create the observed genetic discontinuities.

Ordination methods such as canonical analysis (CA) can be used in a landscape genetics context to investigate relationships between environmental variables and the spatial patterns of genetic diversity (Storfer et al., 2007). In particular spatial coordinates can be incorporated as covariables to determine the extent of genetic variation explained by habitat related variables while accounting for the variation explained by geographic distance (Geffen et al., 2004, 2007; Smith et al., 2008). Due to the controversy surrounding partial mantel tests, canonical analysis, in particular distance based redundancy analysis (dbRDA), provides an alternative to traditional methods to enable population geneticists to quantify variation explained by environmental variables (Anderson, 2003).

Analysis at the landscape level has intuitive criticisms, namely that any landscape will rarely remain constant and that, although the spatial pattern of diversity is defined, the processes that cause an observed pattern of diversity may still be uncertain. Landscape genetic methods, as with many population genetic analyses, require assumptions that are likely to be violated in natural populations, however the extent to which the violations impact the reliability of the conclusions drawn is uncertain (Storfer et al., 2007).

Nevertheless the use of landscape genetics remains an exciting possibility for population geneticists, going some way to answer the questions associated with patterns of
gene flow and spatial patterns of genetic diversity. Although landscape genetics itself requires large scale random sampling strategies, many of its methods can be applied more generally to population genetics studies to help reveal genetic discontinuities and the landscape and environmental features that create them. In terms of conservation the results of such analyses are vital for evidence based approaches to conservation management in real landscapes.

1.8 MAINLAND VERSUS ISLAND POPULATIONS

The importance of islands in overall species diversity is well acknowledged, with speciation following long periods of isolation giving rise to endemic species, as well as many islands acting as repositories for many of the world's threatened taxa (e.g. Darwin, 1859; Paulay, 1994; Myers et al., 2000). Less well recognised is the importance of islands in terms of genetic diversity below the species level, with the isolation of island populations likely to create reservoirs of distinct variation compared to mainland populations.

Extinction of species has long been a focus of island literature and a clear indicator of loss of diversity on islands, occurring more readily in island ecosystems as a response to a small overall area, greater reactions to stochastic processes, genetic drift and lower levels of immigration. These same processes affect island populations of non-endemic species, and indicate the real importance of the conservation of island biotas.

Islands themselves represent an interesting paradigm containing fewer species than the same area on the mainland as a function of their lower overall area, following MacArthur and Wilson's (1967) model of island biogeography, and typically contain lower levels of diversity than their mainland counterparts (Frankham, 1997). Nonetheless it is the rarity of both the species and genetic diversity on islands that defines their conservation potential.
1.8.1 Previous studies of diversity in mainland versus island populations

The importance of islands in speciation and evolution was recognised by Darwin, with islands acting as natural experiments in evolution (MacArthur & Wilson, 1967) and as such they have been the focus of many studies and models particularly in terms of studies into adaptive radiation.

Frankham (1997) defines island isolation, size and life history traits as factors determining the scale of difference in diversity between island and mainland populations, with higher differences found when assessing smaller, isolated islands and for species with low dispersal rates and low adaptability. In terms of this study, the genetic variation on the islands will be greatly impacted by the life history traits of the species (i.e. modes of pollination) and the isolation of the islands, with distance from the mainland complicated by the surrounding islands in the archipelago and intermediate islands between the island of study and the mainland.

In a review of island diversity versus mainland diversity, the majority of island populations were found to be significantly different to mainland populations (see Frankham, 1997). In addition, the level of genetic diversity within populations has been found to be significantly lower in insular populations than in mainland populations (Frankham, 1997). Inoue and Kawahara (1990) report a negative correlation between the total genetic diversity of islands and distance from the mainland for *Campanula punctata*, indicating the more distant island groups are likely to contain lower levels of genetic diversity.

Crop species, or those strongly complicated by trade and human movement have been excluded from review (Frankham, 1997). However in an assessment of the genetic variation of cotton (*Gossypium hirsutum* L.), although low in comparison to non-cultivated species, both geographic and human factors were found to impact its genetic variation (Wendel et al., 1992). It is suggested that the genetic grouping of islands that were previously under British
colonial rule is likely due to germplasm exchange, although other genetic clusters are more influenced by geographic proximity (Wendel et al., 1992).

This difference in genetic diversity between islands and mainland populations, with clear relationships between geography, genetic variation, and cultivation history has major implications for both conservation priorities and future breeding programs (see Inoue & Kawahara, 1990; Wendel & Percy, 1990; Shapcott, 1994). As such this study will focus on the islands surrounding the UK to try to identify hotspots of unique variation in comparison to reference mainland sites.

1.9 Thesis Outline

This thesis attempts to analyse the utility of, and best methods for genetic diversity studies in three species of *Trifolium* across the UK and its associated islands, with a review of genetic diversity methods and studies given in Chapter 2. In order to successfully analyse these species, a comprehensive evaluation of the ecogeographic background of the taxa and the sites was conducted to define the areas targeted for collection. The results of this survey in understanding the social and economic importance of the species and in determining conservation priorities in the absence of genetic information are outlined in Chapter 3. Following the collection of the target species based on the information shown in Chapter 3, a genetic diversity study of all three species was conducted in order to calculate the levels of diversity using AFLP markers. Analysis of the levels of diversity and their relationship to environmental variables is also incorporated into the analysis, which is discussed in Chapter 4. Single nucleotide polymorphisms (SNPs) are purported to become increasingly important as markers in genetic diversity assessments and this potential is reviewed in detail in Chapter 5. The utility of SNPs as genetic markers is assessed in Chapter 6, which defines the identification and analysis of new SNPs in wild populations of *T. pratense*.
These assessments were designed to provide an assessment of genetic diversity in these three *Trifolium* species, but may also provide some information to help the in the study and conservation of widespread species. The implications of the findings in the previous chapters are discussed in Chapter 7.
Chapter 2. AN ECOGEOGRAPHIC BACKGROUND TO THREE *Trifolium* SPECIES IN THE UK

2.1 ECOGEOGRAPHY AS A TOOL FOR CONSERVATION ANALYSES

In order to be able to assess the background of target taxa and to optimise any collection activity it is imperative that researchers have as much prior information as possible. Ecogeographic analysis is a process of synthesising background information on both the target area and the species involved in the study, including taxonomic, ecological, geographical and historical data, using the results to predict collection and conservation strategies (Maxted & Guarino, 1997). Ecogeographic surveys will not only show patterns of infra-specific diversity, but also where and within what environmental constraints the species survives and the population fluctuations of the target taxon. This information can be used to highlight potential areas for collection and conservation, either choosing those habitats which represent heterogenous environments within the taxon’s range, those habitats at the greatest risk from genetic erosion or those areas underrepresented in conservation collections. Hence ecogeographic surveys are a vital decision-making tool in both conservation planning and overall conservation success.

2.2 AIMS OF ECOGEOGRAPHIC SURVEY

This chapter aims to conduct a survey based on available literature to create conservation strategies.

Specifically this chapter aims to:

- provide a background to the target taxa; *Trifolium dubium*, *T. pratense* and *T. repens* including taxonomy, distribution and ecology information
- produce an assessment of the vegetation history of the target island sites
reveal the natural distribution of target taxa within the UK
- assess current conservation status and strategies for the target *Trifolium* species within the UK

2.3 TARGET TAXA

Leguminosae Juss. (Fabaceae Rchb.) is a cosmopolitan family of flowering plants, one of the largest in the world and second only to the grasses in terms of adaptation to wide ranges of habitat diversity (Adams & Pipoly, 1980; Graham & Vance, 2003) and economic importance (Heywood, 1985). Within the legume family, the genus *Trifolium* contains between 250-300 species, and includes some of the most economically important species within the legumes (Williams, 1987; Lewis et al., 2008). The centre of origin of *Trifolium* is generally accepted as the Mediterranean region due to the higher species diversity found in this area, with a secondary centre of distribution in north-eastern America, and is now widely distributed throughout the temperate and subtropical regions of the world (Zohary & Heller, 1984; Caradus, 1995).

Of the circa 250 species of *Trifolium* only 16 are known to have been cultivated on a commercial scale (Taylor & Quesenberry, 1996) including white and red clover (*T. repens* and *T. pratense*), the two most economically important pasture legumes in the UK. *Trifolium* was chosen as the target genus for this study in light of both available taxon expertise in this area and previous analysis indicating the use of species of *Trifolium* as model taxa for analysis of ecogeographic diversity effects on genetic diversity patterns. Many undercollected *Trifolium* species have been identified as priorities for evaluation, collection and conservation, including *T. pratense* and *T. repens* (IBPGR, 1985; Francis, 1999). The increasing interest in sustainable farming and organic fertilisers has highlighted the need to conserve the genetic diversity of clover species. *T. pratense* and *T. repens*, as the most cultivated forms of clover in the UK, have been chosen as target species for this study in light of their growing importance.
to UK agriculture and their ubiquitous distribution across the UK (see Figure 2.1). In addition, *T. dubium*, a widespread wild relative of *T. repens* (Bulińska-Radomska, 2000) is included to help distinguish affects on genetic diversity patterns from cultivation.

### 2.4 TRIFOLIUM GENETICS

Within the *Trifolium* genus, *T. pratense*, *T. repens* and *T. subteranneum* L. are the most studied in terms of genetic diversity. Morphological (Bennett, 2000; Pecetti & Piano, 2002) and isozyme (e.g. Martins & Jain, 1980; Rossiter & Collins, 1989; Hagen & Hamrick, 1998; Lange & Schifino-Wittmann, 2000; Yu et al., 2001; Mosjidis et al., 2004) analysis has been widely used to evaluate genetic diversity in *Trifolium* species, with more limited studies published that use molecular markers for this type of analysis.

Less economically important Trifoleae have been assessed to a smaller extent, with a recent study by Malaviya et al. (2005) providing an isozyme study across 25 species and Rizza et al. (2007) using ISSR markers to study six species of clover for genetic diversity and DNA content. Chandra (2008) has recently proposed a set of 43 microsatellites obtained from *Medicago truncatula* and *T. repens* that can be used with relative success across species and genera, paving the way for future microsatellite studies across less well studied members of the Trifolieae. *T. dubium* has had no prior genetic diversity assessment to the authors knowledge (although see Caradus & Mackay, 1989). However, the vast majority of diversity studies across of members of the Trifolieae have shown a generally high level of diversity within populations, even among other inbreeding species such as *T. subteranneum* (Pecetti & Piano, 2002; Piluzza et al., 2005). In an assessment of five species of *Trifolium* in Turkey, the edaphic and geographic range of the species was found to be the most important factor determining the distribution of variation. Much of this variation has been found to be correlated to environmental factors, with widely distributed species showing higher levels of
variation (Bennett, 2000). There have been mixed results in relating genetic diversity between *Trifolium* populations to geographic distances, as will be attempted in this study, with some finding little correlation (Pecetti & Piano, 2002; Greene *et al*., 2004) and others finding that incorporating topographic and geographic barriers when interpreting the results explained much of the distribution of variation (Mosjidis *et al*., 2004).

2.4.1 *TRIFOLIUM REPENS* GENETIC DIVERSITY

Much of the previous work into the diversity of *T. repens* populations uses either morphology to infer genetic diversity (e.g. Burdon, 1980; Jahufer *et al*., 1997), isozyme analysis (Hamrick & Godt, 1989; Lange & Schifino-Wittmann, 2000) or RAPDs (Gustine & Huff, 1999; Gustine *et al*., 2002; Bortolini *et al*., 2006). Recently there has been a shift towards AFLPs and microsatellites, for example by van Treuren *et al*. (2005), and Kölliker *et al*. (2001) who promote bulked AFLP analysis for analysing *T. repens* cultivars in genebanks, or Dolanska and Curn (2004) and George *et al*. (2006) who promote the use of microsatellites for assessing cultivar genetic diversity. Most recently authors have focused on the identification and validation of SNPs in *T. repens* for use in marker-assisted selection (e.g. Cogan *et al*., 2007; Hand *et al*., 2008; Lawless *et al*., 2009) and it is possible that these polymorphisms may also be useful for genetic diversity studies.

In terms of population genetic diversity in *T. repens*, the literature defines a surprisingly high genetic variation within populations and within very small areas for a species that is known to reproduce clonally (e.g. Cahn & Harper, 1976; Ennos, 1985; Gustine & Sanderson, 2001; Kölliker *et al*., 2001; van Treuren *et al*. 2005). In an assessment on a pasture in north Wales, Burdon (1980) attributes the maintenance of high within population diversity to selection pressures occurring on a micro-scale, for example neighbouring plant size, associated species and sheep grazing pressures.
The persistence of *T. repens* in pastures is attributed to both the establishment of new plants via seedling recruitment and spreading via stolons, with traditionally the greatest importance placed upon regeneration from rooted stolons for spread of *T. repens* (Turkington et al. 1979). Seedling recruitment and establishment in grasslands is rare (Burdon, 1983; Cahn & Harper, 1976), leaving vegetative spread to be the primary method of persistence in grassland swards. However, following population diversity analysis of *T. repens* in grasslands it has been found that genetic diversity is high regardless of its clonal nature. The relative importance of clonal propagation leading to the persistence of one successful genotype and restricting high levels of genetic diversity is therefore perhaps less than originally thought.

Gustine and Huff’s (1999) temporal analysis shows a genetic shift in one population between collecting seasons, with significant differentiation between summer and autumn. This work has been followed by Gustine and Sanderson (2001) who analysed the importance of clonal reproductive growth on genetic variation between two growing seasons. They have attributed this temporal change in genetic diversity to the death/dormancy of clones, seedling recruitment, sampling methodologies and most importantly to ecotypes responses to micro-environmental change over seasons. Seedling recruitment throughout the growing season has been stated as a potential factor, perhaps in response to grazing, however as stated above seedling recruitment is generally low, with few seedlings becoming established (Burdon, 1983). However, another explanation for this observation could be due to the differences in age of the plant at collection and hence variation in DNA extraction leading to the observed differential results. DNA isolation and quality are negatively affected by the build up of chemical defences that can accumulate with age (Katterman and Shattuck, 1983; Moreira & Oliveira, 2011), with cyanogenic glucosides and phenolic concentrations shown to increase with age in *T. repens* (Horrill & Richards, 1986).
This within population variation does not equate to between population variation, with little variation found between populations (Gustine & Huff, 1999; Gustine & Sanderson, 2001; Gustine et al., 2002; Sousa-Correia, 2002). The out-breeding, insect pollinated nature of *T. repens* could account for the lack of between population variation with gene flow and migration leading to the homogenised spread of variation.

This study will focus on naturalised, semi-wild populations in the UK, in comparison to the few molecular marker analyses on *T. repens* which are based on managed swards in north America (Gustine & Huff, 1999; Gustine & Sanderson, 2001). Due to the longer period of naturalisation and proximity to the centre of diversity, UK populations may provide even higher levels of diversity than that shown by North American populations which have most likely undergone a more pronounced founder effect.

2.4.2 *TRIFOLIUM PRATENSE* GENETIC DIVERSITY

Many methods have been used to assess genetic diversity in *T. pratense*, with published studies using morphological (Kouamé & Quesenberry, 1993; Kongkiatngam et al., 1995), RFLP (Milligan, 1991), isozyme (Hagen & Hamrick 1998; Yu et al., 2001; Mosjidis et al., 2004; Dias et al., 2008), RAPD (Campos-de-Quiroz & Ortega-Klose, 2001; Greene et al., 2004; Dias et al., 2008), inter simple sequences repeat (ISSR) (Rizza et al., 2007), microsatellite (Dias et al., 2008) and AFLP markers (Kolliker et al., 2003; Hermann et al., 2005).

As an outbreeding and perennial species, similar in life history traits to *T. repens*, it has been shown to contain relatively high levels of within population diversity compared to most other plant species, with low levels of population divergence (Hagen & Hamrick, 1998; Campos-de-Quiroz & Ortega-Klose, 2001; Mosjidis et al., 2004).
The obligate out-crossing nature of *T. pratense* and the ability to colonise a wide geographic range may account for much of the increased variability in this species, although other species with similar traits have lower levels of genetic diversity than *T. pratense* (Hamrick & Godt, 1996). Hagen and Hamrick (1998) suggest this could be due to the intentional maintenance of diversity in crop species, with Kongkiatngam *et al.* (1995) and Yu *et al.* (2001) finding high levels of genetic diversity within *T. pratense* cultivars.

Low levels of population divergence is mostly attributed to introgression *sensu* Arriola (2005), with the potential seed transfer by animals, humans, birds and farm machinery used to suggest that high levels of gene flow is responsible for the homogenous spread of variation between populations (Hagen & Hamrick, 1998). Mosjidis *et al.* (2004) studied wild *T. pratense* populations from the Caucasus and correlated genetic diversity with topographical features to better understand the relationship between genetic diversity and introgression. Geographical barriers were found to influence gene flow, with the predictions on the distribution of genetic diversity made from the use of GIS-derived maps corresponding to isozyme analysis in every case.

Similar findings to those on *T. repens* indicated that year to year variation is surprisingly high, with Hagen and Hamrick (1998) suggesting that this is due to the high turnover of plants in naturalised grasslands, hence collections should be made within the same year where possible.
Figure 2.1. 10km grid square distribution map of target species across the UK (a) *Trifolium pratense* (b) *T. repens* (c) *T. dubium*. Taken from Preston et al., 2002.
2.5 *Trifolium dubium* Sibth.

*Section Chronosemium*
Annuals. Inflorescence many-flowered and then capitulate-globular to ovoid, rarely few-flowered and –racemed. Flowers shortly pedicellate, reflexed in fruit. Bracts minute to 0. Calyx often campanulate, not growing in fruit, 5-nerved; throat glabrous, open; calyx teeth unequally long, the 2 posterior ones distinctly shorter than the three anterior ones. Petals yellow, purple or pink (never white), persistent turning scarious in fruit with spoon- or boat-shaped standard. Pod stipulate, hidden in fruiting corolla, 1-2-seeded (Zohary & Heller, 1984).

*Subsection Filiformia*

**Accepted name**
*Trifolium dubium* Sibth., Fl. Oxon. 231 (1794)

**Synonyms**
*Chrysaspis dubia* (Sibth.) E.H. Greene
*Chrysaspis dubia* (Sibth.) Desv.
*Trifolium filiforme sensu auct.*
*Trifolium filiforme* L. var. *dubium* (Sibth.) Fiori & Paol.
*Trifolium filiforme* L. subsp. *dubium* (Sibth.) Gams
*Trifolium minus* Sm.
*Trifolium praticeola* Sennen
*Trifolium procumbens sensu auct.*
*Trifolium procumbens* L.

**Common names (UK)**
Suckling Clover, Least Hop Clover, Lesser Hop Clover, Lesser Yellow Trefoil, Red Suckling Clover, Yellow Suckling Clover, Lesser Trefoil, Shamrock.

**Description**
Annual, glabrous or sparingly hairy, 20–40cm. Stems many, often brownish, slightly furrowed, somewhat flexuous, erect to ascending, poorly branching. Leaves very short-petioled; stipules herbaceous, ovate, acute, short-adnate to petioles, 3-5mm long; leaflets 0.8-1 x 0.4-0.7cm, obovate, cuneate at base, rounded or slightly notched at apex, dentate around upper part, bluish-green, terminal ones long petiolulate. Peduncles axillary, filiform, much longer than subtending leaves. Heads 8-9 x 6-7 mm, rather dense, 3-20-flowered, hemispherical. Pedicels 1mm or less, erect, later recurved. Calyx 1.5-2mm; tube campanulate, 5-nerved, glabrous; lower teeth
almost twice as long as tube, upper ones shorter than tube. Corolla about 4mm, yellow, becoming brownish in fruit; standard ovate, smooth, conduplicate with funnel-shaped bundle or nerves in each half, entire or obscurely denticulate; wings clawed, shorter than standard. Ovary long-stipitate, longer than style. Pod 1-seeded, with style one-third to one-fourth the length of pod. Seed ellipsoidal, 1.3 mm long, light brown (Zohary & Heller, 1984). See Figure 2.2.

Flowering and fruiting time
May - October (Zohary & Heller, 1984).

Chromosome Number
2n = 4x = 30. This species has been recently shown to be of allotetraploid origin, combining the genomes of T. campestre Schreb. (Hop trefoil, 2n = 2x = 14) and T. micranthum Viv. (Slender trefoil, 2n = 2x = 16) (Ansari et al., 2008, see also Ellison et al., 2006).

Distribution
Holland, Scandinavia, Belgium, British Isles, France, Portugal, Spain, Italy, Hungary, Poland, Romania, Czechoslovakia, Balkan Peninsula, Turkey, Cyprus, Israel, south and central Russia, Caucasus. Introduced into the USA and Canada (Zohary & Heller, 1984; Gillett & Taylor, 2001). See Figure 2.3.

Habitat
Sandy places, edges of pastures, roadsides from 100 to 1300 m (Zohary & Heller, 1984; Gillett, 1985; Gillett & Taylor, 2001). Frame (2005) asserts that T. dubium is able to naturalise in dry areas, however in very dry areas such as western Australia, Fortune et al. (1995) found that T. dubium was one of the most common annual species in wetter pastures.

Soil
Shallow coarse-textured soils (Frame, 2005).

Reproduction
Self-pollinated (Gillett & Taylor, 2001).
Genepool
There is little reference to *T. dubium* in the literature in terms of genetic relationships to other species, however, in a recent morphological assessment of 15 *Trifolium* species *T. dubium* was found to group most closely with *T. hybridum* L., *T. repens*, *T. strepens* Cr., *T. campestre* Schreb., *T. patens* Schreb. and *T. fragiferum* L. (Bulińska-Radomska, 2000).

Uses
Essentially a wild species, *T. dubium* is little cultivated due to its low competitiveness in highly managed swards, and low/medium tolerance to prolonged drought. However, in terms of grazing utility *T. dubium* adds to early sward yield and can be an important part of seed mixtures to be used in heterogenous sites, and as such been used as a minor part of seed mixtures in sandy soils in dry areas of the USA (Frame, 2005). In fact Hermann (1953) records *T. dubium* as one of a group of minor species of *Trifolium* adapted to cultivation in the USA and Canada. In addition, *T. dubium* is listed as one of the *Trifolium* species used in honey production (Woodgate *et al*., 1999). Of 13 *Trifolium* species evaluated by Caradus (1995), *T. dubium* was one of the most frost tolerant species, indicating potential for use in breeding programs. As a close relative of *T. repens*, (Bulińska-Radomska, 2000) *T. dubium* is a potential gene donor to this economically important crop species.
Figure 2.2: *Trifolium dubium* a) Line drawing (Zohary & Heller, 1984); b) Flower (http://popgen.unimaas.nl) and c) Entire plant (http://www2.uni-jena.de).
Figure 2.3 World distribution of *Trifolium dubium*. Data taken from ILDIS World Database of Legumes (Bisby et al., 2010).
2.6 *Trifolium pratense* L.

*Section Trifolium*

Annuals or perennials. Heads falsely terminal or axillary, sessile or pedunculate. Flowers bractless, very rarely with few bracts at the base of head. Calyx 10-20-nerved, hairy or rarely glabrous with unequal or equal teeth; throat usually closed by bilabiate callosity, if open then provided with hairy ring or narrowed by protruding ring. Corolla mostly partly united. Pod included in calyx tube, 1-, very rarely 2-seeded. Dispersal by single fruiting calyces or by entire fruiting heads (Zohary & Heller, 1984).

*Subsection Pratensia*

*Accepted name*

*Trifolium pratense* L., *Sp. Pl.* 768 (1753)

*Synonyms*

*Trifolium borysthenicum* Gruner

*Trifolium bracteatum* Schousb.

*Trifolium lenkoranicum* (Grossh.) Roskov

*Trifolium pratense* L. var. *lenkoranicum* Grossh

*Trifolium ukrainicum* Opperman

*Common names (UK)*

Red clover

*Description*

Perennial with fusiform root and short stock but without runners, patulous- to apressed-pubescent or glabrescent (with hairs whitish, often arising from tubercules), 20-60cm. Stems many, arising from basal leaves, simple or branched, erect or ascending or decumbent, furrowed to angular. Lower leaves long-petioled, middle and upper ones short-petioled, uppermost ones subsessile; stipules ovate-lanceolate, adnate to petioles, membranous, with green or red nerves, hairy or glabrescent, free portion abruptly mucronulate or cuspidate, usually much smaller than lower part; leaflets very short, petiolulate, mostly 1.5-3(-5) x 0.7-1.5cm, obovate to obovate-oblong or broadly elliptical, rarely (and in upper leaves only) oblong-lanceolate, obtuse or acutish or retuse, almost toothless, appressed hairy on both faces or only beneath, often spotted. Heads terminal, solitary or in pairs, globular or ovoid, mostly involucrate by the stipules of reduced leaves, rarely pedunculate. Flowers many, usually 1.5-1.8 cm long, dense. Calyx tubular-campanulate, 10-nerved, whitish-green, sometimes with a reddish tint, often appressed- or patulous-hairy (rarely tube glabrous); throat with slight annular (epidermal not callous), hairy thickening; teeth unequal, erect, subulate, blunt, often ciliate or patulous hairy, the lower one longer than tube and others. Corolla reddish-purple to pink, rarely yellowish-white or white; standard longer than wings and
keel, notched. Ovary 2(1)-ovuled. Pod ovoid, membranous cartilaginous, shiny upper part. Seed 1, oblong-ovoid, tuberculate, yellow, brownish or violet (Zohary & Heller, 1984). See Figure 2.4.

**Flowering and fruiting time**
March - September (Zohary & Heller, 1984)

**Chromosome Number**
2n = 2x = 14 (Sato et al., 2005).

**Distribution**
Distributed across Europe (except the extreme north), western Asia and the Mediterranean region (except its south east part). Introduced into USA, Australasia and eastern Asia. Very widespread and mostly cultivated throughout the northern hemisphere and elsewhere (Zohary & Heller, 1984; Gillett & Taylor, 2001). See Figure 2.5.

**Habitat**
Meadows, grassy plots, roadsides, near water, glades, borders of fields and forest margins (Zohary & Heller, 1984; Gillett & Taylor, 2001).

**Reproduction**
Cross-pollinated by both Bombus sp. and A. mellifera (Gillett & Taylor, 2001). T. pratense has a gametophytic self-incompatibility system, with populations consisting of heterozygous individuals (Townsend & Taylor, 1985), however some recently bred tetraploid individuals do have self-compatibility (Frame et al., 1998). T. pratense does not proliferate by stolons, with seed most important for spread on a local scale. Seed is important in the colonisation of new areas for T. pratense, spread on a local scale by dehiscence, trampling and to a minor extent wind, and on a larger spatial scale by grazers and birds (Hagen & Hamrick, 1998). Due to the impermeable testa of most legumes, seed has high longevity, with life spans of T. pratense seeds recorded up to 81 years (Gillett & Smith, 1985).
Gene pool
Closely related species are listed as *T. diffusum* Ehrh., *T. pallidum* Waldst. and *T. sarosiense* Hazsl., with studies showing that *T. pratense* is more closely related to annual species than perennials (Cleveland, 1985; Caradus & Williams, 1995).

Soil
*T. pratense* tolerates many soil types and environmental conditions throughout temperate zones, however wet, acid and shallow soils are limiting for this species (Frame, 2005). Grows best on loam soil with good drainage, and while it grows best on soils of pH 6.6-7.6, is better adapted than other forage crops such as *Medicago sativa* at lower levels of pH (Duke, 1981).

Uses
Due to its vigorous growth and high nutritive value, especially in the first year of sowing (Frame, 2005), *T. pratense* is one of the most cultivated species of *Trifolium* used in both monocultures and mixed swards. In addition, due to the deeper tap root of *T. pratense* it is an efficient utiliser of soil water, hence can persist in drier areas in comparison to most other species of *Trifolium*. Originally from south-eastern Europe (Fergus & Hollowell, 1960), the first cultivation of *T. pratense* in northern Europe occurred around 1650, quickly followed by the trading of seed (Merkenschlager, 1934). Historically, due to both its high nutritive value and adaption to wide environmental conditions, the cultivation of *T. pratense* has been high, however production has declined due to high intensity farming systems and the reliance on nitrogen fertilisers (Frame, 2005). However, in light of the recent interest in sustainable, low-input and organic farming as well as its use in medicinal products, there has been a rejuvenation of *T. pratense* cultivation (Frame, 2005).

In addition to its use as forage, *T. pratense* is described as a folk medicine, being used as a diuretic or a sedative, with Native Americans said to use the plant for burns and sore eyes (Duke, 1981).
Figure 2.4: *Trifolium pratense* a) Line drawing (Zohary & Heller, 1984); b) Flower (http://www.naturedirect2u.com) and c) Entire plant (http://www.missouriplants.com).
Figure 2.5 World distribution of *Trifolium pratense*. Data taken from ILDIS World Database of Legumes (Bisby et al., 2010).
2.7 *TRIFOLIUM REPENS* L.

*Section Lotoidea*

Annuals or perennials. Stems simple or branched, some rhizomatous or scapose. Leaves 3-5(-9)-foliate. Inflorescences umbellate, capitate or spicate, sometimes scape-like. Bracts entire, bifid or crenulated, rarely inconspicuous or absent (in Subsect. *Neolagopus*). Pedicels long or short, rarely absent. Calyx normally 10-nerved, rarely 5- or 20- nerved; calyx throat naked, open; calyx teeth generally equal, sometimes unequal. Corolla variously coloured, marcescent. Ovary with 1-12 ovules, sessile, rarely stipulate. Pod indehiscent or opening by ventral suture or also by both sutures (Zohary & Heller, 1984).

*Subsection Lotoidea, series Lotoidea*

*Accepted name*

*Trifolium repens* L. Sp. Pl. 767 (1753)

*Synonyms*

*Amoria repens* (L.) Presl, Symb. Bot.1:47 (1830)

*Common names (UK)*

White Clover, Dutch Clover, Ladino Clover.

*Description*

Perennial, glabrous or glabrescent, 10-30cm. Stems rhizomatous, prostrate, long-creeping and rooting from nodes. Leaves long-petioled, sometimes petioles hairy; stipules broad at base, ovate-lanceolate, with subulate upper part, scarious with reddish or green nerves; leaflets 0.6-2.5(-4) x 1-1.5cm, broadly obovate to orbicular, obcordate, mostly emarginated, margin sharply serrulate, with parallel, lateral veins (featuring light or dark marks) forked towards margin. Inflorescences 1.5-3.5cm in diam., umbellate, 20- to many flowered, rather loose, nearly globular. Peduncles as long as or much longer than subtending leaves, weak. Bracts shorter than pedicels, ovate-oblong, acuminate. Pedicels as long as or longer than the calyx tube, somewhat deflexed from early flowering, strongly deflexed after anthesis, glabrous or hairy. Calyx 3-5mm, glabrous; tube campanulate, 6-10-nerved; teeth about as long as tube, somewhat unequal, triangular-lanceolate. Corolla 4-13mm long, white, yellow or pink; standard ovate-lanceolate, oblong, rounded at apex; wings somewhat spreading. Pod 4-5mm, usually 3-4 seeded, linear-oblong, constricted between seed. Seeds ovoid to reniform, brownish (Zohary & Heller, 1984). See Figure 2.6.

*Flowering and fruiting time*

March - September (Zohary & Heller, 1984; Gillett & Taylor, 2001).
Chromosome Number
2n = 4x = 32 (Hand et al., 2008). Prior to 2006 the identification of the progenitors of \textit{T. repens} has been inconclusive. However a recent phylogenetic study on the genus implicates \textit{T. pallescens} (2n=16) and \textit{T. occidentale} (2n=16) as the potential progenitors (Ellison et al., 2006).

Distribution
Widely distributed in moist temperate zones Europe, central and northern Asia, all Mediterranean countries and north Africa (Frame, 2005). Introduced in north America, southern Latin America, Australasia and Japan (Zohary & Heller, 1984; Frame, 2005). See Figure 2.7.

Habitat
Damp and swampy soils, also on lawns, grassy places, roadsides, pastures etc (Zohary & Heller, 1984). This species is able to grow in most habitats, and is able to tolerate poor conditions better than other species of \textit{Trifolium} (Duke, 1981).

Reproduction
Outcrossing species pollinated primarily by \textit{Bombus} sp., \textit{Apis mellifera}, and to a lesser extent some Lepidoptera. As a general rule \textit{T. repens} is self-incompatible, with the few inbred individuals setting little seed and suffering from high levels of inbreeding depression (Atwood, 1940). Propogation is thought to be primarily through stolons under grazing management, forming large clonal patches, as seedlings in long-term pastures are rare, often not becoming established (Burdon, 1983; Chapman, 1987). Over longer distances seed becomes important in spread, on a local scale by dehiscence, tramping and to a minor extent wind, and on a larger spatial scale by grazers and birds (Williams, 1987).
Gene pool

*T. repens* is closely related to *T. uniflorum* L., *T. occidentale* Coombe and *T. nigrescens* Viv due to the ease of hybridisation with these three species. Hybridisation is difficult but possible with three other species, *T. ambiguum* M.B., *T. isthmocarpum* Brot. and *T. hybridum* L. (Cleveland, 1985; Williams, 1987).

Soil

Grows vigorously across a wide range of soils and environmental conditions, but less well in poorly drained soils or drought prone soils (Duke, 1981). Adequate pH (5.8-6.0 on mineral soils and 5.5-5.8 on peaty soils) is required (Frame, 2005).

Uses

*T. repens* is one of the most nutritious and widely distributed forage legumes crops in the world, used both alone or in mixtures to provide feed and as a cover crop for soil stabilisation (Duke, 1981). In the UK the most economically important forage legume is *T. repens*, although cultivation has been fairly recent, with its first known domestication in the Netherlands in the 16th century, with seed widely traded since the 17th century (Caradus, 1995). Much of the *T. repens* seed sown in the UK in the early 1900s was European varieties (mainly Dutch varieties with some New Zealand varieties) until the superiority of English wild varieties was recognised, generally superseding the foreign varieties (Caradus, 1995). *T. repens* has a high nutritive value and is tolerant to medium intensity grazing and as such is one of the economically important species of pasture legume in moist temperate regions of the world (Frame, 2005). Its use has declined in Europe since the 1950s due the preference for grass swards intensively fertilised with nitrogen however the recently this has been reappraised due to its potential for use in low intensity and low input farming systems (Frame et al., 1998). In addition to its use as forage crop, *T. repens* is used as a folk medicine for such ailments as gout and rheumatism (Duke, 1981).
Figure 2.6. *Trifolium repens* a) Line drawing (Zohary & Heller, 1984); b) Flower (http://www.plant-identification.co.uk) and c) Entire plant (http://webup.univ-perp.fr).
Figure 2.7 World distribution of *Trifolium repens*. Data taken from ILDIS World Database of Legumes (Bisby et al., 2010).
2.8 TARGET ISLAND SITE - ISLES OF SCILLY

The Isles of Scilly are a group of around 200 low lying islands and rocks found 45km from the coast of Lands End, the south-western tip of England. The largest of the islands; St Mary's, Tresco, St Martin's, St Agnes, Bryher and Gugh are inhabited with a total population of 2,153 (Office for National Statistics, 2001).

The Isles of Scilly are formed of a granite cupola, submerged by rising sea levels to become the archipelago seen today (Thomas, 1985). The Isles of Scilly went through several stages of submergence, and is thought to have been known to the Romans as sindos diiaros, the land, describing a large island that remains now as St Mary’s, Samson, Bryher, Tresco, St Helens, St Martins and the Eastern Isles (Thomas, 1985).

Deep water is sometimes found close to shore and, with prevailing strong winds from the south west, the outside of the archipelago is highly exposed to the sea (Cooper, 2006). In comparison, the shores inside the archipelago are protected by warm, shallow straits, with the eastern coasts between islands having very sheltered conditions (English Nature Report, NA113).

The influence of the sea modifies the climate with the range as little as 9°C between the winter and summer months, much lower than that of much of inland UK (see Figure 2.8). The Isle of Scilly’s mild climate is characterised by a rarity of frosts and a higher sunshine record.
in comparison to most of the UK, however the proximity to the sea, along with the
topography and south-westerly winds limits extremely high temperatures with average
temperatures of 19°C in July and August (Met Office, 2005).

The lack of frosts, the maritime influence and the high sunshine record defines a unique
environment in the UK, and hence provides potential for novel genetic diversity on these
islands.

![Graph of Climate of St. Mawgans, average 1971-2000. Source: Met Office data.]

2.8.1 ENVIRONMENTAL HISTORY

When man first arrived on the Isles of Scilly is subject to much debate, however most
are in line with Thomas (1985), who identified 2000 BC when remaining neolithic peoples,
who had been in contact with early Bronze Age cultures, crossed to the islands. This suggests
that seed, and potentially farming methods, will have been transferred from this time, with
reconstructions of Bronze Age boats showing them capable of not only transferring seed but also animals. From this time transfer between the Isles of Scilly and mainland is likely to have been frequent, with the high incidence of Roman goods indicating the Isles were used as a trade route in Roman times (Ashbee, 1974).

At the time of the first settlers, the Isles of Scilly consisted of a wooded landscape, interspersed with heath and little grassland, with pollen analysis of the two remaining peat areas on St Mary’s defining an early history of oak climax community (Dimbleby et al., 1981). The island will have had little resemblance to the archipelago seen today, as some report that the islands still may have been joined as little as 1000 years ago (Cooper, 2006).

From 17th century references describe a treeless Scilly, freeing the land for cultivation and pasture. The timescale of the deforestation is unclear, but there was likely to have been fluctuations with some woodland left until the 1st millennium, with pollen analyses showing the persistence of woodland, although with a more restricted number of species (Dimbleby et al., 1981). Moreover Ashbee (1974) suggests that cultivation and grazing animals played a secondary role to hunting and collecting for the early settlers. The clearance to produce grassland allowed the growth of cattle and sheep numbers on the islands, known to be present from early in the 2nd millennium, coming with the first settlers, however early records also describe the early use of seaweed as forage (Ashbee, 1974; Thomas, 1985). The quality of the soils of the grazing land may have tempered the population expansion after colonisation, with the resources consisting of poor acidic soil, worsened by desiccation and sand blows.

The situation of the Isles on the western tip of England, and therefore their strategic importance during wars and for pirates made life on the islands a hardship. A turbulent history of human colonisations and recolonisations of the islands throughout the subsequent history is described in detail by Thomas (1985) and Ashbee (1974).
Pollen analyses of the upper level of peats show a decrease in arboreal taxa and an increase in herbaceous taxa, indicating an increase in agriculture (Dimbleby et al., 1981). The lack of Leguminosae pollen in the analyses is most likely not due to a lack of the taxa, but due to the pollen being frail and difficult to identify, so an increase in herbaceous pollen has to act as a surrogate indicator for Leguminosae taxa (Dimbleby et al., 1981).

During the Iron age, cultivation and grazing land persisted on the Isles, even in the face of intense soil degradation, which limited farmland to lowland areas. This farmland remained until the advent of flower farming, which overtook most of the farming land on the Isles.

Recent history

The main economy for the Isles is now in the tourist industry (85%), with farming currently the sixth largest employer on the Isles (Office for National Statistics, 2001). The closing of the abattoir and dairy signalled the end to the open habitats on the island (Isle of Scilly Wildlife Trust, 2006). This decline in farming and grazing has led to gorse and bracken overtaking many of the old coastland fields, with grazing in a drastic decline. In the most recent assessment, there are 1.6 km$^2$ of permanent grassland and 10.48 km$^2$ of rough grazing land associated with farms on the Isles of Scilly (Defra, 2005).

2.8.2 Conservation

Conservation designations

The Isles of Scilly are the UKs smallest area of outstanding natural beauty at 16 km$^2$.

Of the 26 Sites of Special Scientific Interest (SSSI) located on the Isles of Scilly (covering a total of around 5.50 km$^2$), 4 SSSI’s namely Annet, Big Pool and Browarth Point (St. Agnes), Higher Moors and Porth Hellick Pool (St. Mary’s) and Tean are defined as covering „neutral grassland – lowland” (0.67 km$^2$).
The entire archipelago is designated a Special Area of Conservation (SAC) however the designation is based upon intertidal zones.

Current threats to diversity

The Isles of Scilly provides a large range of habitats, from intertidal, to grassy heaths and arable lands and is unique to the UK in terms of its isolation, situation and warm maritime climate, which indicates the importance of conservation of plant genetic resources on the Isles of Scilly in UK priorities.

The increase in flower farming and general decline in livestock farming has led to the loss of much of the grassland habitat, in addition to the lack of grazing which has allowed the invasion of scrub and heath of many of the wild grassland sites. One current conservation project involves the reintroduction of grazing to wild sites in an attempt to reduce scrub invasion and maintain grasslands. The Isles of Scilly Wildlife Trust has already reintroduced livestock to areas of Bryher, St. Agnes and St. Mary’s for wild grazing (Isles of Scilly Wildlife Trust, 2006).

2.8.3 Presence of species

The Isles of Scilly, more so than the other islands covered by this assessment contain a high diversity of *Trifolium* species, 26 species in total. All three target species are known to be present on the Isles, with both *T. repens* and *T. dubium* common to all islands (Stace et al., 2003). *T. pratense* is less common on the Isles and now survives in isolated patches, most likely to have arisen from introductions and escapes from cultivation (Parslow, R. 2007, pers. comm. 7 March). *T. repens* is frequent across all islands and in some areas occurs in a form described as var. *townsendii*, with purple flowers (Lousley, 1971).
2.9 Target Island Site - Outer Hebrides

The Outer Hebrides or the Long Isle lies to the North West of Scotland, running for 200km from Lewis in the north to Barra Head in the south. A rising sea level has led to the interruption of the island arc forming a sweeping archipelago of around 119 named islands (Ritchie, 1991). The main islands in the Outer Hebrides; Lewis, Harris, North Uist, Benbecula, South Uist and Barra are surrounded by numerous smaller islands, including St Kilda, Flannan Islands and Sula Sgeir, totalling 3022 km² (General Register Office Scotland, 1991).

The Outer Hebrides are mostly composed of rocks from the Precambrian age, comprising of mostly metamorphic gneisses and igneous rocks, named Lewisian from the Isle of Lewis where they were first described (Gribble, 1991). Lewisian rocks date from between 2800 mya to 1600 mya and are some of the oldest rocks in Britain (Hudson, 1991). The soil on the Outer Hebrides however is much younger, formed after the last glacial period, with some soils such as peat and sandy soils (Machair) formed even more recently. The areas of gently

---

**Table**

<table>
<thead>
<tr>
<th>Latitude</th>
<th>56° 48’ - 58° 31’ N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longitude</td>
<td>6° 8’ - 7° 40’ W</td>
</tr>
<tr>
<td>Distance from mainland</td>
<td>64.37 km</td>
</tr>
<tr>
<td>Total area</td>
<td>3022 km²</td>
</tr>
</tbody>
</table>
undulating Machair that predominate on the western coastlines of the southern islands, provide fertile grazing land amongst the mostly boggy and rocky areas throughout the rest of the islands (Ritchie, 1991). Indeed, the Outer Hebrides provide an area of geographic contrasts, with the islands of Lewis and north Harris in the north of the chain having mountains of an average of 550m high, whilst those on the southern islands, found along the eastern edges rarely top 300m. Whilst Machair covers around 10% of the area in the Uists, the majority of the vegetation in the northern isles is composed of infertile blanket bog and moorland formed on peat soils, providing a distinction in vegetation types between the north and south of the archipelago (Ritchie, 1991).

![Temperature and Rainfall Chart](image)

Figure 2.9. Climate of Stornoway, average 1971-2000. Source: Met Office data.

The cool wet climate of the Outer Hebrides is highly influenced by the sea, with the warmest months (July and August) reflecting the highest sea temperatures, rather than the sunniest months (May and June) (Angus, 1991). The coldest months (January and February)
are similarly affected by the lag in temperatures produced by sea temperatures and, as buffered by the sea, the temperature range in the Outer Hebrides (8.8°C) is one of the smallest in the UK (see Figure 2.9). This low range in temperatures corresponds to both relatively frost-free winters and long periods without frost equalled only by the Isles of Scilly and the southern Inner Hebrides, leading some authors to consider this the basis for the persistence of plants that are more commonly found in more southern areas (Angus, 1991). The oceanic climate leads to higher and more evenly distributed precipitation, and higher humidity and wind speeds, with these islands noted as having the most oceanic climate of Eurasia (Birse, 1971).

2.9.1 ENVIRONMENTAL HISTORY

Due to the lack of unambiguous glacial deposits in the Outer Hebrides, there are no accepted records for vegetation history prior to the last glaciation (Birks, 1991). However, detailed mapping indicates that part of northern Lewis was ice free during the last glaciations, and as such some authors have proposed that there is the possibility that some plants may have survived the last ice age (Harrison, 1939, 1953; Dahl, 1955). Fossil-assemblages on Lewis and Hirta suggest a herb dominated landscape between around 14,000-27,000 before present (BP) (Birks, 1991). Pollen spectra from the late glacial period (13,500-10,000 BP) reflects species poor, subalpine regions of Norway, although it is likely the vegetation is more diverse than is seen today, potentially due to the fertile deposits left in the wake of the retreating glaciers (Birks, 1991). The reconstruction of the floristic history of the islands over the last glacial period suggests that there was never any extensive or continuous tree cover on the Outer Hebrides, rather that trees existed in localised areas. Instead species-rich grasslands dominated by ferns and tall herbs gave way to an increase in acid grasslands, bogs and heaths, (Birks, 1991). Signs of human influence on vegetation history are known from 5000BP, with
the influence adding to the expansion of grassland and pastures, with some evidence of cereal
cultivation at around 1700BP, coinciding with the colonisation of the Isles by Scots. Birks
(1991) defines the vegetational history of the Outer Hebrides as one of “progressive
impoverishment”, with the cumulative effects of acidic bed rock, bog development, distance
from the mainland, limited habitat range, increasing storm frequency and human impact has
led to the extinction of many species previously identified on the Outer Hebrides and the
species poor islands that are found today.

2.9.2 Conservation

Conservation designations

The Outer Hebrides consist of 55 Sites of Special Scientific Interest (SSSI) and 13
Special Areas of Conservation (SAC). Of these SACs, seven focus specifically on the
conservation of grassland, Machair or cliff vegetation, areas that contain the target species of
this study.

Current threats to diversity

Machair, lowland low intensity grassland, provides some of the most diverse and fertile
land on the Outer Hebrides. Machair itself has a long history of management by local
communities, with recent practices involving intensive seasonal grazing with periods of low
input rotational cropping of local barley, oats and potatoes (Ritchie, 1991). This management
sustains the mixture of plants and communities that are found on the Machair and as such
conservation plans incorporate the involvement of managers to maintain traditional farming
practices to preserve the Machair. With the current trend moving away from a system of
agriculture that is only just economically viable, conservationists are working to promote
traditional cropping practices to retain the Machair systems in their current highly diverse
state (Angus & Hansom, 2010).
Between 1960 and 1980, areas of reseeding have taken place on the Outer Hebrides, mostly on Lewis but in some eastern hill areas of the Uists (Boyd & Boyd, 1990; Angus & Elliott, 1992). Shell-sand (Machair) has been transported to these areas and, with the addition of fertilizer and seed, has produced fertile arable pasture land. Surveys in 1975 of 101 reseeded pastures in Lewis found that *Trifolium repens* makes up over 20% of these reseeded areas (Boyd & Boyd, 1990). While reseeding is acknowledged as a concern in terms of conservation as it can reduce species diversity and habitat diversity on a larger scale (UK Biodiversity Group, 1999), only recently have the effects on genetic diversity through reseeding been acknowledged as a threat, with farmers now urged to reseed from local harvests (Machair Life+, 2010). However, much of this is confined to maintaining the local varieties of cereal crops, with no comments on the effects of reseeding to populations of clover and other grassland species.

2.9.3 Presence of species

The Outer Hebrides contain seven *Trifolium* species, of which only four are native, including the three target taxa included in this study. All three species are native and found throughout the main isles of the Outer Hebrides, from Barra to Lewis. Local plant lore names the Seamrag or clover as a weather indicator, “Tha’n t-seamrag a’ pasgadh a còmhdaich ro thuiteam dóirteach” (The shamrock is folding its garments before a heavy downpour) indicating the long history of clover on the islands, as well as the significance to the people (Bennett, 1991). *T. repens* is found all over the Outer Hebrides and is common throughout, as is *T. pratense*, with both species found in the Machair and grasslands. *T. dubium* is found throughout the main island chain but is mostly restricted to roadsides (Mullin & Pankhurst, 1991).
2.10 TARGET ISLAND SITE - SHETLAND ISLES

Shetland comprises over a hundred islands covering a total of 1440 km$^2$, extending from Muckle Fugga in the north to Sumburgh head in the south, in addition to the Fair Isle 39km to the south, which lies between Orkney and Shetland (Ritchie, 1997). Shetland lies at the same latitude and southern Greenland, just 292 km from the west of Norway, and 150km north of the Scottish mainland, the northernmost isles of the British Isles. The main islands; Mainland, Yell, Unst, Fetlar, Whalsay and Bressay, in addition to Burra, Tronda, Papa Stour, Muckle Roe, Fair Isles and Foula, are all permanently inhabited.

Glaciation moulded much of Shetlands landscape, responsible for the deepening of the valleys, moraine deposits and the creation of numerous lakes, with the glacier at its most extreme covering all except the tops of some hills (Small, 1983).
At the time of glacial retreat, Shetland was not joined by land to either Scotland or to northern Europe, hence all flora and fauna will have colonised the land from 15,000 to 20,000 years ago (Berry & Johnston, 1980a). The core of the Shetland landscape consists of Palaeozoic shists and gneisses, contributing the rolling, mainly moor and peat-bog covered landscape, rising in places to over 450m (Small, 1983; Scott & Palmer, 1987). Wide valleys in the central mainland, Yell and Unst arise from limestone bands, providing some of the most continuous fertile soil in Shetland (Small, 1983). Relatively fertile, gently undulating areas occur in the west on the Devonian red sandstone. Scott and Palmer (1987) note Fetlar and Unst and north Ronas Hill as areas with drier heathy habitat, and as such valuable environmental diversity for Shetlands plant communities.

Figure 2.10. Climate of Lerwick, average 1971-2000. Source: Met Office data.

Shetland is extremely exposed and as such is subject to considerable erosion, sand deposition and gales (Small, 1983; Scott & Palmer, 1987). Branching inlets or „voes‘ extend
long distances inland, and prevent any place in Shetland from being more than 5km from the sea (Scott & Palmer, 1987).

The prevailing winds are south-westerly and contribute, along with north Atlantic drift to the relatively warm winters for the latitude. Latitude contributes to the climates, with sunshine hours just over 5 hours on the shortest day, but up to 18 hours in midsummer. The warmer winters, but cool limited summers restrict the variety of cultivation in the Shetland Isles. The generally low relief of the islands contributes to lower rainfall than other such exposed islands, with precipitation over 248 days a year, but totalling only 1025mm (see Figure 2.10).

2.10.1 Environmental history

Orkney to the south of Shetland acted as a stepping stone, allowing the early settlers to arrive around 3,000 BC, with the first evidence of settlement in the Shetland Isles stone age cairns, unique to Scotland (Small, 1983). Around this time there existed a crofting type economy, with the remains of oxen, sheep and large quantities of grain (Small, 1983). The scrub cover of mainly birch and hazel entered a drastic decline from around the time of the first settlers, with consequent spread of moorland species, according the pollen analysis, suggesting clearances undertaken by the first inhabitants of Shetland (Scott & Palmer, 1987).

Orkney, with flatter and more fertile soils consisted of a dispersed agricultural landscape, however the less fertile soils and more extreme climate of Shetland gave rise to the clustering of houses near the few fertile arable areas, surrounded by the large expanses of grazing areas. Large scale migration occurred in the fifth and sixth centuries BC attributed to food shortages throughout Europe, however due to the climate and limited resources, this movement didn’t alter the subsistence economy present on the Isles which had persisted for over 2000 years. At around 800 AD the arrival of the Vikings heralded the biggest change in
the Shetlands anthropomorphic history, with Shetland as the closest land to the source of the newcomers increasing its importance and subsequently population to over 20,000, around which it remains today.

**Recent history**

Under increasing influence from Scotland’s language and customs, Shetland came under Scottish rule in 1469, signed over as dowry in the marriage of James III to Margaret of Denmark. Oil exploration in the 1960s and 70s found oil off the Shetland coast, improving the economic state and halting the decline of population. In a recent assessment, the main contributors to the economy are fisheries and oil production, with agriculture and tourism contributing just a tenth of that generated by the former two industries (Shetlands Island Council, 2005). However, agricultural land is still a major constituent of the Shetland landscape with 1230.67 km² of rough or common grazing.

2.10.2 Conservation

**Conservation designations**

Conservation areas on Shetland consist of 3 National Nature Reserves (NNR), 81 Sites of Special Scientific Interest (SSSI), 12 Special Protection Areas (SPA) and 12 Special Areas of Conservation (SAC) mainly designated as bird breeding grounds. Two SSSIs, namely Aith Meadows and Breckon are designated on the basis of their grassland habitat covering a total of 0.82 km² along with one SAC, the Keen of Hamar covering 0.39 km².

**Current threats to diversity**

Shetland, due to its isolation and extreme environment, supports a fragile ecosystem of small populations with little potential for the replacement of lost genetic and species diversity (Berry & Johnston, 1980b). The flora of Shetland is distinctly impoverished in species terms,
with just 568 established species in comparison to the around 3000 species in the total British flora (Stace, 1991; Scott & Palmer, 1987).

According to Scott and Palmer (1987) sheep overgrazing is contributing to the loss of rare species and habitat decline, with some of the most diverse pastures reduced to barren ground or to the same moorland that covers most of Shetland. The problem of the remaining crofting lifestyle on Shetland and the lack of available grazing land contribute to the conservation dilemma. Scott and Palmer (1987) note that areas of the Unst serpentine and North Roe have been reseeded with a mixture of both *Lolium perenne* and *T. repens*, damaging both the species heterogeneity of the sites and the potential genetic diversity of these species in the Shetland Isles. This reseeding has been undertaken to improve grazing sites across many of the upland heaths of Shetland, and are identifiable as fenced geometric patches of green on the otherwise heath covered hillsides.

### 2.10.3 Presence of Target Species

All three of the target species are listed as present in the Shetland Isles. *T. repens* is common, occurring in pastures and dry sandy soils and also in some damper areas such as marshes. Scott and Palmer (1987) notes the use of *T. repens* in improving some of the unusable peat tracts in the Shetland Isles. *T. pratense* occurs on Shetland in dry pastures, sea cliffs and roadsides. Both species are listed as native, with Mouat and Barclay (1793) providing the first written evidence of existence on the Isles (Unst) stating that „there is little or no sown grass, but the meadows are rich in red and white clover“. *T. dubium* is listed as rare, and is recorded as a firmly established colonist, possibly native, with the first evidence a collection in 1924 from Scousburgh (Scott & Palmer, 1987). *T. dubium* persists in sandy areas and arable ground around the Bay of Quendale (SSSI) and in disturbed areas throughout the
rest of Shetland, formerly existing near the airport in Sumburgh before runway expansion (Scott & Palmer, 1987).

2.11 TARGET ISLAND SITE - SKYE, INNER HEBRIDES

The Isle of Skye is the largest island in the Inner Hebrides, which lie close to the west coast of Scotland, running for 240km from Skye in the north to Islay in the south. The main islands in the Inner Hebrides; Skye, Mull, Jura and Islay are surrounded by numerous smaller islands, with the isles of Raasay, South Rona and the Small Isles included in Skye’s parishes, totalling 1736.83 km$^2$ (Darling, 1955).

The predominant geology of Skye was formed around 60 million years ago, when the Cuillins in the south and the northern and central basalt plateau were formed. The geology of the Isle was subsequently moulded by the ice sheets, which divided Rassay from Skye and Skye from the mainland (Armit, 1996). The basalt plateau in the north of Skye forms brown loam soils which, to the west, produce the Glens of Tottrenish, Glen Haultin, Glen Romesdal, Glen Hinnisdal and Glen Uig which are some of the best pasture

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Latitude</td>
<td>57° 16’ N</td>
</tr>
<tr>
<td>Longitude</td>
<td>6° 12’ W</td>
</tr>
<tr>
<td>Total area</td>
<td>1736.83 km$^2$</td>
</tr>
</tbody>
</table>
grounds for cattle and sheep on Skye (Darling, 1955). In areas such as the Vaternish peninsula, Duirinish and Bracadale the basalt creates relatively fertile grazing land, however here is a greater cover of peat than in the steeply sloping Totternish. The Cuillins, a spectacular hill chain, form the central part of Skye and are composed of Gabbro and red granites. The southern part of the islands consists of limestones and gneiss rocks which, where the drainage is good, yield valuable agricultural land, especially west of Sleat at Ord (Darling, 1955). Raasay, lying between Skye and the mainland, is one of the few areas in the Western Isles with deep soils, especially on the east of the island, however this deep soil has afforded both deep tilling and direct reseeding of the Glam of Raasay (Darling, 1955).

The Isle of Skye has a cool oceanic climate, tempered slightly by North Atlantic drift which maintains the relatively mild climate over the year (Birks, 1973). The coast of Skye is deeply indented with no part of the island more than 8km from the sea, underlining the maritime influence on the Isle (see Figure 2.11). Both high winds and rainfall dominate the

![Figure 2.11. Climate of Tirree, average 1971-2000. Source: Met Office data.](image-url)
Hebridean climate, with the steep hills of parts of Skye impacting local conditions, increasing precipitation and lowering temperatures (Armit, 1996; Birks, 1973). Where the high precipitation and humidity coincide with poor drainage peat rapidly forms, as seen in the southern part of the island.

2.11.1 ENVIRONMENTAL HISTORY

By the late Devensian and early Flandrian, initial hazel and birch scrub was succeeded by dense birch-pine and hazel woodlands, enriching the remaining soils, allowing the colonisation of herbaceous species. Birks (1973) comprehensive pollen analysis indicates that, within this colonising herbaceous flora, *Trifolium* pollen can be identified from the peat strata, although at a very low level.

The first permanent inhabitants to extensive areas of Scotland are likely to have arrived in the Mesolithic (7000 BC), with some of the best evidence of their habitation found on sites in the Inner Hebrides, including An Corran on Skye (Armit, 1996). Evidence for the first farming communities on Skye remain elusive, however the close proximity to the mainland suggests that Neolithic farmers arrived at a similar time to those of western Scotland after 4000 BC. Around this time deforestation started on the Inner Hebrides contributing to peat spread and soil impoverishment, and the treeless nature is still prevalent on Skye, with trees surviving only in small sheltered bays, along with a few recent plantations (Birks, 1973).

Following these first colonisations, the Hebrides came under various influences, for example becoming part of the Pictish state, before coming under Viking rule in 800 AD, however in terms of the environment the people of the Hebrides were likely to have continued with a farming lifestyle, cultivating small areas with fishing, cattle rearing and fowling contributing significantly to the economy (Armit, 1996). Skye, in comparison to the other
islands in this study is in close proximity to the mainland, and as such is likely to have high levels of population movement, likely to bring with it novel farming practices and resources.

Evidence indicates that there were a relatively large number of pre-eighteenth century pre-crofting structures on Skye, reflecting the increase in population at this time (Armit, 1996). By the end of the 18th century the population reached a high of over 230000, with the economy focused on crofters, fishing and kelp-burning (Birks, 1973). After 1840 sheep farming was introduced, accelerating tree loss, with many crofters forced to leave their land as part of the highland clearances (Birks, 1973), however Skye lays claim to the Battle of the Braes, a rebellion against the landlords, culminating in Parliament passing laws to give crofters more security.

Recent history

The population of Skye has continued to decrease due in part to the impoverished agricultural land, resulting from deforestation and over-grazing leading to the spread of bracken in many areas reduces grazing potential (Barclay & Darling, 1955). Now, most land on Skye is upland hill pasture grazed mainly by sheep, with higher densities found on the fertile Trotternish Peninsula (Birks, 1973). The main focus of the economy now based on tourism and pastoral agriculture, with crofting settlements still the main type of townships in the area (Stanton, 1996). The increasing role of tourism on Skye has started to increase population numbers on the Isle and will likely impact on the environment, already seen in the increase of traffic and widening of the small numbers of roads, although at present development remains small-scale (Stanton, 1996).
2.11.2 Conservation

Conservation designations

Skye contains 27 Sites of Special Scientific Interest (SSSI) and 10 Special Areas of Conservation (SAC), however of these only one SAC, Totternish Ridge, specifically focuses on the conservation of grassland, covering 31.70 km$^2$.

Current threats to diversity

Grazing is one on the primary concerns for species diversity on Skye with over-grazing small areas leading to loss of species diversity, and a reduction in sheep numbers contributing to the spread of heath areas. Overgrazing, along with muirbum (heather burning) and afforestation contribute to erosion which is seen in areas around crofting areas and on moderate to steep slopes. In addition, the recent increase in tourism will likely impact on the environment of Skye (Stanton, 1996).

2.11.3 Presence of species

Traditionally the agriculture of Skye followed a seven year rotational system, with the last three years dedicated to hay conservation or grazing, however since the early 1970s there has been a move from this system to a greater grassland production for grazing (Grant & MacLeod, 1983). A comprehensive assessment of the flora of Skye places $T. repens$ as a major constituent of 11 of the 65 recognised communities, with $T. pratense$ listed as a secondary constituent of many communities (Birks, 1973). All three taxa are currently listed as present in the Inner Hebrides, with $T. dubium$ less frequent, however correspondence with the Biological Society of the British Isles (BSBI) recorder defines its presence on areas within Skye and particularly on the south west coast of Raasay, nearest to Skye (Bungard, S. 2007, pers. comm.1 May, 2008).
2.12 TARGET ISLAND SITE - ST KILDA, OUTER HEBRIDES

The St Kilda archipelago consists of four main islands, lying 160km north-west of the Scottish mainland, 64km to the west of the Outer Hebrides. In addition to Hirta, Boreray, Soay and Dun there are numerous sea stacs, of which Stac an Armin and Stac Lee (191 m and 165 m high respectively) are the highest in the British Isles (Buchanan, 1995). The largest, and perhaps only until recently permanently habitable of the islands, Hirta, covers 6.28 km² and contains the remains of the village evacuated in 1930. The village is situated in a horseshoe bay surrounded by a ring of five hills, which acts as an entrance to the grassy slopes and cliffs that encompass much of the remainder of the island.

To the south of Hirta lies Dun (0.32 km²), consisting of a narrow ridge, protecting the village from the prevailing south-westerly winds. To the north west of Hirta lies Soay, with a central plateau surrounded by steep cliffs (0.97 km²). Boreray (0.87 km²) lies apart from the islands to the

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latitude</td>
<td>57° 49’ N</td>
</tr>
<tr>
<td>Longitude</td>
<td>08° 33’ W</td>
</tr>
<tr>
<td>Distance from the mainland</td>
<td>160 km</td>
</tr>
<tr>
<td>Total area</td>
<td>8.44 km²</td>
</tr>
</tbody>
</table>
north east of Hirta, and whose steep rocky nature makes it home to one of the largest colonies of North Atlantic gannets.

The islands are what remain of an extinct volcano ring, with the centre of the volcano, lying between Boreray and Hirta, having collapsed inwards (Coates, 1990). The jagged cliffs of Dun and Boreray display the gabbro rocks produced during early volcanic activity, with the smoother island contours of Hirta defined by the granophyre rock produced from later phases of volcanic activity around 55 million years ago (Harding et al., 1984; Buchanan, 1995).

The oceanic climate, modified by the steep cliffs of the islands, creates high rainfall (1400 mm pa) (Small, 1979), and a high frequency of gales (Clutton-Brock et al., 2004). Winter temperatures are higher than that expected for the latitude due to the contribution of North Atlantic drift, with winters alternatively warm and wet to drier and colder. Fluctuations in the climate are strongly linked to the fluctuations over the North Atlantic, which affect much of this region, with Post and Stenseth (1999) and Clutton-Brock et al. (2004) finding the North Atlantic oscillations responsible for earlier and longer flowering times and grass availability.

Although St Kilda has a mainly oceanic climate, the driest period falls between April and June, coinciding with anti-cyclones approaching from the Atlantic (St Kilda Management Plan, 2003). In addition, to the generally high humidity of the archipelago, sea spray is a major contributing factor to the flora, with halophytic Plantago swards to be found hundreds of meters above sea level.

The flora itself is similar to that found on the Outer Hebrides, with no evidence of geographical subspeciation as there is scarce difference in the climates between the two archipelagos (Poore & Robertson, 1948). The moorland vegetation covering much of the upland areas bears a high resemblance to that of the Hebrides, with some notable species missing, illustrating the niche expansion prevalent in the St Kildan flora. The other two main
vegetation classes on Hirta are maritime grassland found around the edges of the island, and the now *Holcus-Agrostis* grassland found on the formerly cultivated area around village bay, both displaying a distinct lack of species in comparison to Hebridean and mainland grasslands.

In terms of grazing, concentrated before the evacuation around the village and Gleann Mor to the north, the numbers of sheep have been steadily increasing from the original introduction of 107 sheep from Soay after the evacuation (Poore & Robertson, 1948).

The most highly productive grazing area of *Holcus-Agrostis* dominated vegetation that occupies the area above the head dyke on the formerly cultivated land is the most frequented by the Soay sheep that now freely roam Hirta (Jones *et al*., 2006). This area of previously cultivated land, around 0.15 to 0.3 km\(^2\) was intensively manured to improve the formerly poor soils of the area with large amounts of peat ash and both human and seabird waste. Recent assessments of the soils of Hirta, in light of the unique manuring practices of these isles, have discovered high levels of Pb and Zn contaminants (Meharg *et al*., 2006).

The oceanic nature of the archipelago, situated to withstand the full force of the Atlantic, produces a site of high natural and historical significance, and in terms of genetic diversity is of great interest, being the most isolated archipelago in the United Kingdom.

2.12.1 ENVIRONMENTAL HISTORY

Dating of the first habitation of the islands tends to centre on the European Bronze age (Campbell, 1974), however some authors suggest that, as the archipelago can be viewed from the Outer Hebrides on clear days, foraging trips may have been made to islands earlier, during the Neolithic era (Harman, 1995).

The Outer Hebrides, the origin of the first settlers, was not an isolated culture with many influences, including agricultural, from both the mainland and Orkney and Shetland
Islands (Fleming, 2005). The first settlers would have encountered a *graminae-plantago* dominated sward, much as it is seen today (Walker, 1984). Anthropomorphic effects on the vegetation of St Kilda appear to have been secondary to the impact of regional climatic disturbances, which resulted in the temporal variation seen in the composition of St Kilda’s grassy swards. The decrease in the already small proportion of woody pollen in the pollen assemblage could be attributed to small scale clearances, although this remains speculative (Walker, 1984). In terms of sheep grazing, the removal of the improved breeds in 1934 contributed to the spread of *Calluna* in the upland areas, and decrease in *Nardus* grassland (Gwynne *et al*., 1974) indicating the impact grazing has on the islands recent vegetational history.

Acid, peaty soil predominate on the islands (Gwynne *et al*., 1974) with blanket peats found on the flat areas on Hirta, Mullach Mor and on the west side of Gleann Mor. The soils of Hirta were fertile, although in the 0.32 km² under cultivation in 1758 in the immediate vicinity of the village the fertility was aided by the addition of bird remains, urine and ashes to the compost (Macaulay, 1764), with hill pastures common and used for sheep and cattle grazing (Seton, 1878).

The isolation of St Kilda was a main factor in its anthropomorphic history, with one of the earliest accounts of the peoples of St Kilda, made in 1697 when the only means of transport to island was an open long boat during calm weather (Martin, 1981). There are many accounts of the islanders exporting produce during the late 18th and 19th century, in particular from the large seabird colonies, as well as producing hand tilled barley and oats. There are records of yearly visits from the minister from the Hebrides visiting the islands, as well as the factor, to collect rent, in addition to tourists from 1834. Hence, trips to and from the mainland and Hebrides, and conceivably material transfer, will have been necessary, but most likely infrequent.
Cattle were kept on Hirta, along with imported black faced sheep on Hirta and Boreray, mainly providing produce for the islanders’ use, although the population remained small, with the earliest record of population size numbering just 180 (Martin, 1981). Many of the earliest accounts of the islands refer to the large number of sheep on the islands, indicating the high intensity of grazing pressure on Hirta (Buchanan, 1995; Fleming, 2005).

Recent history

An 18th century outbreak of smallpox reduced the population to 30, which then climbed to 100 following repopulation from Skye and Harris at which it remained until the late 18th century (Fleming, 1999; Clutton-Brock et al., 2004). Following a large emigration to Australia, the population continued to lessen, culminating in a requested evacuation of the population and domesticated animals when numbers were too few to till the fields and maintain the islands boat.

The islands were bought after the evacuation by Lord Dumfries in 1931, later the fifth Marquess of Bute, which he designated as a bird reserve and permitted visits from limited numbers of tourists and naturalists (Harman, 1995). In his will he bequeathed the islands to the National Trust for Scotland, keen to preserve the history and the fauna of the islands.

2.12.2 Conservation

Conservation designations

In 1957 the ownership was passed to the National Trust for Scotland. At this time, just after the Second World War, the archipelago also became home to a radar station, so human occupation of Hirta, though small, returned to the archipelago.

St Kilda was designated a National Nature Reserve (NNR) in 1964, and a world heritage site in 1986, under the management of the Nature Conservancy Council, leased by
the National Trust for Scotland. In addition, St Kilda was designated as a Site of Special Scientific Interest (SSSI) in 1984, and has been designated a Special Area of Conservation (SAC) due in part to its severe maritime environment and hence the extreme forms of Atlantic maritime vegetation.

Currently St Kilda comes under shared management between the National Trust for Scotland, Scottish Natural Heritage (previously the Nature Conservancy Council), and the Ministry of Defence (previously the Air Ministry).

Current threats to diversity

In terms of diversity St Kilda provides an interesting dichotomy, suffering from „niche expansion“, where a smaller number of species occupy a wider niche space, whilst simultaneously acting as a potential refuge from introgression due to its extreme isolation from the mainland.

Due to the protected area designations of the archipelago the entire area falls under a series of management plans, limiting the threats to the island, with, in terms of flora diversity, grazing regimes and natural processes continuing without intervention, strict limits to the number of visitors and prescriptions in place to avoid the introduction of alien species and genotypes (St Kilda Management Plan, 2003). Most recently however, the Ministry of Defence (MoD) which mans a radar station on Hirta has announced plans to remotely manage the station, a plan that the National Trust say will be to the detriment of St Kilda (BBC, 2009). Plans are in their early stages, so it remains to be seen whether the removal of the MoD presence will have a negative impact on the management of St Kilda.

2.12.3 Presence of target species

One of the earliest mentions of *Trifolium* comes from Macaulay (1764), describing a valley on the north west of the Island, where “one may see intermixed with the more common
kinds of grass, a great and beautiful variety of the richest plants, clover white and red”. Interestingly, in relation to pollination of the target species, Martin (1981) asserts that “…nor is ever a bee seen at any time”.

In the most recent assessment of the flora of St Kilda (Crawley, 2004), the only recorded *Trifolium* present on St Kilda are *T. repens* and *T. pratense*. Leguminosae in general is limited with the only other species noted to have occurred on St Kilda as *Vicia sepium* and *V. hirsuta*, both of which may now only exist in the seed bank (Crawley, 2004). *T. repens* is recorded as abundant in fertile grasslands within the Head Dyke, as frequent in the village, the army base and in the vicinity of cleit doors, as well as in *Plantago* swards and short seaside turf, and finally as occasional in heaths and drier heathy grasslands. *T. pratense* however, is noted as rare and found only in the fertile grasslands within the Head Dyke (within village bay).

### 2.13 CURRENT THREAT STATUS

The level of threat to a species is most often linked to the risk of extinction with rarity, decline and habitat fragmentation often used as indicators of the level of threat (Hartley & Kunin, 2003). In terms of rarity, in particular in relation to the International Union for the Conservation of Nature (IUCN) Red List assessment, rarity is most often used to describe the geographic extent of a species and/or the number of individuals within that range (Hartley & Kunin, 2003). However, it should be noted here that it does not always follow that a rare species is threatened, as some species are able to persist in low numbers (Mace & Kershaw, 1997). The IUCN (2001) Red List (see Table 2.1) is an attempt to provide an international, transparent and rigorous set of criteria to assess the conservation status of species.
Table 2.1 IUCN categories system, version 3.1 (IUCN, 2001).

<table>
<thead>
<tr>
<th>Classification</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extinct (EX)</td>
<td>A taxon is Extinct when there is no reasonable doubt that the last individual has died.</td>
</tr>
<tr>
<td>Extinct in the wild (EW)</td>
<td>A taxon is Extinct in the Wild when it is known only to survive in cultivation, in captivity or as a naturalized population (or populations) well outside the past range.</td>
</tr>
<tr>
<td>Critically endangered (CR)</td>
<td>A taxon is Critically Endangered when the best available evidence indicates that it meets any of the criteria A to E for Critically Endangered and it is therefore considered to be facing an extremely high risk of extinction in the wild.</td>
</tr>
<tr>
<td>Endangered (EN)</td>
<td>A taxon is Endangered when the best available evidence indicates that it meets any of the criteria A to E for Endangered and it is therefore considered to be facing a very high risk of extinction in the wild.</td>
</tr>
<tr>
<td>Vulnerable (VU)</td>
<td>A taxon is Vulnerable when the best available evidence indicates that it meets any of the criteria A to E for Vulnerable and it is therefore considered to be facing a high risk of extinction in the wild.</td>
</tr>
<tr>
<td>Near threatened (NT)</td>
<td>A taxon is Near Threatened when it has been evaluated against the criteria but does not qualify for Critically Endangered, Endangered or Vulnerable now, but is close to qualifying for or is likely to qualify for a threatened category in the near future.</td>
</tr>
<tr>
<td>Least concern (LC)</td>
<td>A taxon is Least Concern when it has been evaluated against the criteria and does not qualify for Critically Endangered, Endangered, Vulnerable or Near Threatened. Widespread and abundant taxa are included in this category.</td>
</tr>
<tr>
<td>Data deficient (DD)</td>
<td>A taxon is Data Deficient when there is inadequate information to make a direct, or indirect, assessment of its risk of extinction based on its distribution and/or population status.</td>
</tr>
<tr>
<td>Not evaluated (NE)</td>
<td>A taxon is Not Evaluated when it is has not yet been evaluated against the criteria.</td>
</tr>
</tbody>
</table>

The quantification of range sizes and numbers of individuals form a large part of the Red List assessment, measuring the extent of occurrence and area of occupancy, in addition to basing classifications on available information on population size, and levels of habitat fragmentation. While the IUCN (2008) promote the use of information that quantifies population viability, the lack of this type information means that the majority of assessments will not use this criterion. This, in addition to defining rarity by either extent of occurrence or
area of occupancy, provides some obvious questions about the methodology of the red list assessments, with rarity relating to habitat specificity, ephemerality and genetic diversity not incorporated into the criteria for assessment.

Table 2.2 Summary of the five criteria (A-E) used to evaluate if a taxon belongs in a threatened category (Critically Endangered, Endangered or Vulnerable). Taken from IUCN (2008).

<table>
<thead>
<tr>
<th>Use any of the criteria A-E</th>
<th>Critically Endangered</th>
<th>Endangered</th>
<th>Vulnerable</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. Population reduction</strong></td>
<td>Declines measured over the longer of 10 years or 3 generations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>&gt; 90%</td>
<td>&gt; 70%</td>
<td>&gt; 50%</td>
</tr>
<tr>
<td>A2, A3 &amp; A4</td>
<td>&gt; 80%</td>
<td>&gt; 50%</td>
<td>&gt; 30%</td>
</tr>
<tr>
<td><strong>A1.</strong> Population reduction observed, estimated, inferred, or suspected in the past where the causes of the reduction are clearly reversible AND understood AND ceased based on and specifying any of the following:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) direct observation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b) an index of abundance appropriate to the taxon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(c) a decline in area of occupancy (AOO), extent of occurrence (EOO) and/or habitat quality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(d) actual or potential levels of exploitation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(e) effects of introduced taxa, hybridisation, pathogens, pollutants, competitors or parasites.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A2.</strong> Population reduction observed, estimated, inferred, or suspected in the past where the causes of reduction may not have ceased OR may not be understood OR may not be reversible, based on any of (a) to (e) under A1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A3.</strong> Population reduction projected or suspected to be met in the future (up to a maximum of 100 years) based on any of (b) to (e) under A1.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A4.</strong> An observed, estimated, inferred, projected or suspected population reduction (up to a maximum of 100 years) where the time period must include both the past and the future, and where the causes of reduction may not have ceased OR may not be understood OR may not be reversible, based on any of (a) to (e) under A1.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>B. Geographic range in the form of either B1 (extent of occurrence) OR B2 (area of occupancy)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1. Either extent of occurrence</td>
<td>&lt; 100 km²</td>
<td>&lt; 5,000 km²</td>
<td>&lt; 20,000 km²</td>
</tr>
<tr>
<td>B2. or area of occupancy</td>
<td>&lt; 10 km²</td>
<td>&lt; 500 km²</td>
<td>&lt; 2,000 km²</td>
</tr>
<tr>
<td>and 2 of the following 3:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) severely fragmented or # locations</td>
<td>= 1</td>
<td>≤ 5</td>
<td>≤ 10</td>
</tr>
<tr>
<td>(b) continuing decline in (i) extent of occurrence (ii) area of occupancy, (iii) area, extent and/or quality of habitat,(iv) number of locations or subpopulations and (v) number of mature individuals</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(c) extreme fluctuations in any of (i) extent of occurrence, (ii) area of occupancy, (iii) number of locations or subpopulations and (iv) number of mature individuals.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>C. Small population size and decline</strong></td>
<td>&lt; 250</td>
<td>&lt; 2,500</td>
<td>&lt; 10,000</td>
</tr>
<tr>
<td>Number of mature individuals and either C1 or C2:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1. An estimated continuing decline of at least up to a maximum of 100 years</td>
<td>25% in 3 years or 1 generation</td>
<td>20% in 5 years or 2 generations</td>
<td>10% in 10 years or 3 generations</td>
</tr>
<tr>
<td>C2. A continuing decline and (a) and/or (b)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a i) # mature individuals in largest subpopulation</td>
<td>&lt; 50</td>
<td>&lt; 250</td>
<td>&lt; 1,000</td>
</tr>
<tr>
<td>(a ii) or % mature individuals in one subpopulation = 90-100%</td>
<td>95-100%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>(b) extreme fluctuations in the number of mature individuals</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>D. Very small or restricted population</strong></td>
<td>&lt; 50</td>
<td>&lt; 250</td>
<td>&lt; 1,000</td>
</tr>
<tr>
<td>Either (1) number of mature individuals</td>
<td>na</td>
<td>na</td>
<td>typically: AOO &lt; 20km² or # locations ≤5</td>
</tr>
<tr>
<td>or (2) restricted area of occupancy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>E. Quantitative Analysis</strong></td>
<td>50% in 10 years or 3 generations (100 years max)</td>
<td>20% in 20 years or 5 generations (100 years max)</td>
<td>10% in 100 years</td>
</tr>
</tbody>
</table>
For our purposes here however the Red List assessment for these target taxa, based on available information in a regional context, provides an interesting base line for conservation planning, and will serve as a comparison for conservation strategies based on genetic diversity assessments.

Within the UK there is evidence of habitat fragmentation of grasslands (Fuller, 1987) and as such the potential for population decline (Criterion A) however there is no quantitative information on decline on these three species. Indeed for *T. repens* and *T. pratense* it may be the case that natural wild populations are being replaced by semi-natural populations derived from cultivation, but this is both difficult to assess without a frame of reference and in any case would not be assessed under the IUCN criteria. For *T. dubium*, there does appear to be anecdotal evidence that the populations in the extreme north are in decline, however it is also suggested that these populations fall outside of its natural range and represent introductions (Bungard, S. 2007, pers. comm. 1 May, 2008).

Using the criteria defined in Table 2.2 and categorising under a regional framework, all three species cover the whole of the United Kingdom totalling 242,514 km$^2$ (Office of National Statistics, 2004), with the area of occupancy only just under that of their extent of occurrence (Figure 2.1). Therefore the three target taxa, *T. dubium*, *T. pratense* and *T. repens*, as widespread and abundant species can be categorised as Least Concern under the Red List categories for a UK wide regional assessment (see also Dines et al., 2005).

### 2.14 Conservation status on the basis of ecogeographic information

The widespread occurrence and abundant nature of the three target taxa ensures that many populations will exist within protected areas within the UK (Hawker & Hawker, 2005). In addition, this high area of occupancy ensures that the three taxa are categorised as Least Concern.
under the Red List classification. Basing a conservation strategy purely on the ecogeographic information above, it could be suggested that the abundance of the three species across a wide range of habitat types and therefore their increased ability to cope with threats to their habitats signifies that specific conservation measures are little required. However, does abundance necessarily mean that these species require no conservation action compared to more rare species? While the taxa are present in protected areas, this does not necessarily equate to taxon specific management activities designed for their conservation, as widespread and abundant species they are often not priority species and thus not specifically managed. Indeed management activities within protected areas designed to maintain other species can have negative effects on Trifolium, for example re-seeding or over and undergrazing. The nature of the three target species as crop wild relatives highlights their value and thus their need for conservation. The widespread use of cultivated T. repens and T. pratense indicates the potential for genetic erosion or swamping of the wild species genomes, giving species such as these, grown alongside a conspecific crop, high priority for conservation, even if it is little acknowledged in the classic conservation literature. Thus while these species are little threatened at the species level of diversity in the UK, they may require different conservation strategies on the basis of genetic diversity, a question that is addressed in the following chapters.
Chapter 3. **GENETIC STRUCTURE OF THREE *TRIFOLIUM* SPECIES IN THE UK USING AFLP**

### 3.1 INTRODUCTION

The conservation of natural resources is often limited by the lack of technical, scientific and financial resources. Consequently researchers are forced to prioritise biodiversity for conservation based on a scientifically sound assessment of the "value" of each component of biodiversity, whether at the ecosystem level, the species or the genetic level of diversity.

Within a species these limited conservation resources dictate which populations should be prioritised for conservation. When the aim is to conserve the maximum diversity of a widespread species, genetic diversity is typically a major determinant of the value of populations for conservation. Spatial structuring of diversity has been found to occur in natural populations of plant species, through both the non-random mating of genotypes and the occupation of heterogeneous environments (Heywood, 1991). This spatial genetic structure can produce high levels of variation within populations, with resultant conservation priority lying in those populations that contain the highest levels of diversity.

Measures of diversity alone are not adequate when prioritising populations. Selecting more than one population for conservation will require additional assessments of the distinctiveness of the populations to allow for redundancy in the dataset (Petit *et al*., 1998). A further dimension to population selection also lies in the ability to retain diversity in the population through isolation and the provision of management. Priority populations can thus be defined as a function of their diversity, their divergence and their isolation from factors that threaten their biodiversity (Petit *et al*., 1998; Gaston *et al*., 2002; Margules *et al*., 2002).
Island populations could be identified in advance as priority areas for conservation through their potential for isolation, and as such, potential for unique diversity compared to mainland populations. The process of divergence, not always at the level of speciation but at the population level, can lead to island populations being both phenotypically and genotypically distinct from their corresponding mainland populations (Whittaker & Fernández-Palacios, 2007). Island isolation can also create a refuge for genetic diversity, shielding insular populations against the possible gene flow and genetic swamping from introduced or cultivated varieties that may be prevalent in their mainland counterparts.

Islands provide an interesting dichotomy in terms of conservation, both forming a refuge whilst simultaneously restricting the overall size of populations. Extinctions occur more readily on islands as a consequence of small overall area and associated population size restriction, i.e. through a greater reaction to stochastic processes, genetic drift and lower levels of immigration (Frankham 1995, 1997; Whittaker & Fernández-Palacios, 2007). Hence, the conservation importance of island biotas lies both in their potential uniqueness, and their susceptibility for diversity loss.

In this chapter, the spatial genetic structure and the factors that affect it are assessed using three species of Trifolium; T. dubium Sibth., T. pratense L. and T. repens L. in order to indicate priority areas for conservation in the UK including the surrounding islands. T. dubium is an annual, inbreeding species native to Europe and western Asia, which has spread to many parts of the world (Zohary & Heller, 1984; Frame, 2005). Although characterized in terms of morphological and flowering variation (Caradus & Mackay, 1989), to the author’s knowledge this is the first report of a population genetic study in T. dubium. T. pratense is a perennial self-incompatible species widely distributed across temperate zones of the world, and cultivated throughout the northern hemisphere (Williams & Williams, 1947; Zohary & Heller, 1984). T. pratense is characterized by large genetic variation within and between
populations, with large numbers of locally adapted genotypes permitting its persistence in many diverse parts of the world (Joshi et al., 2001; Dias et al., 2008). *T. repens* L. was chosen as a model native species which has historically been cultivated across most of the UK, with seed traded since the 17th century across Europe (Caradus, 1995). In the UK, *T. repens* a long-lived mixed-mating perennial, is found in a diverse range of grassy habitats, in either sown mixtures in pastures or as wild and semi-natural populations in uncultivated areas (Zohary & Heller, 1984; Frame, 2005). This adaptation of *T. repens* to a broad range of habitats highlights the high inherent genetic diversity of the species (Gustine & Huff, 1999; Kölliker et al., 2001). All three species are ubiquitous within the UK and as such are often given little conservation importance; however this lack of conservation priority may be mistaken. Differentiation through isolation, across all types of geographic barriers can occur in widespread species, creating unique pockets of genetic diversity of importance for both future conservation and breeding programs.

This study aims to: 1) evaluate the spatial pattern of diversity in *T. dubium*, *T. pratense* and *T. repens* across the UK; 2) compare the patterns of diversity between the three related species; 3) assess the importance of island groups of the UK for genetic isolation and conservation; and 4) indicate priority areas for conservation. For *T. repens* this study also aims to answer the following question: Can wild populations of *T. repens* remain genetically distinct from cultivated varieties of the same species, and do some UK islands offer refugia from crop gene flow?
3.2 METHODS

3.2.1 POPULATION SAMPLING

The most effective sampling strategy would be designed on the basis of prior knowledge of the genetic variation within a species and within populations (Namkoong, 1988; Gapare et al., 2008). In the absence of such prior information, as is the case in many wild species, theoretical models have been developed to define effective sampling strategies (e.g. Kimura & Crow, 1964; Brown & Marshall, 1995; Lawrence et al., 1995a). The most effective of these models has been the subject of considerable debate, particularly in terms of the number of individuals to collect per site (notably Allard, 1970; Marshall & Brown, 1975, 1983; Yonezawa, 1985; van Reehan et al., 1993; Lawrence et al., 1995a, b). Controversy arises from some authors advocating the increase in the number of individuals to include the collection of, or more correctly to increase probability of, collecting rare alleles that are likely to be missed when using a more pragmatic approach (Krusche & Geburek, 1991; van Reehan et al., 1993; Lawrence et al., 1995a). However, any plant collecting strategy must be ultimately constrained by limiting resources; with Brown and Marshall (1995) noting that in terms of the resources required by a collector, the increase in alleles per individual plant collected is directly proportional to sample size. An efficient and cost effective strategy is clearly vital due both to the need to design conservation strategies before the loss of diversity becomes too great, as well as due to the cost of maintaining and assessing large samples of redundant germplasm (Richards et al., 2007).

In the present study the aim of the collection was to collect a representative sample from each selected area to assess the genetic diversity in Trifolium species across the UK. The collection would then act as a baseline study for future collections of genetic material of these species for use in breeding programs. Marshall and Brown (1999) define a representative
sample as the number of plants required to collect, with a 95% certainty, a copy of all the common alleles in that population with a frequency of over 0.05. This number varies according to breeding system with those from a fully outbreeding species, such as *T. pratense* and *T. repens*, numbering around 30 random plants per population. In a self-fertilising species, such as *T. dubium*, the suggested optimum sample size is increased to 59 random individuals (Brown & Marshall, 1995). While this is a more pragmatic approach compared to the larger numbers quoted by other authors (Krusche & Geburek, 1991; Lawrence 1995a), increasing the certainty level to above 95% or ensuring the collection of extremely rare alleles would necessitate large increases in sample sizes, with diminishing returns per sample. As the aim of this collection is to collect a representative sample with limited resources, large sample sizes were not feasible (see Table 3.1 for sample sizes per site). In the absence of the presence of all three species at a site, potentially due to their varying niche requirements, collections were made in as close proximity as possible.

Random sampling of genotypes is generally the most desirable method of sampling particularly in species which show little or no sub-population structure. Wild species however often show some sub-structure, with the available literature for *T. repens* and *T. pratense* indicating high genetic diversity within populations (Gustine & Huff, 1999; Mosjidis *et al.*, 2004; van Treuren *et al.*, 2005; Dias *et al.*, 2008). Hence, where a site showed clear habitat heterogeneity, a stratified random sampling method was used, sampling random individuals from selected micro-sites. Due to the small leaf size of *T. dubium*, entire plants were collected and placed in labelled zip lock bags with indicator silica gel. As *T. dubium* is largely a self-fertilising species, individuals were collected a minimum distance apart (1.5m or more, dependent on total population area). All *T. pratense* individuals were collected as leaf samples and placed immediately in labelled zip lock bags with indicator silica gel. Where possible
individuals were chosen at random to avoid any bias associated with intermittent variation in a population (Brown & Briggs, 1991; Brown & Marshall, 1995).

Two *T. repens* reference cultivars, English Dutch and Kent Wild White, were included in this study, chosen in consultation with clover specialists and taken from the gene bank at IGER, considering both history of the landraces and availability of seed. These represent historical landraces grown in the UK and can be considered forerunners of commercial cultivars used in Britain, with Kent Wild White still used commercially in the UK. The two cultivars were grown from seed and randomly chosen individuals were sampled from each of the two grown reference landraces. In wild populations individuals were collected >5m apart due to the clonal nature of *T. repens*. Patch size is known to range from <1 – 5m², although in general 1.5 to 6 clones can be found per 10m² (Harberd, 1963; Cahn & Harper, 1976). Sackville Hamilton and Chorlton (1995) describe populations that have consisted of just one or two clones, but these are postulated to occur at the limits of the species, e.g. in shrubland and tall grassland.

Populations collected from Benbecula, St Kilda and Skye were collected as vegetative samples, bagged separately per site. The plants were potted in separate containers on return and kept at 16 hours daylight and 20/10°C. Each sample from populations in South Uist, Sussex, Devon and Dorset were collected and placed immediately in labelled zip lock plastic bags with indicator silica gel. Replicates were taken from the same stolon to ensure the same genotype was sampled.

Collection sites were selected to represent wild, and in the case of *T. repens* semi-natural, populations across the UK, encompassing both islands and mainland reference populations (see Table 3.1 and Figure 3.1, Figure 3.2 and Figure 3.3 for provenance information). As most genetic differentiation is related to geographic heterogeneity, Marshall and Brown (1999) advocate that when overall sample number is restricted the number of sites
should be increased over number of individuals per site. As such, three populations of 15 or more individuals were selected in each area, dependent on species availability.
Table 3.1 Collection site data for all populations. TD corresponds to *T. dubium*, TP to *T. pratense* and TR to *T. repens*. All collections made by Serene Hargreaves otherwise by 1S Hargreaves and David Jacoby; 2Nigel Maxted; 3Maria Scholten; 4S Hargreaves, Ian Thomas and Sue Dalton; 5I Thomas and S Dalton; 6Ryoko Hirano, I Thomas and S Dalton.

<table>
<thead>
<tr>
<th>Population</th>
<th>Code</th>
<th>Date</th>
<th>Lat.</th>
<th>Long.</th>
<th>Site description</th>
<th>No. collected</th>
<th>TD</th>
<th>TR</th>
<th>TP</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Dutch</td>
<td>ED1</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kent Wild White</td>
<td>KWW1</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Wild/semi natural populations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>St Martins, Isles of Scilly(^1)</td>
<td>IOS1</td>
<td>23/06/07</td>
<td>49.96669</td>
<td>-6.30043</td>
<td>Coastal grassland path around Tinkers point.</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>St Mary's, Isles of Scilly(^2)</td>
<td>IOS2</td>
<td>23/06/07</td>
<td>49.91805</td>
<td>-6.30167</td>
<td>Small walled grazed fields on Carn Morval point.</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Bryher, Isles of Scilly(^1)</td>
<td>IOS3</td>
<td>24/06/07</td>
<td>49.95044</td>
<td>-6.3517</td>
<td>Path between pool and Stinking Porth.</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Bryher, Isles of Scilly(^1)</td>
<td>IOS4</td>
<td>24/06/07</td>
<td>49.9504</td>
<td>-6.35088</td>
<td>Central Bryher, next to recently laid paths to Bay Hotel.</td>
<td>15</td>
<td></td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>St Mary's, Isles of Scilly(^1)</td>
<td>IOS5</td>
<td>26/06/07</td>
<td>49.90134</td>
<td>-6.30081</td>
<td>Path to Penennis head from Hugh Town.</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Branscombe, Devon(^2)</td>
<td>DEV1</td>
<td>11/01/07</td>
<td>50.69389</td>
<td>-3.15944</td>
<td>Around ramparts of hill fort between cliff and valley.</td>
<td>12</td>
<td></td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Berry Head NNR, Devon</td>
<td>DEV2</td>
<td>17/08/07</td>
<td>50.40406</td>
<td>-3.48445</td>
<td>Clifftop paths on Berry Head.</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Branscombe, Devon</td>
<td>DEV3</td>
<td>17/08/07</td>
<td>50.68584</td>
<td>-3.11899</td>
<td>Gay's farm, upper fields overlooking sea.</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Kingcombe NR, Dorset(^2)</td>
<td>DOR1</td>
<td>13/01/07</td>
<td>50.78794</td>
<td>-2.63407</td>
<td>Wet meadows, permanent pasture.</td>
<td>14</td>
<td></td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Abbotsbury Cast., Dorset(^2)</td>
<td>DOR2</td>
<td>13/01/07</td>
<td>50.67736</td>
<td>-2.62834</td>
<td>Iron age hill fort, around ramparts. Permanent pasture.</td>
<td>22</td>
<td></td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Lamberts Cast., Dorset(^2)</td>
<td>DOR3</td>
<td>14/01/07</td>
<td>50.77713</td>
<td>-2.90908</td>
<td>SSSI, damp meadows, permanent pasture, many springs.</td>
<td>20</td>
<td></td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Rye, East Sussex(^3)</td>
<td>RYE1</td>
<td>02/07/04</td>
<td>50.9667</td>
<td>0.7683</td>
<td>Long established grazing marshes</td>
<td>21</td>
<td></td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Population</td>
<td>Code</td>
<td>Date</td>
<td>Lat.</td>
<td>Long.</td>
<td>Site description</td>
<td>No. collected</td>
<td>TD</td>
<td>TR</td>
<td>TP</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------</td>
<td>-----------</td>
<td>---------</td>
<td>----------</td>
<td>----------------------------------------------------------------------------------</td>
<td>---------------</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>N Walney Island, Cumbria</td>
<td>LKD1</td>
<td>31/07/07</td>
<td>54.11775</td>
<td>-3.25211</td>
<td>Field to the right of the road before Walney airport.</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N Walney Island, Cumbria</td>
<td>LKD2</td>
<td>31/07/07</td>
<td>54.11775</td>
<td>-3.26681</td>
<td>Coastal path bordering Walney airport.</td>
<td>15 15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dobshell Wood, Cumbria</td>
<td>LKD3</td>
<td>31/07/07</td>
<td>54.18539</td>
<td>-2.835</td>
<td>Lower pasture, off road to Arnside Knott.</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arnside Knott, Cumbria</td>
<td>LKD4</td>
<td>31/07/07</td>
<td>54.18547</td>
<td>-2.835</td>
<td>Arnside Knott pasture, at the entrance to the Knott.</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hutton Roof, Cumbria</td>
<td>LKD5</td>
<td>01/08/07</td>
<td>54.18564</td>
<td>-2.65214</td>
<td>Path to Park Wood NNR from road.</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hutton Roof, Cumbria</td>
<td>LKD6</td>
<td>01/08/07</td>
<td>54.18506</td>
<td>-2.65267</td>
<td>Meadow at entrance to Park Wood NNR.</td>
<td>15 15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morvich, NW Scotland</td>
<td>NWS1</td>
<td>24/10/06</td>
<td>57.2338</td>
<td>-5.3696</td>
<td>Entrance to Gleann Lichd, Wet boggy area beside path.</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morvich, NW Scotland</td>
<td>NWS2</td>
<td>24/10/06</td>
<td>57.2342</td>
<td>-5.3842</td>
<td>One of a number of small fields in a broad river valley.</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ardelve, NW Scotland</td>
<td>NWS3</td>
<td>24/10/06</td>
<td>57.2805</td>
<td>-5.5315</td>
<td>Path alongside shingle spit of sea loch.</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nordie, north west Scotland</td>
<td>NWS4</td>
<td>03/08/07</td>
<td>57.26825</td>
<td>-5.56844</td>
<td>On path side leading to houses and jetty.</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elgol, Skye</td>
<td>SKY1</td>
<td>24/10/06</td>
<td>57.1457</td>
<td>-6.1061</td>
<td>Moorland opposite jetty, near stream.</td>
<td>15 32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Torrin, Skye</td>
<td>SKY2</td>
<td>24/10/06</td>
<td>57.2216</td>
<td>-6.0298</td>
<td>Near old sheep pens, between road and loch shore.</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aird of Slet, Skye</td>
<td>SKY3</td>
<td>04/08/07</td>
<td>57.03347</td>
<td>-5.952</td>
<td>Road leading to Aird of Slet from main road, last 100m.</td>
<td>15 15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merkadale, Skye</td>
<td>SKY5</td>
<td>05/08/07</td>
<td>57.28592</td>
<td>-6.3</td>
<td>B8009 leading to Merkadale, at end of Loch Harport.</td>
<td>15 15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benbecula</td>
<td>BEN1</td>
<td>27/06/05</td>
<td>57.60255</td>
<td>-7.52322</td>
<td>Balranald (RSPB reserve). Rough machair, near shore.</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peninerine, South Uist</td>
<td>UIS2</td>
<td>16/01/07</td>
<td>57.2903</td>
<td>-7.4201</td>
<td>Near Peninerine on machair/pasture.</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Howmore, South Uist</td>
<td>UIS4</td>
<td>18/01/07</td>
<td>57.3028</td>
<td>-7.3978</td>
<td>Dunes north of Howmore, river estuary at top of dunes.</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhenigidale, Harris</td>
<td>UIS5</td>
<td>06/08/07</td>
<td>57.91719</td>
<td>-6.66889</td>
<td>Roadside in Rhenigidale village.</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manish, Harris</td>
<td>UIS6</td>
<td>06/08/07</td>
<td>57.78606</td>
<td>-6.86839</td>
<td>Manish, road leading to township from main road.</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td>Code</td>
<td>Date</td>
<td>Lat.</td>
<td>Long.</td>
<td>Site description</td>
<td>No. collected</td>
<td>TD</td>
<td>TR</td>
<td>TP</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-------</td>
<td>--------</td>
<td>-------</td>
<td>--------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>---------------</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Lochskipport, South Uist</td>
<td>UIS7</td>
<td>08/08/07</td>
<td>57.31875</td>
<td>-7.28356</td>
<td>Road to Lochskipport, before turn off to Salmon farm.</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gleann, Barra</td>
<td>UIS8</td>
<td>09/08/07</td>
<td>56.95243</td>
<td>-7.45220</td>
<td>Road end (last 50m) and car park at end of road.</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glean Mor, St.Kilda</td>
<td>STK1</td>
<td>29/06/05</td>
<td>57.81859</td>
<td>-8.59317</td>
<td>Site of old sealings used as shelters by sheep.</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bagh a' Bhaire, St.Kilda</td>
<td>STK2</td>
<td>30/06/05</td>
<td>57.81291</td>
<td>-8.57099</td>
<td>Former cultivated area near abandoned village.</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tobar Childa, St.Kilda</td>
<td>STK3</td>
<td>30/06/05</td>
<td>57.81471</td>
<td>-8.57166</td>
<td>Rocky ground between 'cleits' and walled enclosures.</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ruabhal, St Kilda</td>
<td>STK4</td>
<td>30/06/05</td>
<td>57.80448</td>
<td>-8.57944</td>
<td>Steeply sloping coastal promontory.</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ruabhal, St Kilda</td>
<td>STK5</td>
<td>30/06/05</td>
<td>57.80413</td>
<td>-8.57388</td>
<td>Grazed area on lower slopes of promontory.</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whiteness Penins., Shetland</td>
<td>SHT1</td>
<td>24/08/07</td>
<td>60.12628</td>
<td>-1.26034</td>
<td>East coast of Whiteness peninsula, coastal path.</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whiteness Penins., Shetland</td>
<td>SHT2</td>
<td>24/08/07</td>
<td>60.18585</td>
<td>-1.28511</td>
<td>End of road nearest to A491 - last 50m of roadside.</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nibon, Shetland</td>
<td>SHT3</td>
<td>25/08/07</td>
<td>60.43591</td>
<td>-1.40248</td>
<td>Road to Nibon, before reaching Trolladale water.</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brae, Shetland</td>
<td>SHT4</td>
<td>25/08/07</td>
<td>60.43371</td>
<td>-1.38543</td>
<td>Road between Brae and Nibon.</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bay of Quendale, Shetland</td>
<td>SHT5</td>
<td>25/08/07</td>
<td>59.88031</td>
<td>-1.31689</td>
<td>Dune flats to the east of Quendale Bay</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bay of Quendale, Shetland</td>
<td>SHT6</td>
<td>25/08/07</td>
<td>59.88566</td>
<td>-1.30271</td>
<td>Field next to public path to the east of Quendale Bay</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3.1 Locations of the 11 *T. dubium* collection sites across the UK: a. Collections from the Inner and Outer Hebrides, b. Isles of Scilly collection sites.
Figure 3.2 Locations of the 16 *T. pratense* collection sites across the UK: a. Shetland collection sites, b. North West Scotland collections and associated islands, c. Isles of Scilly collection sites.
Figure 3.3 Locations of the 27 *T. repens* collection sites across the UK: 
a. North West Scotland collections with associated islands, 
b. Shetland collection sites, 
c. St Kilda collection sites, 
d. Isles of Scilly collection sites.
3.2.2 Amplified Fragment Length Polymorphism

DNA was extracted from leaf material of *T. repens* (50 mg fresh weight, 10 mg dry weight) using the DNeasy 96 Plant Kit (Qiagen). The standard Qiagen protocol was slightly modified by increasing the incubation period at -20°C from 10 to 20 min to increase the DNA quality. Leaf material from *T. dubium* and *T. pratense* was extracted using a CTAB extraction protocol modified from Gawel and Jarret (1991), see Appendices 1 and 2 for full protocol and stock solution information. DNA was quantified on a NanoDrop ND-1000 Full-spectrum UV/Vis Spectrophotometer after extraction.

The AFLP protocol used was based on that described by Vos *et al.* (1995) and adapted for capillary electrophoresis and fluorescent detection as described by Skøt *et al.* (2005). Around 110 ng of total genomic DNA was digested to completion with the restriction enzymes EcoRI and MseI, followed by ligation of EcoRI and MseI adapters, with incubation for 2 h at 37°C. Pre-amplification was carried out in 20 μl, including 4 μl of ligated DNA, 15 μl AFLP Amplification Core Mix (ABI) and 1 μl Preselective Primer Mix (ABI) containing primers with one selective base (EcoRI-A and MseI-C). The selective amplification was carried out in 10 μl including 1.5 μl pre-amp product, 5.8 μl sterile distilled water, 1μl 10x AmpliTaq buffer (ABI), 0.04 μl of 5 units/μl of AmpliTaq Gold (ABI), 0.6 μl of 25 nM MgCl2, 0.08 μl of 25 mM dNTPs mix, and 0.5 μl of unlabelled MseI and fluorescently labelled EcoRI selective primers. Primer pairs were selected for primer optimization on the basis of a literature search and previous AFLP analysis at IGER (Kölliker *et al.*, 2003; Herrmann *et al.*, 2005; Hirano, 2005). Following optimization primer pairs were selected on the basis of highest polymorphism detected and peak detection quality (Table 3.2).

Pre-amplification PCR was carried out in PE 9700 Gene Amp System for 2 min at 72°C, followed by 20 cycles of 20 s of initial denaturation at 94°C, 30 s of annealing at 56°C,
2 min of extension at 72°C, and 45 min of elongation at 60°C. Selective amplification was for 10 min of denaturation at 95°C, followed by 13 cycles of 20 s denaturation at 94°C, 30 s of annealing at 66-56°C (decreasing by 0.7°C per cycle), 2 min of extension at 72°C, followed by a further 20 cycles of 20 s of denaturation at 94°C, 30 s of annealing at 56°C and 2 min of extension at 72°C, followed by 30 min of elongation at 60°C. 10 μl of each reaction was run on a 1% agarose gel, stained with ethidium bromide to ensure amplification had occurred. All centrifugations were carried out on a Sigma Laboratory Centrifuge 4-15C (Qiagen).

The selective amplification products were run on an ABI3130xl Genetic Analyzer (Applied Biosystems) and visualised with GENEMAPPER (version 4.0), typed as present or absent (above or below a threshold intensity).

Table 3.2 Primer optimization, detailing the number of polymorphisms observed and peak detection quality for each primer pair across eight randomly chosen individuals. *Primer pairs chosen for further analysis.

<table>
<thead>
<tr>
<th>Species</th>
<th>MseI-CAC</th>
<th>EcoRI-AAC</th>
<th>MseI-CAC</th>
<th>EcoRI-AAC</th>
<th>MseI-CAC</th>
<th>EcoRI-AAC</th>
<th>MseI-CAC</th>
<th>EcoRI-AAC</th>
<th>MseI-CAC</th>
<th>EcoRI-AAC</th>
<th>MseI-CAC</th>
<th>EcoRI-AAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trifolium dubium</td>
<td>24*</td>
<td>n/a</td>
<td>n/a</td>
<td>20</td>
<td>n/a</td>
<td>21</td>
<td>41*</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak detection quality</td>
<td>High</td>
<td>n/a</td>
<td>n/a</td>
<td>High</td>
<td>n/a</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Trifolium repens</td>
<td>n/a</td>
<td>n/a</td>
<td>146*</td>
<td>205*</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak detection quality</td>
<td>n/a</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trifolium pratense</td>
<td>42*</td>
<td>n/a</td>
<td>n/a</td>
<td>28</td>
<td>n/a</td>
<td>49</td>
<td>66*</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak detection quality</td>
<td>High</td>
<td>n/a</td>
<td>n/a</td>
<td>High</td>
<td>n/a</td>
<td>Medium</td>
<td>High</td>
<td>High</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Presence (1) or absence (0) of AFLP markers was entered into a binary matrix. Bins were set to 1 base pair (bp) and profiles were assessed from 50 to 500 bp in size. All AFLP profiles were checked visually to correct any misinterpretations of the Genemapper output and only peaks and individuals that could be scored unambiguously were included in the
analysis. To assess the reproducibility of the AFLP genotyping, replicate AFLP profiles were produced from 20 randomly chosen individuals for each species and each primer pair. The duplicate AFLP fingerprints for both primer pairs were compared visually following the method outlined by (Bonin et al., 2004), with the results shown in Table 3.3.

Table 3.3: AFLP error rate for three species of *Trifolium* based on the method outlined by Bonin et al. (2004).

<table>
<thead>
<tr>
<th>Species</th>
<th>Differences</th>
<th>Total comparisons</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Trifolium dubium</em></td>
<td>39</td>
<td>1401</td>
<td>2.78%</td>
</tr>
<tr>
<td><em>Trifolium pratense</em></td>
<td>53</td>
<td>1421</td>
<td>3.73%</td>
</tr>
<tr>
<td><em>Trifolium repens</em></td>
<td>58</td>
<td>1328</td>
<td>4.36%</td>
</tr>
</tbody>
</table>

3.2.3 Population genetic analysis

Due to the dominant nature of AFLP, observed heterozygosity cannot be determined and expected heterozygosity cannot be determined directly, hence allele frequencies were estimated by a Bayesian method (Zhivotovsky, 1999), assuming both a non-uniform prior distribution of allele frequencies and Hardy-Weinburg proportions using AFLP-SURV (version 1.0) (Vekemans, 2002). This method of estimating allele frequencies was chosen due to its relatively unbiased estimates compared to other methods; taking into account the sample sizes used in this study (see Krauss, 2000 for discussion; Isabel et al., 1999; Zhivotovsky, 1999). It should be noted that these methods are designed for a maximum of two alleles per locus, and using them in a tetraploid species such as *T. dubium* and *T. repens* can lead to an overestimate of allele frequencies and an underestimate of within-population genetic diversity (Lynch & Milligan, 1994; Krauss, 2000). However, as yet, there are no programs available that enable AFLP analysis in allotetraploid species.
Mating was assumed to be random in *T. pratense* and *T. repens*, and therefore not to deviate from Hardy-Weinburg proportions, due to their self-incompatibility systems (Atwood, 1942, 1944; Williams & Williams, 1947; Lawrence, 1996). However, due to the potential of a mixed-mating system in *T. repens* through clonal reproduction, analyses were also carried out assuming an $F_{IS}$ of 0.5 to determine to what extent the results are dependent on the assumption of outbreeding. For *T. dubium*, a predominantly self-fertilising species, a measure of the fixed deviation from Hardy-Weinberg proportions ($F_{IS}$) was used. In the absence of a published $F_{IS}$ value for this species, the self-fertilisation rate ($s$) of 0.97 (Dhar *et al*., 2006) was used to calculate the expected equilibrium value of $F_{IS}$ under a mixed mating model (see Hartl & Clark, 1989):

Estimates of allele frequencies were used to calculate genetic diversity and population genetic structure following the methods outlined by Lynch and Milligan (1994). Notations follow Lynch and Milligan (1994), with $H_j$ analogous to Nei’s (1978) unbiased heterozygosity $H_e$ and $H_w$ analogous to $H_s$. Wright’s (1951) $F_{ST}$ was calculated for the overall sample to test for differentiation between populations, in which a test of significance was performed comparing observed $F_{ST}$ with the distribution of $F_{ST}$ assuming no genetic structure, obtained using 1000 permutations of individuals among groups.

Allelic richness, a measure of genetic diversity identified as of high importance in prioritizing populations for conservation (Marshall & Brown, 1975; Petit *et al*., 1998), was determined using AFLPDiv (version 1.1) (Coart *et al*., 2005). Allelic richness is dependent on sample size, with uneven sample sizes biasing estimates. Larger sample sizes and intensively sampled regions may show a higher allelic richness or greater private alleles than smaller, less sampled populations and areas (Kalinowski, 2004). This bias can be overcome using
rarefaction methods to standardize allelic richness to a fixed sample size. In this study rarefaction was used to compare allelic richness at the smallest population size in each species.

To determine how distinct island populations are compared to mainland populations within the UK, unweighted pair group method with arithmetic mean (UPGMA) analysis was conducted based on matrices of pairwise genetic distances based on Nei’s (1972) measure of genetic distance after Lynch and Milligan (1994). Dendrograms were constructed using the NEIGHBOR program of PHYLIP (version 3.67) (Felsenstein, 2004). A majority rule consensus tree was constructed in the consense package in PHYLIP (Felsenstein, 2004), using 1000 replicated matrices produced in AFLP-SURV (Vekemans, 2002). For *T. repens* a model-based Bayesian clustering method was used to assign individuals to populations by the program STRUCTURE (version 2.2) to further determine levels of gene flow between crop and wild species (Pritchard *et al*., 2000; Falush *et al*., 2003, 2007). This method identifies *K* (unknown) populations within a dataset and assigns each population/individual to one or more population/cluster if the individual is admixed. To determine the most probable number of *K* an admixture model with correlated allele frequencies was run fifteen times for each value of *K*, with a burn-in period of $10^3$ for $10^6$ iterations. Different levels of *K* were examined based on the number of collection groups/areas plus 2 (*K* = 2-10). The rate of change of the log probability between successive values of *K* ($\Delta K$) was evaluated according to the method defined by Evanno *et al.* (2005) to determine the correct estimation of the number of clusters.

Analysis of molecular variance was carried out using ARLEQUIN (version 2.0) (Schneider *et al*. 2000) to estimate the partition of variation among regions, among populations within regions and within populations. Population regions were identified as groups defined by clustering analysis.
3.2.4 Isolation by Distance

To test for isolation by distance, pairwise $F_{ST}$ values transformed to $F_{ST}/(1- F_{ST})$ were compared with log-transformed geographic distances in a Mantel test following Rousset (1997). Mantel tests were carried out in GENALEX (version 6.1) (Peakall & Smouse, 2006), with significance tested using 999 permutations.

To further determine spatial structure, spatial autocorrelation was used to assess kinship over defined distance classes (using a distance matrix created in GENALEX). Kinship coefficients are most easily described as the probability of identity by descent of loci, with the use of kinship coefficients with dominant data discussed by Hardy (2003). Using this method, negative kinship coefficients can occur when two individuals are less related than two individuals taken at random (Hardy & Vekemans, 2003). In the absence of a consensus on the spatial scale at which to study genetic diversity patterns or at which to determine distance classes (see Escudero et al., 2003; Manel et al., 2003 for reviews), pairwise spatial distances (km) were grouped into classes taking into account the similarity of sample sizes in each class (see Appendix 3 for distance matrices). This method ensures that only pairs separated by less than half the maximum distance were considered in each distance class (Le Corre et al., 1998). Significance of the spatial structure was tested using 1000 permutations, with analysis conducted using SPAGEDI (Hardy & Vekemans, 2002). This method follows the equal frequency method outlined by Escudero et al. (2003), a method comprising of unequal distance lags with equal pairwise data in each class.

*T. dubium*: Average kinship coefficients $F_{ij}$ (Hardy, 2003) were computed for the following 5 distance classes; $\leq 40.60$km, $\leq 378.36$km, $\leq 509.05$km, $\leq 787.03$km and $\leq 825.53$km. An inbreeding coefficient ($F_{IT}$) is required to compute an estimate of kinship coefficients between population groups using SPAGEDI (Hardy & Vekemans, 2002). As such
the inbreeding coefficient of 0.97 \( (F_{IS}, \text{Dhar et al., 2006}) \) and calculated measure of \( F_{ST} \) (generated in AFLP-SURV) were used to calculate \( F_{IT} \) as per Hamilton (2009). However Hardy and Vekemans (2002) note that the estimate of kinship coefficient is robust to some error in inbreeding coefficient.

**T. pratense**: Average kinship coefficients \( F_{ij} \) (Hardy, 2003) were computed for the following 8 distance classes; \( \leq 39.75\text{km} \), \( \leq 208.28\text{km} \), \( \leq 397.81\text{km} \), \( \leq 429.45\text{km} \), \( \leq 521.82\text{km} \), \( \leq 752.26\text{km} \), \( \leq 860.52\text{km} \) and \( \leq 1211.33\text{km} \).

**T. repens**: Average kinship coefficients \( F_{ij} \) (Hardy, 2003) were computed for the following 7 distance classes; \( \leq 38.88\text{km} \), \( \leq 165.01\text{km} \), \( \leq 372.73\text{km} \), \( \leq 476.98\text{km} \), \( \leq 748.78\text{km} \), \( \leq 863.62\text{km} \) and \( \leq 1209.49\text{km} \).

Due the distribution of sample sites there are clearly uneven lags between the distance classes for all species. Manual manipulation of spatial classes to ensure similar distances are covered by each distance class (equal interval method) showed equivalent results. However, using this method numbers of pairwise comparisons within each interval were uneven and occasionally lower than the 30 pairs of data per class recommended by Legendre and Fortin (1989).

### 3.2.5 Environmental impact on genetic diversity

To determine the extent to which environmental variables affect patterns of genetic diversity, over and above geographical distance, a distance based redundancy (dbRDA)
method was used (Legendre & Anderson, 1999; McArdle & Anderson, 2001; Anderson, 2001). dbRDA, a method of multivariate multiple regression, can be performed directly on a distance or dissimilarity matrix. For this analysis two measures of genetic variation, Nei’s genetic distance and $F_{ST}$, were used as response variables to various sets of environmental predictor variables, including geographic distance (see Table 3.4). Due to the relatively small number of populations, predictor variables were considered separately as it was not feasible to include all variables in a single regression model.

Table 3.4 Variable sets used in dbRDA analyses. a Data obtained from nearest weather stations to collection sites, 1971-2001 climate averages, however see text for air frost levels in Isles of Scilly (Met Office, 2010). b Data obtained from mapped climate averages of 1971-2001 weather data (Met Office, 2010). c Data not available for St Kilda populations.

<table>
<thead>
<tr>
<th>Set</th>
<th>Variables included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response variable</td>
<td>Generated in AFLP-SURV</td>
</tr>
<tr>
<td>Nei’s genetic distance</td>
<td>Days where minimum temp &lt;0 (days)</td>
</tr>
<tr>
<td>$F_{ST}$</td>
<td>Altitude in meters above sea level</td>
</tr>
<tr>
<td>Predictor variable</td>
<td>Geographic distance</td>
</tr>
<tr>
<td>Air frost$^{ab}$</td>
<td>Grass minimum temperature$^{b}$</td>
</tr>
<tr>
<td>Altitude</td>
<td>Island population</td>
</tr>
<tr>
<td>Distance</td>
<td>Island population indicator:</td>
</tr>
<tr>
<td>Geographic distance</td>
<td>pH$^{c}$</td>
</tr>
<tr>
<td>Grass minimum temperature$^{b}$</td>
<td>Rainfall$^{a}$</td>
</tr>
<tr>
<td>Island population</td>
<td>Snow$^{b}$</td>
</tr>
<tr>
<td>pH$^{c}$</td>
<td>Sunshine duration$^{a}$</td>
</tr>
<tr>
<td>Rainfall$^{a}$</td>
<td>Temperature$^{a}$</td>
</tr>
</tbody>
</table>

Two sets of analyses were conducted, i) marginal tests on all sets of predictor variables ii) conditional tests using geographical co-ordinates as covariables in the model to determine the extent that environmental variables can determine genetic distance when accounting for
the variation explained by spatial distance. Significance of the model was determined by running 9999 permutations of the rows and columns of the residual matrix under the full model for both marginal and conditional tests (Anderson & Legendre, 1999). All dbRDA analyses were conducted using the program DISTLM (McArdle & Anderson, 2001).

St Kilda climate data was available from 1999-2008 (J. Pemberton pers. comm. April, 2008). A correlation was performed to compare available weather data from St Kilda with Stornoway weather data collected by the Meteorological (Met) Office, an island approximately 100km from St Kilda. Maximum and minimum daily temperature, precipitation, sunshine duration and days of air frost were averaged from weather stations on St Kilda and compared with equivalent data from Stornoway weather station (Table 3.5).

Table 3.5 Correlation ($r^2$) between Stornoway/St Kilda weather station data from 2000-2007 and Tirree/Skye and St Mawgan/Isles of Scilly weather data from 2001-2006. *** indicates significance level P<0.001. a Data not available; b See note in text about air frost days in Scilly.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Maximum temperature (°C)</th>
<th>Minimum temperature (°C)</th>
<th>Precipitation (mm)</th>
<th>Sun (hours/year)</th>
<th>Air frost (days/year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>St Kilda/Stornoway</td>
<td>95.3%***</td>
<td>90.0%***</td>
<td>74.5%***</td>
<td>86.3%***</td>
<td>57.1%***</td>
</tr>
<tr>
<td>Skye/Tirree</td>
<td>96.6%***</td>
<td>97.7%***</td>
<td>65.7%***</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>Isle of Scilly/St Mawgan</td>
<td>97.4%***</td>
<td>97.5%***</td>
<td>70.2%***</td>
<td>a</td>
<td>b</td>
</tr>
</tbody>
</table>

High significance between the two areas confirm that the Stornoway weather data captures local conditions closely and can be used as a surrogate for historic weather data from St Kilda (see also Hallett et al., 2004). Additional tests were performed for Skye and the Isles of Scilly as the nearest available weather station data was deemed to be potentially too far from the collection sites to be representative. Data was obtained direct from the Met Office for Skye (Lusa) and Isles of Scilly (St Mary’s Airport) covering the years 2001 to 2006 and compared with corresponding years data from Tirree and St Mawgan respectively (see Table
3.5). Due to the high correlation the historical data available from these weather stations was deemed sufficient surrogate for historical data from either Skye or the Isles of Scilly. However, due to the lack of a correlation of air frost days between the available data for the Isles of Scilly and St Mawgan it was decided to use the published figure of an average of 2 days per year for the Isles of Scilly (Met Office, 2010).

3.2.6 ISLAND VERSUS MAINLAND HETEROZYGOSITY

A comparison between island and mainland expected heterozygosity ($H_i$) was determined according to Frankham (1997). Following this method, in the presence of multiple mainland sites, mainland expected heterozygosity levels are averaged and compared to each island population to obtain a ratio of cases where mainland populations have a higher versus lower genetic variation than island populations. In addition allelic richness was compared between island and mainland populations, standardised to account for differences in sample size. Due to the proximity of Skye to the mainland, and therefore a lack of significant water barrier to gene flow, populations from Skye were included in mainland heterozygosity levels.
3.3 RESULTS - *TRIFOLIUM DUBIUM*

3.3.1 DESCRIPTIVE POPULATION GENETICS

A total of 127 loci were scored across 165 individuals of *T. dubium*, 70 from the MseI-CAC/EcoRI-AAC primer pair and 57 from the MseI-CTA/EcoRI-AGA primer pair, with on average 41.87% polymorphic per population. The expected heterozygosity (*Hj*) ranged from 0.059 (UIS7 population) to 0.216 (IOS3 population), with the average within population expected heterozygosity (*Hw*), 0.149 ± 0.015 over all populations. Detailed descriptive statistics for each population is given in Table 3.6.

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample size (n)</th>
<th>Polymorphic loci (%) a</th>
<th>Allelic richness (<em>A15</em>)</th>
<th>Expected heterozygosity (<em>Hj</em>)</th>
<th>S.E.(<em>Hj</em>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEV2</td>
<td>15</td>
<td>52</td>
<td>1.575</td>
<td>0.186</td>
<td>0.017</td>
</tr>
<tr>
<td>DEV3</td>
<td>15</td>
<td>42.5</td>
<td>1.512</td>
<td>0.167</td>
<td>0.018</td>
</tr>
<tr>
<td>IOS1</td>
<td>15</td>
<td>50.4</td>
<td>1.504</td>
<td>0.184</td>
<td>0.017</td>
</tr>
<tr>
<td>IOS2</td>
<td>15</td>
<td>52</td>
<td>1.52</td>
<td>0.193</td>
<td>0.018</td>
</tr>
<tr>
<td>IOS3</td>
<td>15</td>
<td>62.2</td>
<td>1.622</td>
<td>0.216</td>
<td>0.017</td>
</tr>
<tr>
<td>LKD2</td>
<td>15</td>
<td>44.1</td>
<td>1.441</td>
<td>0.143</td>
<td>0.016</td>
</tr>
<tr>
<td>LKD5</td>
<td>15</td>
<td>31.5</td>
<td>1.315</td>
<td>0.102</td>
<td>0.013</td>
</tr>
<tr>
<td>SKY1</td>
<td>15</td>
<td>34.6</td>
<td>1.433</td>
<td>0.107</td>
<td>0.013</td>
</tr>
<tr>
<td>SKY3</td>
<td>15</td>
<td>37.8</td>
<td>1.378</td>
<td>0.151</td>
<td>0.018</td>
</tr>
<tr>
<td>SKY5</td>
<td>15</td>
<td>35.4</td>
<td>1.354</td>
<td>0.131</td>
<td>0.016</td>
</tr>
<tr>
<td>UIS7</td>
<td>15</td>
<td>18.1</td>
<td>1.315</td>
<td>0.059</td>
<td>0.011</td>
</tr>
</tbody>
</table>

3.3.2 POPULATION SUBDIVISION

Genetic differentiation among populations, with an *F*<sub>ST</sub> value of 0.303 (P<0.001), indicates a moderately high and significant level of differentiation among populations. A three level AMOVA was used to partition the total variation into among groups (geographic
regions), among populations within regions and within populations (Table 3.7). Populations were separated into two geographic regions based on the clustering observed in the UPGMA tree; northern and southern UK. The largest component of total variation was found within populations (63.59%), with 25.92% found among populations.

Table 3.7 Analysis of molecular variance (AMOVA) of variation across 165 *T. dubium* individuals conducted in ARLEQUIN (Schneider et al., 2000). Probability tested using 1000 permutations.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>Sum of squares</th>
<th>Variance components</th>
<th>Percentage of variation</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among regions</td>
<td>1</td>
<td>171.85</td>
<td>1.42</td>
<td>10.48</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Among populations within regions</td>
<td>9</td>
<td>540.79</td>
<td>3.52</td>
<td>25.92</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Within populations</td>
<td>154</td>
<td>1293.69</td>
<td>8.63</td>
<td>63.59</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

3.3.3 GENETIC RELATIONSHIPS AMONG POPULATIONS

UPGMA analysis revealed some structure, dividing the populations into two geographic regions; the first including those populations from northern England and Scotland, with the second grouping the Isles of Scilly with southern England (Figure 3.4). The southern UK grouping shows clear delineation between the population clusters of the Isles of Scilly and Devon, while those populations from central and northern UK show no particular grouping according to geographic region.
Figure 3.4 Unrooted unweighted pair group method with arithmetic mean (UPGMA) dendrogram based on Nei's genetic distances after Lynch and Milligan (1994). Values at the nodes represent 1000 bootstrap values shown as percentages; only values over 50% are shown.
3.3.4 **Isolation by Distance**

A significant relationship between pairwise $F_{ST}$ and geographic distance was observed across all populations ($r = 0.282$, $P<0.01$) (Figure 3.5a). No significant relationship between genetic and geographic distance was detected in mainland populations, when excluding populations from the Isles of Scilly and Outer Hebrides (Figure 3.5b).

![Graph a](image1.png)

![Graph b](image2.png)

Figure 3.5 Mantel tests of isolation by distance, plotting transformed $F_{ST}$ (Rousset, 1997) against log transformed distance values; a all populations b all mainland populations (including Skye).
Spatial autocorrelation analysis produced a negative correlation with average kinship coefficients decreasing over increasing distance (Figure 3.6). Significant positive kinship coefficients were identified between individuals in the distance class less than 40.60km apart (with a mean distance 10.33km), suggesting that *T. dubium* may be able to form a related group within this distance. Clearly this distance reflects the sampling strategy used and therefore geographic distance between selected sampling sites and indicates only a preliminary genetic patch size, with further analysis on a continuously sampled population required for a more accurate representation of kinship.

![Figure 3.6](image.png)

**Figure 3.6** Pairwise average kinship coefficient of 165 *T. dubium* individuals plotted against geographic distance (km) using SPAGEDI (Hardy & Vekemans 2002); * indicates significant deviation from random mating among individuals (*P*<0.05).

### 3.3.5 ENVIRONMENTAL IMPACTS ON PATTERNS OF GENETIC DIVERSITY

Using dbRDA, marginal tests showed that a large proportion of variation in both *F*<sub>ST</sub> and Nei’s genetic distance is explained by deviation in spatial distribution in confirmation of
the above analysis. Further, marginal tests indicate the influence of various climatic factors on variance in genetic distance between populations (Table 3.8).

Table 3.8 Redundancy based analysis of the effect of environmental factors on genetic differentiation in T. dubium populations. All marginal tests are shown on the left, with conditional tests on the right (including spatial distance as covariables in the analysis). Values in bold represent significant $P$ values ($<0.1$). $r^2$ values represent the proportion of variation explained by each environmental variable.

<table>
<thead>
<tr>
<th>Variable set</th>
<th>F</th>
<th>P</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nei’s genetic distance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance</td>
<td>3.697</td>
<td>0.014</td>
<td><strong>0.480</strong></td>
</tr>
<tr>
<td>Air frost</td>
<td>1.682</td>
<td>0.191</td>
<td>0.158</td>
</tr>
<tr>
<td>Altitude</td>
<td>1.111</td>
<td>0.382</td>
<td>0.101</td>
</tr>
<tr>
<td>Grass</td>
<td>4.397</td>
<td>0.023</td>
<td><strong>0.328</strong></td>
</tr>
<tr>
<td>Island</td>
<td>1.499</td>
<td>0.230</td>
<td>0.143</td>
</tr>
<tr>
<td>pH</td>
<td>0.752</td>
<td>0.524</td>
<td>0.008</td>
</tr>
<tr>
<td>Rainfall</td>
<td>4.879</td>
<td>0.001</td>
<td><strong>0.352</strong></td>
</tr>
<tr>
<td>Snow</td>
<td>2.784</td>
<td>0.060</td>
<td><strong>0.236</strong></td>
</tr>
<tr>
<td>Sun hours</td>
<td>6.536</td>
<td>0.003</td>
<td><strong>0.421</strong></td>
</tr>
<tr>
<td>Temperature</td>
<td>4.535</td>
<td>0.005</td>
<td><strong>0.531</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conditional tests</th>
<th>Variable set</th>
<th>F</th>
<th>P</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air frost</td>
<td>2.857</td>
<td>0.102</td>
<td>0.151</td>
<td></td>
</tr>
<tr>
<td>Altitude</td>
<td>1.109</td>
<td>0.364</td>
<td>0.071</td>
<td></td>
</tr>
<tr>
<td>Grass</td>
<td>1.832</td>
<td>0.216</td>
<td>0.108</td>
<td></td>
</tr>
<tr>
<td>Island</td>
<td>3.942</td>
<td>0.062</td>
<td><strong>0.187</strong></td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td>0.641</td>
<td>0.569</td>
<td>0.044</td>
<td></td>
</tr>
<tr>
<td>Rainfall</td>
<td>0.671</td>
<td>0.569</td>
<td>0.045</td>
<td></td>
</tr>
<tr>
<td>Snow</td>
<td>0.461</td>
<td>0.656</td>
<td>0.032</td>
<td></td>
</tr>
<tr>
<td>Sun hours</td>
<td>2.099</td>
<td>0.171</td>
<td>0.120</td>
<td></td>
</tr>
<tr>
<td>Temperature</td>
<td>1.867</td>
<td>0.194</td>
<td>0.199</td>
<td></td>
</tr>
</tbody>
</table>

Accounting for spatial distance (conditional tests); grass minimum temperature, rainfall, hours of sunshine and temperature become non significant, most likely due to their correlation to spatial distance in the areas studied for these species. By fitting spatial distance as covariables in the analysis only the distinction of island populations, differentiating
populations separated by a significant water barrier, remains significant for GD, accounting for 21% of the variation. Geffen et al. (2004) suggest that variables with P-values lower than 0.10 can be used to identify variables that may be significant to account for the small number of populations and therefore the potential lack of power in the analyses (see also Smith et al., 2007). As such, both air frost and the number of sun hours may also show a significant relationship with genetic variation after accounting for geographic distance.

3.3.6 ISLAND VERSUS MAINLAND HETEROZYGOSITY

Island populations of *T. dubium* are shown to have, in general, higher expected heterozygosity and allelic richness than mainland populations, with all populations from the Isles of Scilly maintaining high heterozygosity levels. In contrast, the population from Uist has the lowest heterozygosity level per population (Table 3.9). However these results may be deceptive as they only relate to two island groups in the UK, and are biased towards the three populations collected from the Isles of Scilly. Populations were extremely rare on the other island groups visited in this study and further collections would be required to confirm these results.

Table 3.9 Island versus mainland heterozygosity following the method outlined by Frankham (1997) for within population expected heterozygosity/allelic richness in *T. dubium*. *a* On the left are the number of occasions where mainland population expected heterozygosity/allelic richness is greater than island populations, on the right are occasions when island expected heterozygosity/allelic richness is greater than mainland populations.

<table>
<thead>
<tr>
<th>Island population</th>
<th>$H_j$</th>
<th>M &gt; Is: M &lt; Is $^a$</th>
<th>$A$</th>
<th>M &gt; Is: M &lt; Is $^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOS1</td>
<td>0.184</td>
<td>0:1</td>
<td>1.504</td>
<td>0:1</td>
</tr>
<tr>
<td>IOS2</td>
<td>0.193</td>
<td>0:1</td>
<td>1.52</td>
<td>0:1</td>
</tr>
<tr>
<td>IOS3</td>
<td>0.216</td>
<td>0:1</td>
<td>1.622</td>
<td>0:1</td>
</tr>
<tr>
<td>UIS7</td>
<td>0.059</td>
<td>1:0</td>
<td>1.315</td>
<td>1:0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1:3</td>
<td>1.429</td>
<td></td>
</tr>
</tbody>
</table>
3.4 Results - *Trifolium pratense*

3.4.1 Descriptive Population Genetics

A total of 244 loci were scored across 231 individuals of *T. pratense*, 130 from the MseI-CAC/EcoRI-AAC primer pair and 114 from the MseI-CTA/EcoRI-AGA primer pair, with an average 44.81% polymorphic per population. The expected heterozygosity ($H_e$) ranged from 0.128 (SHT6 population) to 0.169 (LKD1 population). The average within population expected heterozygosity, $H_w = 0.154 \pm 0.003$ over all populations (for heterozygosity of each sample site see Table 3.10).

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample size (n)</th>
<th>Polymorphic loci (%) a</th>
<th>Allelic richness ($A_9$)</th>
<th>Expected heterozygosity ($H_e$)</th>
<th>S.E.($H_e$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEV2</td>
<td>15</td>
<td>43.4</td>
<td>1.443</td>
<td>0.154</td>
<td>0.011</td>
</tr>
<tr>
<td>DEV3</td>
<td>15</td>
<td>48.8</td>
<td>1.445</td>
<td>0.160</td>
<td>0.011</td>
</tr>
<tr>
<td>IOS4</td>
<td>15</td>
<td>44.3</td>
<td>1.423</td>
<td>0.153</td>
<td>0.011</td>
</tr>
<tr>
<td>IOS5</td>
<td>9</td>
<td>52.9</td>
<td>1.475</td>
<td>0.169</td>
<td>0.011</td>
</tr>
<tr>
<td>LKD1</td>
<td>12</td>
<td>56.6</td>
<td>1.51</td>
<td>0.166</td>
<td>0.011</td>
</tr>
<tr>
<td>LKD3</td>
<td>15</td>
<td>45.9</td>
<td>1.522</td>
<td>0.168</td>
<td>0.01</td>
</tr>
<tr>
<td>LKD6</td>
<td>15</td>
<td>43.4</td>
<td>1.389</td>
<td>0.149</td>
<td>0.01</td>
</tr>
<tr>
<td>NWS4</td>
<td>15</td>
<td>45.5</td>
<td>1.417</td>
<td>0.156</td>
<td>0.011</td>
</tr>
<tr>
<td>SKY3</td>
<td>15</td>
<td>46.3</td>
<td>1.426</td>
<td>0.157</td>
<td>0.010</td>
</tr>
<tr>
<td>SKY5</td>
<td>15</td>
<td>45.1</td>
<td>1.46</td>
<td>0.159</td>
<td>0.011</td>
</tr>
<tr>
<td>UIS5</td>
<td>15</td>
<td>45.5</td>
<td>1.429</td>
<td>0.154</td>
<td>0.011</td>
</tr>
<tr>
<td>UIS6</td>
<td>15</td>
<td>38.1</td>
<td>1.398</td>
<td>0.153</td>
<td>0.012</td>
</tr>
<tr>
<td>UIS8</td>
<td>15</td>
<td>44.7</td>
<td>1.419</td>
<td>0.161</td>
<td>0.011</td>
</tr>
<tr>
<td>SHT2</td>
<td>15</td>
<td>40.2</td>
<td>1.403</td>
<td>0.147</td>
<td>0.011</td>
</tr>
<tr>
<td>SHT4</td>
<td>15</td>
<td>37.3</td>
<td>1.336</td>
<td>0.128</td>
<td>0.011</td>
</tr>
<tr>
<td>SHT6</td>
<td>15</td>
<td>38.9</td>
<td>1.355</td>
<td>0.128</td>
<td>0.011</td>
</tr>
</tbody>
</table>
3.4.2 Population Subdivision

Genetic differentiation among populations, with an $F_{ST}$ value of 0.099 ($P<0.001$), indicates moderate but significant differentiation among populations. A three level AMOVA was used to partition the total variation into among geographic regions, among populations within regions and within populations (Table 3.11). Populations were separated into the four geographic regions based on the clustering observed in the UPGMA tree, the first including southern England populations including those from the Isles of Scilly, the second including populations from North West Scotland (including UIS5), the third encompassing two populations from the Outer Hebrides and the fourth including populations from Shetland. The largest component of total variation was found within populations (78.82%), with only 10.41% found among populations. This high level of variation within populations is in accordance with previous assessments of genetic diversity in *T. pratense* (Hagen & Hamrick, 1998; Mosjidis *et al.*, 2004).

![Table 3.11 Analysis of molecular variance (AMOVA) of variation across 231 *T. pratense* individuals conducted in ARLEQUIN (Schneider *et al.*, 2000). Probability tested using 1000 permutations.](image)

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>Sum of squares</th>
<th>Variance components</th>
<th>Percentage of variation</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among regions</td>
<td>3</td>
<td>517.65</td>
<td>2.44</td>
<td>10.77</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Among populations within regions</td>
<td>12</td>
<td>588.92</td>
<td>2.36</td>
<td>10.41</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Within populations</td>
<td>215</td>
<td>3459.97</td>
<td>17.88</td>
<td>78.82</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

3.4.3 Genetic Relationships Among Populations

UPGMA analysis of *T. pratense* populations reveal the populations grouped according to geographic region, excluding one population from the Outer Hebrides which is more genetically similar to populations from the Inner Hebrides and the Lake District (Figure 3.7). Shetland remains the most unique of the geographic regions analysed, followed by two
populations from the Outer Hebrides, with the rest of the populations grouped in the third cluster. This final group is separated along a north and south divide.

Figure 3.7 Unrooted unweighted pair group method with arithmetic mean (UPGMA) dendrogram based on Nei’s genetic distances after Lynch and Milligan (1994). Values at the nodes represent 1000 bootstrap values shown as percentages; only values over 50% are shown.
3.4.4 ISOLATION BY DISTANCE

A significant relationship between pairwise distance and distance was observed across all populations \( (r = 0.507, P \leq 0.001) \) (Figure 3.8a). Excluding populations from the Isles of Scilly, Outer Hebrides and Shetland significant isolation by distance \( (r = 0.789, P \leq 0.001) \) was observed between populations on the mainland (Figure 3.8b).

![Mantel tests of isolation by distance](image)

Figure 3.8 Mantel tests of isolation by distance, plotting transformed \( F_{ST} \) (Rousset, 1997) against log transformed distance values: \( ^a \) all populations \( ^b \) all mainland populations (including Skye collections).
Spatial autocorrelation analysis produced a negative correlation with average kinship coefficient decreasing over increasing distance, confirming the above results (Figure 3.9). Significant kinship coefficients were identified between individuals less than 208.28km apart, suggesting that *T. pratense* may be able to form a related group up to this distance. Clearly this distance reflects the sampling strategy used and indicates only a preliminary genetic patch size, with further analysis on a continuously sampled population required for a more accurate representation of kinship.

![Figure 3.9 Pairwise average kinship coefficient of 231 *T. pratense* individuals plotted against geographic distance (km) using SPAGEDi (Hardy & Vekemans 2002); * indicates significant deviation from random mating among individuals (P<0.05).](image)

### 3.4.5 ENVIRONMENTAL IMPACTS ON PATTERNS OF GENETIC DIVERSITY

As shown above, significant isolation by distance was determined across the populations sampled indicating a significant proportion of variation is explained by variation in geographic distance. Marginal dbRDA tests confirm this pattern across both the $F_{ST}$ and Nei’s genetic distance data sets with up to 51% of variation explained by differences in
latitude and longitude. In addition many environmental variables appear to explain significant proportions of variation in genetic diversity (Table 3.12).

Table 3.12 Redundancy based analysis of the effect of environmental factors on genetic differentiation in *T. pratense* populations. All marginal tests are shown on the left, with conditional tests on the right (including spatial distance as covariables in the analysis). Values in bold represent significant *P* values (<0.1). *r*² values represent the proportion of variation explained by each environmental variable.

<table>
<thead>
<tr>
<th>Variable set</th>
<th>F</th>
<th>P</th>
<th>r²</th>
<th>Variable set</th>
<th>F</th>
<th>P</th>
<th>r²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nei’s genetic distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance</td>
<td>6.705</td>
<td>0.000</td>
<td><strong>0.508</strong></td>
<td>Air frost</td>
<td>2.915</td>
<td>0.072</td>
<td><strong>0.096</strong></td>
</tr>
<tr>
<td>Air frost</td>
<td>1.199</td>
<td>0.349</td>
<td>0.079</td>
<td>Altitude</td>
<td>0.935</td>
<td>0.472</td>
<td>0.036</td>
</tr>
<tr>
<td>Altitude</td>
<td>0.429</td>
<td>0.732</td>
<td>0.029</td>
<td>Grass</td>
<td>0.487</td>
<td>0.687</td>
<td>0.019</td>
</tr>
<tr>
<td>Grass</td>
<td>4.139</td>
<td>0.010</td>
<td><strong>0.228</strong></td>
<td>Island</td>
<td>4.575</td>
<td>0.017</td>
<td><strong>0.136</strong></td>
</tr>
<tr>
<td>Island</td>
<td>1.278</td>
<td>0.329</td>
<td>0.084</td>
<td>pH</td>
<td>2.309</td>
<td>0.119</td>
<td>0.079</td>
</tr>
<tr>
<td>pH</td>
<td>1.982</td>
<td>0.135</td>
<td>0.124</td>
<td>Rainfall</td>
<td>2.146</td>
<td>0.073</td>
<td><strong>0.075</strong></td>
</tr>
<tr>
<td>Rainfall</td>
<td>2.192</td>
<td>0.113</td>
<td>0.135</td>
<td>Snow</td>
<td>0.242</td>
<td>0.830</td>
<td>0.009</td>
</tr>
<tr>
<td>Snow</td>
<td>4.876</td>
<td>0.008</td>
<td><strong>0.258</strong></td>
<td>Sun hours</td>
<td>2.559</td>
<td>0.083</td>
<td><strong>0.087</strong></td>
</tr>
<tr>
<td>Sun hours</td>
<td>5.754</td>
<td>0.004</td>
<td><strong>0.291</strong></td>
<td>Temperature</td>
<td>3.072</td>
<td>0.045</td>
<td><strong>0.176</strong></td>
</tr>
<tr>
<td>Temperature</td>
<td>3.042</td>
<td>0.017</td>
<td><strong>0.319</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance</td>
<td>6.375</td>
<td>0.000</td>
<td><strong>0.495</strong></td>
<td>Air frost</td>
<td>2.811</td>
<td>0.059</td>
<td><strong>0.096</strong></td>
</tr>
<tr>
<td>Air frost</td>
<td>1.226</td>
<td>0.324</td>
<td>0.081</td>
<td>Altitude</td>
<td>1.045</td>
<td>0.419</td>
<td>0.041</td>
</tr>
<tr>
<td>Altitude</td>
<td>0.497</td>
<td>0.709</td>
<td>0.034</td>
<td>Grass</td>
<td>0.744</td>
<td>0.564</td>
<td>0.029</td>
</tr>
<tr>
<td>Grass</td>
<td>3.891</td>
<td>0.011</td>
<td><strong>0.218</strong></td>
<td>Island</td>
<td>4.541</td>
<td>0.009</td>
<td><strong>0.139</strong></td>
</tr>
<tr>
<td>Island</td>
<td>1.486</td>
<td>0.247</td>
<td>0.096</td>
<td>pH</td>
<td>1.978</td>
<td>0.142</td>
<td>0.072</td>
</tr>
<tr>
<td>pH</td>
<td>1.774</td>
<td>0.161</td>
<td>0.113</td>
<td>Rainfall</td>
<td>1.891</td>
<td>0.105</td>
<td>0.069</td>
</tr>
<tr>
<td>Rainfall</td>
<td>2.013</td>
<td>0.125</td>
<td>0.126</td>
<td>Snow</td>
<td>0.399</td>
<td>0.779</td>
<td>0.016</td>
</tr>
<tr>
<td>Snow</td>
<td>4.238</td>
<td>0.011</td>
<td><strong>0.232</strong></td>
<td>Sun hours</td>
<td>2.507</td>
<td>0.066</td>
<td><strong>0.087</strong></td>
</tr>
<tr>
<td>Sun hours</td>
<td>5.291</td>
<td>0.004</td>
<td><strong>0.274</strong></td>
<td>Temperature</td>
<td>3.102</td>
<td>0.028</td>
<td><strong>0.182</strong></td>
</tr>
</tbody>
</table>

By accounting for variation explained by spatial distance in the form of covariables in the analysis (conditional tests) days of snow lying and grass minimum temperature become non significant, perhaps due to the correlation to latitude. Average monthly temperature and number of sun hours per year remain significant, indicating that these variables may be
important in determining genetic dissimilarities between populations. By fitting spatial
distance as covariables both air frost and the distinction of a population occurring on an island
become significant factors in the structuring of genetic diversity accounting for 9 and 13% of
the variation in genetic diversity respectively. The results of these tests show some relatively
high F values whilst remaining insignificant in terms of P values. This may be due to the lack
of statistical power involved in using small data sets with larger data sets required for the
comprehensive rejection of the null hypotheses.

3.4.6 ISLAND VERSUS MAINLAND HETEROZYGOSITY

In agreement with Frankham’s (1997) comprehensive review of mainland versus island
heterozygosities *T. pratense* shows that mainland populations can maintain higher levels of
heterozygosity within populations than island populations (Table 3.13).

Table 3.13 Island versus mainland heterozygosity following the method outlined by Frankham (1997) for within
population expected heterozygosity in *T. pratense*. * On the left are the number of occasions where mainland
population expected heterozygosity/allelic richness is greater than island populations, on the right are occasions
when island expected heterozygosity/allelic richness is greater than mainland populations.

<table>
<thead>
<tr>
<th>Island population</th>
<th>$H_j$</th>
<th>$M &gt; Is$: $M &lt; Is^a$</th>
<th>$A$</th>
<th>$M &gt; Is$: $M &lt; Is^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOS4</td>
<td>0.153</td>
<td>1:0</td>
<td>1.423</td>
<td>1:0</td>
</tr>
<tr>
<td>IOS5</td>
<td>0.169</td>
<td>0:1</td>
<td>1.475</td>
<td>0:1</td>
</tr>
<tr>
<td>SHT2</td>
<td>0.147</td>
<td>1:0</td>
<td>1.403</td>
<td>1:0</td>
</tr>
<tr>
<td>SHT4</td>
<td>0.128</td>
<td>1:0</td>
<td>1.336</td>
<td>1:0</td>
</tr>
<tr>
<td>SHT6</td>
<td>0.128</td>
<td>1:0</td>
<td>1.355</td>
<td>1:0</td>
</tr>
<tr>
<td>UIS5</td>
<td>0.154</td>
<td>1:0</td>
<td>1.429</td>
<td>1:0</td>
</tr>
<tr>
<td>UIS6</td>
<td>0.153</td>
<td>1:0</td>
<td>1.398</td>
<td>1:0</td>
</tr>
<tr>
<td>UIS8</td>
<td>0.161</td>
<td>0:1</td>
<td>1.419</td>
<td>1:0</td>
</tr>
<tr>
<td>Total</td>
<td>6:2</td>
<td>7:1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

0.159

Average mainland heterozygosity ($H_j$) 0.159

Average mainland allelic richness ($A$) 1.452
3.5 RESULTS - *TRIFOLIUM REPENS*

3.5.1 DESCRIPTIVE POPULATION GENETICS

A total of 351 loci were scored across a total of 602 individuals of *T. repens*, 216 from the MseI-CCT/EcoRI-ACA primer pair and 135 from the MseI-CAC/EcoRI-ACT primer pair. The expected heterozygosity ranged from 0.082 (SHT1 population) to 0.118 (STK2 population), with the mean gene diversity within populations ($H_w$) 0.107 ± 0.002, and 0.110 when assuming inbreeding (Table 3.14). In general these values are lower than those found in comparative studies on dominant markers, where outcrossing, long-lived perennial species had on average slightly higher levels of genetic diversity with means of 0.242 for long-lived perennials and 0.260 for outcrossing species (Nybom & Bartish, 2000).

3.5.2 POPULATION SUBDIVISION

There was moderate differentiation among populations, $F_{ST} = 0.108$, $P<0.001$ and 0.153, $P<0.001$ when assuming inbreeding. A three level AMOVA was used to partition the total variation into among groups (geographic regions), among populations within groups and within populations (Table 3.15). Populations were separated into four geographic regions based on the clustering observed in the UPGMA tree (Figure 3.10), the first comprising mainland English populations (DEV1, DOR1-3, LKD2, 4, 6 and RYE1), the Isles of Scilly (IOS1-3) and the two landraces (ED1 and KWW1), the second grouping from North West Scotland (NWS1-3) and the Hebrides (SKY1-2, BEN1, UIS2 and UIS4), the third grouping from Shetland (SHT1,3,5), and the fourth grouping made up of those from St Kilda (STK1-5). The most significant component of total variation was found within populations (78.82%), in concordance with other assessments of genetic diversity in *T. repens* (Gustine & Huff, 1999; Kölliker *et al.*, 2001).
Table 3.14 Population genetic analysis of *T. repens* based on two AFLP primer pairs. Polymorphic loci and expected heterozygosity calculated in AFLP-SURV. *a* proportion of polymorphic loci with allelic frequencies lying within range 0.05 to 0.95.

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample size (n)</th>
<th>Polymorphic loci (%)&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Allelic richness ($A_{12}$)</th>
<th>Expected heterozygosity ($H_j$)</th>
<th>S.E.($H_j$)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cultivated varieties</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ED1</td>
<td>24</td>
<td>33.3</td>
<td>1.376</td>
<td>0.113</td>
<td>0.008</td>
</tr>
<tr>
<td>KWW1</td>
<td>24</td>
<td>31.3</td>
<td>1.389</td>
<td>0.111</td>
<td>0.008</td>
</tr>
<tr>
<td><strong>Wild / semi natural varieties</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEV1</td>
<td>12</td>
<td>25.1</td>
<td>1.35</td>
<td>0.099</td>
<td>0.008</td>
</tr>
<tr>
<td>DOR1</td>
<td>14</td>
<td>27.9</td>
<td>1.405</td>
<td>0.106</td>
<td>0.008</td>
</tr>
<tr>
<td>DOR2</td>
<td>22</td>
<td>31.3</td>
<td>1.382</td>
<td>0.103</td>
<td>0.008</td>
</tr>
<tr>
<td>DOR3</td>
<td>20</td>
<td>35.6</td>
<td>1.393</td>
<td>0.110</td>
<td>0.008</td>
</tr>
<tr>
<td>IOS1</td>
<td>15</td>
<td>32.2</td>
<td>1.385</td>
<td>0.111</td>
<td>0.008</td>
</tr>
<tr>
<td>IOS2</td>
<td>15</td>
<td>33.0</td>
<td>1.452</td>
<td>0.114</td>
<td>0.008</td>
</tr>
<tr>
<td>IOS3</td>
<td>15</td>
<td>31.9</td>
<td>1.388</td>
<td>0.113</td>
<td>0.008</td>
</tr>
<tr>
<td>LKD2</td>
<td>15</td>
<td>33.9</td>
<td>1.39</td>
<td>0.114</td>
<td>0.008</td>
</tr>
<tr>
<td>LKD4</td>
<td>15</td>
<td>29.1</td>
<td>1.364</td>
<td>0.103</td>
<td>0.008</td>
</tr>
<tr>
<td>LKD6</td>
<td>15</td>
<td>29.9</td>
<td>1.391</td>
<td>0.104</td>
<td>0.008</td>
</tr>
<tr>
<td>RYE1</td>
<td>21</td>
<td>39.6</td>
<td>1.398</td>
<td>0.113</td>
<td>0.008</td>
</tr>
<tr>
<td>BEN1</td>
<td>24</td>
<td>30.5</td>
<td>1.356</td>
<td>0.114</td>
<td>0.009</td>
</tr>
<tr>
<td>NWS1</td>
<td>31</td>
<td>35.0</td>
<td>1.374</td>
<td>0.113</td>
<td>0.008</td>
</tr>
<tr>
<td>NWS2</td>
<td>30</td>
<td>36.2</td>
<td>1.405</td>
<td>0.114</td>
<td>0.008</td>
</tr>
<tr>
<td>NWS3</td>
<td>27</td>
<td>30.2</td>
<td>1.335</td>
<td>0.102</td>
<td>0.008</td>
</tr>
<tr>
<td>SKY1</td>
<td>32</td>
<td>31.1</td>
<td>1.365</td>
<td>0.112</td>
<td>0.008</td>
</tr>
<tr>
<td>SKY2</td>
<td>32</td>
<td>32.8</td>
<td>1.376</td>
<td>0.113</td>
<td>0.008</td>
</tr>
<tr>
<td>UIS2</td>
<td>20</td>
<td>35.9</td>
<td>1.375</td>
<td>0.112</td>
<td>0.008</td>
</tr>
<tr>
<td>UIS4</td>
<td>16</td>
<td>33.0</td>
<td>1.342</td>
<td>0.110</td>
<td>0.008</td>
</tr>
<tr>
<td>SHT1</td>
<td>15</td>
<td>24.5</td>
<td>1.266</td>
<td>0.082</td>
<td>0.007</td>
</tr>
<tr>
<td>SHT3</td>
<td>15</td>
<td>25.1</td>
<td>1.286</td>
<td>0.085</td>
<td>0.008</td>
</tr>
<tr>
<td>SHT5</td>
<td>15</td>
<td>26.5</td>
<td>1.322</td>
<td>0.092</td>
<td>0.008</td>
</tr>
<tr>
<td>STK1</td>
<td>24</td>
<td>31.6</td>
<td>1.336</td>
<td>0.109</td>
<td>0.008</td>
</tr>
<tr>
<td>STK2</td>
<td>24</td>
<td>33.3</td>
<td>1.333</td>
<td>0.118</td>
<td>0.009</td>
</tr>
<tr>
<td>STK3</td>
<td>24</td>
<td>31.1</td>
<td>1.34</td>
<td>0.113</td>
<td>0.009</td>
</tr>
<tr>
<td>STK4</td>
<td>23</td>
<td>26.2</td>
<td>1.275</td>
<td>0.093</td>
<td>0.008</td>
</tr>
<tr>
<td>STK5</td>
<td>23</td>
<td>28.8</td>
<td>1.297</td>
<td>0.104</td>
<td>0.009</td>
</tr>
</tbody>
</table>
3.5.3 Genetic Relationships Among Populations

The UPGMA analysis revealed distinct clustering, dividing populations amongst four geographic regions; St Kilda, Shetland, North West Scotland and the Hebrides, and English mainland and reference cultivars, with the first separation formed of populations from St Kilda (Figure 3.10). There was no difference to population structure when assuming inbreeding in *T. repens*. The branch lengths among the St Kilda populations are longer than those separating other groups indicating a higher genetic distance between populations. The Isles of Scilly and mainland English wild populations and the two landraces, English Dutch and Kent Wild White cluster closely, with similar clustering shown between populations from the Scottish mainland and the Hebrides.

The model-based Bayesian clustering analysis based on the AFLP data conducted in the program *STRUCTURE* found Δ*K*=6 (Figure 3.11). This grouping divides the populations into six groups, with generally a very low, but consistent level of admixture between groups as shown in Table 3.16. In particular, the outer island groups of Scotland, Shetland and St Kilda, show particularly low levels of admixture with the other genetic population groups. Further, the Bayesian analysis suggests that hybridisation between the populations in Scotland and the historically cultivated species assessed in this study is absent. This level of Δ*K* closely follows the structure shown in the UPGMA clustering analysis. It must be stated here that the authors

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>Sum of squares</th>
<th>Variance components</th>
<th>Percentage of variation</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among regions</td>
<td>3</td>
<td>1465.301</td>
<td>3.019</td>
<td>11.75</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Among populations within regions</td>
<td>25</td>
<td>1760.287</td>
<td>2.424</td>
<td>9.43</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Within populations</td>
<td>573</td>
<td>11606.868</td>
<td>20.246</td>
<td>78.82</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Table 3.15 Analysis of molecular variance (AMOVA) of variation across 602 *T. repens* individuals conducted in *ARLEQUIN* (Schneider et al., 2000). Probability tested using 1000 permutations.
acknowledge that the underlying STRUCTURE model is not well suited to data where the individuals in question are under isolation by distance, as is the case in this study (Pritchard et al. 2007; Schwartz & McKelvey, 2009). Hence these results must be interpreted with caution.

Figure 3.10 Unrooted unweighted pair group method with arithmetic mean (UPGMA) dendrogram based on Nei’s genetic distances after Lynch and Milligan (1994). Values at the nodes represent 1000 bootstrap values shown as percentages; only values over 50% are shown.
Figure 3.11 STRUCTURE analysis of AFLP data for *T. repens* across the UK. Values of $\Delta K$ (Evanno et al., 2005) are plotted against $K=2$-10.

Table 3.16 Estimated membership coefficients for each population to each cluster as defined by STRUCTURE for $K=6$, based on an admixture model with correlated allele frequencies (Pritchard et al., 2000). Highest membership coefficients for each population are emboldened for emphasis.

<table>
<thead>
<tr>
<th>Population</th>
<th>1</th>
<th>2</th>
<th>Inferred cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>DEV1</td>
<td>0.821</td>
<td>0.129</td>
<td>0.009</td>
</tr>
<tr>
<td>DOR1</td>
<td>0.872</td>
<td>0.021</td>
<td>0.019</td>
</tr>
<tr>
<td>DOR2</td>
<td>0.956</td>
<td>0.017</td>
<td>0.008</td>
</tr>
<tr>
<td>DOR3</td>
<td>0.939</td>
<td>0.018</td>
<td>0.023</td>
</tr>
<tr>
<td>ED1</td>
<td>0.922</td>
<td>0.018</td>
<td>0.014</td>
</tr>
<tr>
<td>KWW1</td>
<td>0.916</td>
<td>0.05</td>
<td>0.009</td>
</tr>
<tr>
<td>RYE1</td>
<td>0.964</td>
<td>0.011</td>
<td>0.008</td>
</tr>
<tr>
<td>IOS1</td>
<td>0.036</td>
<td>0.932</td>
<td>0.009</td>
</tr>
<tr>
<td>IOS2</td>
<td>0.067</td>
<td>0.906</td>
<td>0.007</td>
</tr>
<tr>
<td>IOS3</td>
<td>0.034</td>
<td>0.937</td>
<td>0.01</td>
</tr>
<tr>
<td>LKD2</td>
<td>0.11</td>
<td>0.836</td>
<td>0.024</td>
</tr>
<tr>
<td>LKD4</td>
<td>0.035</td>
<td>0.829</td>
<td>0.052</td>
</tr>
<tr>
<td>LKD6</td>
<td>0.053</td>
<td>0.781</td>
<td>0.021</td>
</tr>
<tr>
<td>BEN1</td>
<td>0.007</td>
<td>0.006</td>
<td>0.906</td>
</tr>
<tr>
<td>UIS2</td>
<td>0.039</td>
<td>0.017</td>
<td>0.858</td>
</tr>
<tr>
<td>UIS4</td>
<td>0.008</td>
<td>0.012</td>
<td>0.906</td>
</tr>
<tr>
<td>NWS1</td>
<td>0.012</td>
<td>0.016</td>
<td>0.018</td>
</tr>
<tr>
<td>NWS2</td>
<td>0.086</td>
<td>0.053</td>
<td>0.015</td>
</tr>
<tr>
<td>NWS3</td>
<td>0.055</td>
<td>0.024</td>
<td>0.065</td>
</tr>
<tr>
<td>SKY1</td>
<td>0.008</td>
<td>0.015</td>
<td>0.13</td>
</tr>
<tr>
<td>SKY2</td>
<td>0.008</td>
<td>0.015</td>
<td>0.082</td>
</tr>
<tr>
<td>SHT1</td>
<td>0.004</td>
<td>0.007</td>
<td>0.011</td>
</tr>
<tr>
<td>SHT3</td>
<td>0.007</td>
<td>0.021</td>
<td>0.02</td>
</tr>
<tr>
<td>SHT5</td>
<td>0.034</td>
<td>0.011</td>
<td>0.041</td>
</tr>
<tr>
<td>STK1</td>
<td>0.01</td>
<td>0.063</td>
<td>0.014</td>
</tr>
<tr>
<td>STK2</td>
<td>0.014</td>
<td>0.005</td>
<td>0.012</td>
</tr>
<tr>
<td>STK3</td>
<td>0.009</td>
<td>0.007</td>
<td>0.015</td>
</tr>
<tr>
<td>STK4</td>
<td>0.003</td>
<td>0.003</td>
<td>0.006</td>
</tr>
<tr>
<td>STK5</td>
<td>0.005</td>
<td>0.005</td>
<td>0.115</td>
</tr>
</tbody>
</table>
3.5.4 ISOLATION BY DISTANCE

A significant positive relationship was observed between pairwise genetic distance and geographic distance over total population samples ($r = 0.416$, $P<0.001$), excluding the two landraces. Significant isolation by distance was detected amongst all mainland UK populations ($r = 0.804$, $P<0.001$) and also amongst populations of the North West Scotland region ($r = 0.876$, $P<0.001$) (see Figure 3.12). No isolation by distance was detected between populations on Shetland or St Kilda (graphs not shown).

Figure 3.12 Mantel tests of isolation by distance, plotting transformed $F_{ST}$ (Rousset, 1997) against log transformed distance values; a all populations b all mainland populations (including Skye collections).
Spatial autocorrelation analysis produced a negative correlation with average kinship coefficient decreasing over increasing distance (Figure 3.13). Significant kinship coefficients were identified between individuals less than 165.01 km apart, suggesting that white clover may be able to form a related group up to this distance. Clearly these distances reflect the sampling strategy used and indicate only a preliminary genetic patch size, with further analysis on a continuously sampled population required for a more accurate representation of kinship.

![Figure 3.13 Pairwise average kinship coefficient of 602 T. repens individuals plotted against geographic distance (km) using SPAGEDi (Hardy & Vekemans 2002); * indicates significant deviation from random mating among individuals (P<0.05).](image)

3.5.5 ENVIRONMENTAL IMPACTS ON PATTERNS OF GENETIC DIVERSITY

Redundancy based analysis confirmed the importance of spatial distance on genetic diversity in *T. repens* with latitude and longitude explaining 64 and 69% of the variation in *F*<sub>ST</sub> and Nei’s genetic diversity (Table 3.17).

In addition some climatic factors appear to explain a proportion of variation in genetic diversity. Accounting for variation explained by spatial distance (conditional tests), rainfall
becomes non significant, perhaps due to its orthogonal relationship to spatial distance. Temperature, grass minimum temperature and days of snow lying remain significant, indicating that these variables may be important in determining genetic dissimilarities between populations. By fitting spatial distance as covariables, altitude and the distinction of a water barrier become significant factors in determining a small proportion of genetic variation between populations of *T. repens*.

**Table 3.17** Redundancy based analysis of the effect of environmental factors on genetic differentiation in *T. repens* populations. All marginal tests are shown on the left, with conditional tests on the right (including spatial distance as covariables in the analysis). Values in bold represent significant $P$ values (<0.1). $r^2$ values represent the proportion of variation explained by each environmental variable.

<table>
<thead>
<tr>
<th>Variable set</th>
<th>Marginal tests</th>
<th>Conditional tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Variable set</td>
<td>F</td>
</tr>
<tr>
<td>Nei’s genetic distance</td>
<td>Distance</td>
<td>26.077</td>
</tr>
<tr>
<td></td>
<td>Air frost</td>
<td>1.423</td>
</tr>
<tr>
<td></td>
<td>Altitude</td>
<td>0.625</td>
</tr>
<tr>
<td></td>
<td>Grass</td>
<td>5.349</td>
</tr>
<tr>
<td></td>
<td>Island</td>
<td>3.235</td>
</tr>
<tr>
<td></td>
<td>Rainfall</td>
<td>4.144</td>
</tr>
<tr>
<td></td>
<td>Snow</td>
<td>8.429</td>
</tr>
<tr>
<td></td>
<td>Sun hours</td>
<td>8.021</td>
</tr>
<tr>
<td></td>
<td>Temperature</td>
<td>3.018</td>
</tr>
</tbody>
</table>

$F_{ST}$

<table>
<thead>
<tr>
<th>Variable set</th>
<th>Marginal tests</th>
<th>Conditional tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Variable set</td>
<td>F</td>
</tr>
<tr>
<td>Distance</td>
<td>20.889</td>
<td>0.000</td>
</tr>
<tr>
<td>Air frost</td>
<td>1.439</td>
<td>0.217</td>
</tr>
<tr>
<td>Altitude</td>
<td>0.753</td>
<td>0.498</td>
</tr>
<tr>
<td>Grass</td>
<td>5.046</td>
<td>0.013</td>
</tr>
<tr>
<td>Island</td>
<td>3.129</td>
<td>0.044</td>
</tr>
<tr>
<td>Rainfall</td>
<td>4.115</td>
<td>0.020</td>
</tr>
<tr>
<td>Snow</td>
<td>7.249</td>
<td>0.001</td>
</tr>
<tr>
<td>Sun hours</td>
<td>7.212</td>
<td>0.004</td>
</tr>
<tr>
<td>Temperature</td>
<td>3.075</td>
<td>0.030</td>
</tr>
</tbody>
</table>
3.5.6 ISLAND VERSUS MAINLAND HETEROZYGOSITY

Contrary to other studies of island versus mainland populations, *T. repens* is shown to maintain higher expected heterozygosity levels in island populations in comparison to average expected heterozygosity in mainland populations (Table 3.18). However, when accounting for differences in population size using the rarefaction method for allelic frequency mainland populations of *T. repens* contain higher levels of allelic richness than island populations.

Table 3.18 Island versus mainland heterozygosity following the method outlined by Frankham (1997) for within population heterozygosity in *T. repens*. *a* On the left are the number of occasions where mainland population expected heterozygosity/allelic richness is greater than island populations, on the right are occasions when island expected heterozygosity/allelic richness is greater than mainland populations.

<table>
<thead>
<tr>
<th>Island population</th>
<th>$H_j$</th>
<th>$M &gt; Is$: $M &lt; Is^a$</th>
<th>$A$</th>
<th>$M &gt; Is$: $M &lt; Is^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEN1</td>
<td>0.114</td>
<td>0:1</td>
<td>1.356</td>
<td>1:0</td>
</tr>
<tr>
<td>IOS1</td>
<td>0.111</td>
<td>0:1</td>
<td>1.385</td>
<td>0:1</td>
</tr>
<tr>
<td>IOS2</td>
<td>0.114</td>
<td>0:1</td>
<td>1.452</td>
<td>0:1</td>
</tr>
<tr>
<td>IOS3</td>
<td>0.113</td>
<td>0:1</td>
<td>1.388</td>
<td>0:1</td>
</tr>
<tr>
<td>SHT1</td>
<td>0.082</td>
<td>1:0</td>
<td>1.266</td>
<td>1:0</td>
</tr>
<tr>
<td>SHT3</td>
<td>0.086</td>
<td>1:0</td>
<td>1.286</td>
<td>1:0</td>
</tr>
<tr>
<td>SHT5</td>
<td>0.092</td>
<td>1:0</td>
<td>1.322</td>
<td>1:0</td>
</tr>
<tr>
<td>STK1</td>
<td>0.109</td>
<td>0:1</td>
<td>1.336</td>
<td>1:0</td>
</tr>
<tr>
<td>STK2</td>
<td>0.118</td>
<td>0:1</td>
<td>1.333</td>
<td>1:0</td>
</tr>
<tr>
<td>STK3</td>
<td>0.113</td>
<td>0:1</td>
<td>1.340</td>
<td>1:0</td>
</tr>
<tr>
<td>STK4</td>
<td>0.093</td>
<td>1:0</td>
<td>1.275</td>
<td>1:0</td>
</tr>
<tr>
<td>STK5</td>
<td>0.104</td>
<td>1:0</td>
<td>1.297</td>
<td>1:0</td>
</tr>
<tr>
<td>UIS2</td>
<td>0.112</td>
<td>0:1</td>
<td>1.375</td>
<td>1:0</td>
</tr>
<tr>
<td>UIS4</td>
<td>0.110</td>
<td>0:1</td>
<td>1.342</td>
<td>1:0</td>
</tr>
<tr>
<td>Total</td>
<td>5:9</td>
<td></td>
<td>11:3</td>
<td></td>
</tr>
</tbody>
</table>
3.6 DISCUSSION

3.6.1 THE SPATIAL PATTERN OF DIVERSITY IN *TRIFOLIUM* SPECIES ACROSS THE UK

Comprehensive reviews of allozyme diversity identify the link between the distribution of genetic diversity and life history traits of species (Gottlieb, 1977; Hamrick *et al*., 1979; Hamrick & Godt, 1989, 1996). Both breeding system (defined as selfing, mixed mating or outcrossing), and life form (annual, short-lived perennial or long-lived perennial) account for 46% of the variation in population genetic diversity explained by life history traits (Hamrick & Godt, 1989). Inbreeding populations are in general homozygous, with complete inbreeding reducing the effective population size (*N*e) by half (Pollak, 1987). Thus, through smaller *N*e, and therefore greater susceptibility to drift and reduced gene flow, inbreeding species are predicted to show higher levels of genetic variation among populations and lower heterozygosity within populations than outcrossing species (Wright, 1951; Allard *et al*., 1968; Loveless & Hamrick, 1984). Differences in life form are thought to influence the partitioning in genetic variation as perennial species are less susceptible to the effects of drift and will lose genetic variation more slowly than short-lived species. Therefore longer-lived species are more likely to have higher levels of heterozygosity within populations and lower population divergence (Antonovics, 1968; Brown, 1979; Loveless & Hamrick, 1984). Nonetheless, in spite of its predominantly inbreeding nature, *T. dubium* was found to contain relatively high levels of diversity within some populations compared to the mixed mating and outbreeding species assessed in the study. This inconsistency may be in part due to the presence of differing life forms between the two species studied, as correlations between single traits and their associated effect on genetic diversity become complicated in the presence of combinations of several different life history characteristics (Hamrick & Godt, 1996). Indeed,
in two trait analyses of within species diversity, measures of diversity due to in- and outbreeding were shown to be affected by life form, with annual inbreeding species having equivalent levels of within species variation as outbreeding short-lived perennials and mixed-mating long-lived perennials (Hamrick & Godt, 1996). Hence the high levels of diversity within some populations of *T. dubium* in comparison to *T. pratense* and *T. repens* may be attributable to both its annual life form and potentially a low level of outbreeding in natural populations. It should be noted here that these levels of expected heterozygosity are only high in comparison to the other species assessed in this study, as Nybom (2004), in a study of dominant marker data finds that the average value reported for $H_w$ is 0.23, as such the values reported for all species are lower than could be expected.

The highest level of expected heterozygosity is seen in only the southern populations of *T. dubium* included in this study, with northern populations containing lower within population diversity. This high variability in *T. dubium* within population diversity levels compared to the relatively stable within population heterozygosity of *T. pratense* and *T. repens* is in accordance with other studies comparing among population diversity with differing breeding systems (Schoen & Brown, 1991). In addition, Schoen & Brown (1991) found that, in some species, increased within population variation correlated with greater effective population size ($N_e$) in inbreeding populations, with low $N_e$ corresponding to low levels of diversity. This effect is seen in the *T. dubium* populations assessed in this study as *T. dubium*, although distributed across the UK, is more restricted in its distribution in the north, where smaller populations were observed (see also Preston *et al.*, 2002). These smaller populations were found to have both lower heterozygosity and allelic richness within populations than in southern areas where the species was found in greater numbers. This effect is most pronounced in northern islands of the UK, with the Outer Hebrides having the lowest within population diversity. The extremely low within population heterozygosity in
this population compared to the rest of those sampled suggests that this population could be
derived from a single founder event.

The spatial pattern of diversity in *T. dubium* conforms to the „abundant centre’ model
which dictates that species are expected to have the highest abundance at the centre of their
ranges where optimum conditions occur for survival and reproduction (Brussard, 1984; Lesica
& Allendorf, 1995; Eckert *et al*., 2008). Conversely populations at the edge of ranges become
smaller and more isolated; often corresponding to a lower genetic diversity within
populations. Hence, it is suggested that the non-continuous distribution of the species in the
north and the pattern of increasing population diversity towards the south of the range may
indicate that the north of the UK is the edge of the range for *T. dubium* in the UK.
Alternatively, the potential edge of range observed here may be a consequence of glacial
retreat in the UK (Hampe & Petit, 2005). As such the southern populations survive as
genetically rich source populations that have given rise to a number of distinct marginal
populations in a species that has not yet reached stability following glaciations.

*T. pratense* and *T. repens* have a relatively consistent level of heterozygosity within
populations, as expected in more outbreeding species (Schoen & Brown, 1991). Variability
appears in the *T. pratense* dataset when including island populations, where allelic richness,
normalized to account for differences in population size, shows on average lower values on
island populations compared to mainland populations. In a widespread review of island versus
mainland heterozygosity, island populations were shown to contain lower genetic diversity
than mainland populations, resulting from founder effects and loss of diversity due to finite
population size (Frankham, 1997). Only populations collected from the Isles of Scilly show
higher diversity within populations than mainland populations for all three species, suggesting
that populations on the Isles of Scilly may have sufficient population size and gene flow from
the mainland to maintain levels of diversity. Island populations are shown to be more
diverged from their mainland counterparts in both *T. pratense* and *T. repens* due to their isolation, genetic drift and natural selection. However, northern islands are more diverged from their mainland counterparts than the southern island populations included in this study. This may support the previous assessment that gene flow between mainland and the Isles of Scilly has been more prevalent than in northern islands.

Limited gene dispersal in inbreeding populations is likely to provide larger differentiation between populations than in outbreeding or mixed mating populations (Brown, 1978; Hamrick & Godt, 1989; Schoen & Brown, 1991). Hence as expected population differentiation, measured by $F_{ST}$, was greater in *T. dubium* than in *T. pratense* and *T. repens*. However the values given here are lower overall than those identified by Nybom (2004) in a review of $G_{ST}$ values (an analogue of $F_{ST}$) from dominant marker studies. Significant isolation by distance was confirmed in the species studied, both including and excluding island populations. Redundancy analysis confirms that geographic distance is the largest determinant of genetic distance for all species. Geographic distance is a clear obstacle to gene dispersal, but geographic distance may not be the sole cause of genetic differentiation and divergence between populations. The populations assessed here are located on a latitudinal gradient with associated climatic regime changes, shown in *T. pratense* and *T. repens* to be factors in determining spatial patterns of diversity, even when controlling for spatial distance. The only factor common to all species shown to have a significant impact on genetic diversity patterns outside of geographic diversity is whether the population is present on an island, and thus is separated by a significant water barrier, highlighting the importance of isolated populations to overall genetic diversity in species.
3.6.2  *T. repens* – GENETIC EXCHANGE IN A CROP-WILD SPECIES COMPLEX

These results emphasise the unique genetic diversity of *T. repens* contained within the more remote (Shetland and St Kilda) UK islands. The moderate level of population subdivision and low genetic distance found between populations of *T. repens*, in conjunction with the widespread nature of the sampling sites, serves to suggest that overall there is a high homogenisation of the gene pool within the UK, especially in those populations within close proximity of each other. Analyses of isolation by distance (IBD) serve to show a highly significant IBD detected across all populations, across the North West Scotland region and across mainland UK populations. Vicariance, in relation to any geographic barrier is also more likely to be detected over increasing geographic distance, so the observed pattern of IBD is likely to be both a function of distance and geographic barriers. Nonetheless this observed IBD indicates that genetic differentiation is associated with geographic distance across the range of *T. repens* within the UK. Consequently fragmentation and ultimately geographic isolation can maintain levels of differentiation across the range of a native, widespread species.

This fragmentation is reflected by the Bayesian clustering analyses where populations are split into geographic groups, as well as in the dendrogram topology where populations are clearly separated into four „regions”: St Kilda, Shetland, North West Scotland and the Hebrides, and an English mainland grouping including the two reference landraces. The clustering of populations from North West Scotland suggests that, in the absence of long distance dispersal in this species over sea barriers, anthropomorphic dispersal will have contributed to the similarity between the Outer Hebrides and Scottish mainland populations. The absence of admixture between populations from Scotland and the cultivated forms used in this study suggests that the consequences of past crop-wild hybridization have been minimal or non-existent outside of the southern UK. It should be noted here that this
assessment is based on two landraces and further experimentation using a wider sample of cultivated species is needed to clarify this issue. The clustering of historically widely cultivated varieties with southern English mainland wild populations indicates their close genetic proximity, and points to gene flow from these crop varieties to wild species, although ancestral similarity between the populations as a result of common ancestry cannot be discounted. Genetic similarity through common ancestry cannot be easily differentiated from gene flow using AFLP analysis; however it does seem unlikely that crop-wild gene flow will not have occurred where native and cultivated forms occur in a sympatric distribution. Measurements of pollen flow in *T. repens* have shown that, in accordance with the IBD results, it is reduced over longer distances (Osborne *et al.*, 2000). However, while bees regularly forage over small distances, honey bees have been found to forage over distances greater than 9.5 km when differences in forage patch size and quality are large (Beekman & Ratnieks, 2000). Thus habitat fragmentation and loss of unimproved grassland in the UK (Fuller, 1987) may lead to an increase in longer distance gene flow amongst bee-pollinated plants such as *T. repens*. In relation to long distance dispersal, both large herbivores and migratory birds are thought to be involved in *T. repens* seed dispersal (Williams, 1987), with *T. repens* seed identified as being carried by larger herbivores both in epi- and endozoochory (Couvreur *et al.*, 2004, 2005). Nonetheless, human influence and transportation is surely a significant factor for seed dispersal in this cultivated species (Nathan *et al.*, 2008).

St Kilda maintains a lower level of allelic richness (when accounting for sample sizes) than mainland populations as would be expected in an island population (MacArthur & Wilson 1967; Frankham, 1997). However St Kilda provides a particularly interesting example of an island flora as the low level of genetic differentiation amongst populations in close geographic proximity is not reproduced within St Kilda, which maintains relatively high distinctiveness between populations irrespective of their close geographic relationship. The
distinction of the north Scottish island populations of Shetland and St Kilda indicates that in such a widespread and commonly cultivated "homogeneous" native species as *T. repens*, isolated areas in the UK can harbour reservoirs of substantial extant variation that may have once existed, but no longer can be found in mainland UK.

3.6.3 **Priority areas for conservation**

Priority populations for conservation can be defined in terms of their value to biodiversity, through higher levels of diversity, divergence and isolation from threats that threaten diversity. In terms of levels of diversity, the Isles of Scilly and mainland UK populations contain the highest diversity across the UK, with northern island populations typically containing lower levels of variation than mainland populations. However, does it necessarily follow that these low diversity populations should be given a lower priority in terms of conservation? Whether peripheral populations of lower diversity should be given precedence in terms of conservation has been the subject of considerable debate, particularly when the populations are peripheral in relation to a political unit while remaining globally common (Hunter & Hutchinson, 1994; Lesica & Allendorf, 1995; Eckert *et al*., 2008). It is acknowledged that peripheral populations of plants, while often containing lower levels of genetic variation than central populations, can be important for conservation efforts in spite of their lower population size and frequency, through their value to overall diversity in terms of divergence and isolation. Unique diversity inherent in isolated populations may be important for species adaptation to environmental change, assuming that the level of variation observed is reflected in quantitative traits (Hunter & Hutchinson, 1994; Eckert *et al*., 2008). In addition, many authors consider isolated and peripheral populations to be the progenitors of speciation events contributing to the generation of biological diversity (Mayr, 1954; Levin, 1970, 1993). Hence, while island populations in this study show in general a lower diversity than mainland
populations, the importance of isolated island populations for conservation lies in their divergence and isolation from mainland populations, and their consequent potential for containing, and maintaining, unique diversity.

This study has shown that *T. repens* has been able to maintain some level of genetic differentiation across its native range irrespective of widespread cultivation of the species since the 17th century (Caradus, 1995), however in England the relationship between geographic and genetic differentiation is less pronounced, in the authors opinion due to the likely homogenising effect of gene flow between cultivated germplasm and wild populations. As yet, the more isolated areas of the UK in north western Scotland remain differentiated from southern populations and the cultivated variants assessed in this study. With no direct management of *T. repens* genetic diversity it is unlikely that this will continue. Widespread use of cultivated forms in outbreeding and mixed-mating species that are closely genetically related to forms found in other parts of the country highlights the difficulties faced when defining conservation areas (Greene *et al.*, 2008). In terms of conservation strategy, this would suggest that where native species commonly occur in close proximity to a con-specific crop, priority conservation sites for genetic diversity should be allocated to areas where little or no cultivation has occurred and may not in the future. The most remote of the UK islands in this study, St Kilda, holds the most unique genetic diversity in this species, and due to its isolation both through natural barriers and from human influence will be the most likely to be able to retain its diversity in the future.

For *T. pratense* the potential major threat to wild diversity, like *T. repens*, comes from cultivation and the associated effects of genetic swamping and homogenization of the gene pool. Consequently, as for *T. repens*, it follows that the level of isolation of populations from these threats is an important determinant of conservation priority in this species. *T. pratense* maintains distinct island populations on Shetland and in some populations in the Outer
Hebrides compared to mainland populations. However, the genetic proximity of one population on the Outer Hebrides to populations from North West Scotland indicates the potential for gene flow, potentially through the re-seeding of areas in the Outer Hebrides. While sites in Shetland and protected sites in the Outer Hebrides could be designated as priority areas in terms of their divergence from mainland populations, other complementary areas will need to be included to maintain total genetic diversity. Both the presence of isolation by distance and the high diversity in the south suggests appropriate complementary areas should be situated in the south of the UK. The Isles of Scilly is a notable area for conservation in the south, as, while divergence is low, populations here maintain some of the highest diversity within populations. For *T. dubium*, as a wild species, both diversity and divergence levels become of more importance when assessing conservation priority; hence conservation precedence would necessarily focus on the genetically diverse southern UK populations. However, it is suggested that an appropriate strategy should include some of the populations located on its range edge, potentially vital in terms of adaption to environmental change.

Defining conservation priorities and attaching any measure of value to populations is always problematic, balancing the integration of different fields of biology with limited resources, in addition to social concerns. Despite this, in any attempt to identify populations that deserve priority, this study highlights the importance of baseline genetic diversity assessments, with all species assessed in this study requiring differing strategies in response to their genetic diversity assessments, life form and cultivation history. Whether isolated populations should be given priority in the face of restricted funds is still contentious, however this study suggests that island populations provide an important opportunity for conservation, both protecting populations from threats that endanger diversity and allowing
populations to diverge from their mainland counterparts, providing the building blocks for future diversity.
Chapter 4. **Applying SNPs to Conservation Questions; A Case Study of *T. pratense* in the UK**

4.1 **Introduction**

The expansion of modern agriculture, where genetic uniformity has overtaken numerous diverse local varieties, has led to the erosion of the diversity in crop gene pools. With low levels of genetic diversity it is likely that crop yields will struggle to maintain current levels due the likely effects of a changing climate, a decline in pollinators from habitat fragmentation and the continual adaptation of pests (Ehrlich, 1988). This known vulnerability in our essential plant resources has underlined the need to identify and conserve progenitors of crop plants, their wild relatives and traditional landraces, the conservation of which will help to maintain a diverse gene pool to allow future breeding programs to adapt to the changing needs of farmers, consumers and the environment.

A wide range of markers are available to study the genetic variation in wild species and landraces. Molecular markers have evolved since the first true molecular marker, allozymes, which measure the variation in enzymes, to markers that directly assess DNA variation. These markers include those that make use of restriction enzymes to amplify regions of the genome to assess variation such as restriction fragment polymorphisms (Botstein *et al*., 1980) and amplified fragment length polymorphisms (Vos *et al*., 1995), to microsatellites (Tautz, 1989; Weber & May, 1989) which measure the variation in repeat regions in the genome. Known limitations of these markers, which can limit further research development have highlighted the need to produce markers than have the potential to advance population genetic analyses (Zhang & Hewitt, 2003). Following the increased sequencing efforts in crop plants, particularly in producing expressed sequence tag (EST) databases, it has become possible to identify and assess polymorphisms directly, at the nucleotide level, using single nucleotide
polymorphisms or SNPs. Random neutral markers, both SNPs assessed in non-functional regions of the genome and more traditional markers, are of great importance to population genetics and evolutionary studies (Syvänen, 2001), allowing assessments on the impacts of gene flow, inbreeding and genetic drift (Ouborg et al., 2010a). However there are questions as to whether these neutral genetic markers can reliably reflect the differences in the underlying adaptive traits (for reviews see Hedrick, 2001; Reed & Frankham, 2001). Van Tienderen et al. (2002) note that there may be only a limited set of genes that enable an ecotype or species to survive in a particular niche, so the variation in neutral markers may not adequately reflect the variation in the traits of interest. This has signalled a move away from random genetic markers, towards more „functional markers”, including SNPs within coding regions of a candidate gene, that can be causally related to a phenotypic trait. This movement is of considerable importance to the development of conservation genetics, directly examining the link between variation and the quantitative traits that are likely to be required to ensure future persistence of the gene pool. With the conservation questions to date often focusing on the effects of habitat fragmentation, the need for functional markers is even more relevant in light of the recognition of more, or greater, threats posed by environmental degradation, climate change and the deterioration of species-species interactions (Ouborg et al., 2010b).

Andersen & Lübberstedt (2003) separate SNPs into gene-targeted and functional markers, where functional markers (FMs) define those where their affect on a phenotypic trait is experimentally determined, through association or mutant studies. Determining FMs is therefore an expensive and timely task, finding the trait of interest, identifying the traits involved and developing markers within or flanking the genes. Thus while FMs provide unequivocal evidence of how diversity affects traits of interest, for conservation questions this may not be a feasible method for diversity studies where financial and time constraints are
often limiting. However, gene-targeted markers (GTMs), those that are derived from polymorphisms within genes, provide markers within a candidate gene and thus can provide more information on adaption than more randomly assigned markers, although this relationship is more ambiguous than that defined by FMs. Patterns of selection in the data derived from GTMs can be assessed following genotyping using a number of different statistical techniques to filter out the imprints of selection from random patterns in the data (Schlötterer 2002; van Tienderen et al. 2002; Oleksyk et al., 2010).

In this chapter GTMs are used to assess spatial genetic variation in a wild *Trifolium* species, *T. pratense*, to identify priority areas for conservation in the UK and surrounding islands. *T. pratense* is one of the most economically important forage crop species in the northern hemisphere and is widely cultivated throughout (Kölliker et al., 2003). To my knowledge this is the first study using SNPs as GTMs to assess variation in wild populations of *Trifolium*. This study aims to 1) generate GTMs in *T. pratense*; 2) evaluate the genetic diversity in *T. pratense* on the basis of these markers 3) assess the feasibility of using GTMs to assess the diversity in wild species such as *T. pratense*. 

4.2 METHODS

The study into adaptive variation in *T. pratense* was carried out at the University of Birmingham, UK and Aberystwyth University, Gogerddan Campus, UK, formerly the Institute of Grassland and Environmental Research (IGER).

4.2.1 CANDIDATE GENE SELECTION

To assess potential adaptive diversity in *T. pratense* nine gene loci were selected for further analysis, representing genes associated with drought-stress. From a preliminary survey conducted at IGER by Leif Skøt to find stress related genes in published *T. pratense* EST sequences, a total of 20 sequences were found to be of interest (see Table 4.1). Reference numbers for these 20 are reported as received from IGER. These consist of both Genbank accession numbers (those beginning BB) and TIGR Plant Transcript Assemblies accession numbers (those beginning TA) which describe large assembled transcripts of multiple ESTs (Appendix 4 contains further information on transcript assemblies used in this study). So as not to bias statistical analysis through unequal loci lengths, in addition to the increased likelihood of sequencing error with longer sequences, individual EST sequences were selected from each assembly for further analysis.

Each locus received from IGER was compared with the available database of sequence information using BLAST (Basic Local Alignment Search Tool) (Zhang *et al.*, 2000) to confirm the putative function of each locus, in conjunction with a review of the published information on each locus. The number of potential candidate loci was reduced to nine by confining candidate genes for further study to genes specifically related to drought stress (see Table 4.1).
Table 4.1 Blast search of nucleotide database (nr/nt) for candidate genes for putative *T. pratense* drought stress associated genes conducted on 28/6/09 using BLAST v2.2.21 (Zhang *et al*., 2000). Underlined EST references indicate the nine loci selected for amplification. Bold and underlined EST references indicate the five loci used for further analysis. EST reference numbers relate to Genbank (BB) and TIGR transcript assembly database (TA) accession numbers. a The number of alignments expected by chance (a reflection of the size of the database and scoring system used); b Indication of the proportion of the alignment that contain identical nucleotide pairs; c The percent of the query sequence that is matched by the aligned segments. d Obtained using Megablast; an algorithm optimised for highly similar sequences; e Obtained using blastn; an algorithm optimised for somewhat similar sequences.

<table>
<thead>
<tr>
<th>T. pratense EST reference</th>
<th>Template gene</th>
<th>Species</th>
<th>Function</th>
<th>BLAST ref</th>
<th>E value&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Max. ident&lt;sup&gt;b&lt;/sup&gt; (%)</th>
<th>Query coverage&lt;sup&gt;c&lt;/sup&gt; (%)</th>
<th>Direct function in drought stress</th>
<th>Potential function in drought stress</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB910055</td>
<td><em>Medicago sativa</em></td>
<td>Sucrose-phosphate synthase mRNA</td>
<td>AF322116</td>
<td>0.0&lt;sup&gt;d&lt;/sup&gt;</td>
<td>92</td>
<td>100</td>
<td>Preferential partitioning of carbon to sucrose occurs under osmotic stress</td>
<td></td>
<td></td>
<td>Quick <em>et al</em>., 1989; Zrenner &amp; Stitt, 1991</td>
</tr>
<tr>
<td>BB914596</td>
<td><em>Arabidopsis thaliana</em></td>
<td>SIP3 (SOS3-interacting protein 3)</td>
<td>NM119244</td>
<td>2e-93&lt;sup&gt;d&lt;/sup&gt;</td>
<td>80</td>
<td>81</td>
<td>Interacts with SOS3, a Ca2+ sensor involved in salt stress</td>
<td></td>
<td></td>
<td>Halfter <em>et al</em>., 2000</td>
</tr>
<tr>
<td>BB914880</td>
<td><em>Arabidopsis thaliana</em></td>
<td>C2H2 zinc-finger protein SERRATE (SE) mRNA</td>
<td>AF311221</td>
<td>2e-70&lt;sup&gt;e&lt;/sup&gt;</td>
<td>79</td>
<td>55</td>
<td>Some zinc-finger proteins have roles in drought tolerance</td>
<td></td>
<td></td>
<td>Prigge &amp; Wagner 2001; Huang &amp; Zhang, 2007; Xu <em>et al</em>., 2008;</td>
</tr>
<tr>
<td>BB920885</td>
<td><em>Medicago sativa</em></td>
<td>mRNA for heat shock protein (HSP)</td>
<td>X58711</td>
<td>1e-174&lt;sup&gt;d&lt;/sup&gt;</td>
<td>86</td>
<td>96</td>
<td>Transcription of HSPs induced by osmotic shock</td>
<td></td>
<td></td>
<td>Gyorgyey <em>et al</em>., 1991</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th><strong>T. pratense EST reference</strong></th>
<th><strong>Template gene</strong></th>
<th><strong>Species</strong></th>
<th><strong>Function</strong></th>
<th><strong>BLAST ref</strong></th>
<th><strong>E value</strong></th>
<th><strong>Max ident</strong></th>
<th><strong>Query coverage</strong></th>
<th><strong>Direct function in drought stress</strong></th>
<th><strong>Potential function in drought stress</strong></th>
<th><strong>References</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>BB922071</td>
<td>LjM3Kalpha mRNA for mitogen-activated kinase kinase kinase alpha</td>
<td><em>Lotus japonicus</em></td>
<td></td>
<td>AB167408</td>
<td>6e-93&lt;sup&gt;a&lt;/sup&gt;</td>
<td>78</td>
<td>96</td>
<td>Some MaPKKK genes have roles in stress signalling</td>
<td></td>
<td>Kim <em>et al.</em>, 2003; Kinoshita <em>et al.</em>, 2004; Nakagami <em>et al.</em>, 2004</td>
</tr>
<tr>
<td>BB926818</td>
<td>Heat shock transcription factor (HSFA)</td>
<td><em>Pisum sativum</em></td>
<td></td>
<td>AJ010643</td>
<td>7e-162&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>86</td>
<td>98</td>
<td>Some HSFs have roles in drought signalling</td>
<td></td>
<td>Sakuma <em>et al.</em>, 2006; von Koskull-Doering <em>et al.</em>, 2007</td>
</tr>
<tr>
<td>TA1010_57577 (BB924456)</td>
<td>PsEXT8 mRNA for xyloglucan endo-transglucosylase</td>
<td><em>Pisum sativum</em></td>
<td></td>
<td>AB270623</td>
<td>4e-115&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>89</td>
<td>58</td>
<td>Over-expression of XTHs increase drought tolerance</td>
<td></td>
<td>Bacon, 1999; Cosgrove, 2005; Cho <em>et al.</em>, 2006;</td>
</tr>
<tr>
<td>TA1106_57577</td>
<td>LBD39 (Lateral organ domain containing protein 39) mRNA</td>
<td><em>Arabidopsis thaliana</em></td>
<td></td>
<td>NM119918</td>
<td>9e-78&lt;sup&gt;e&lt;/sup&gt;</td>
<td>81</td>
<td>28</td>
<td>No known role in stress response</td>
<td></td>
<td>Shuai <em>et al.</em>, 2002</td>
</tr>
<tr>
<td>T. pratense EST reference</td>
<td>Template gene</td>
<td>Species</td>
<td>Function</td>
<td>BLAST ref</td>
<td>E value</td>
<td>Max ident(^b) (%)</td>
<td>Query coverage(^c) (%)</td>
<td>Direct function in drought stress</td>
<td>Potential function in drought stress</td>
<td>References</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------</td>
<td>---------</td>
<td>----------</td>
<td>-----------</td>
<td>---------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>-------------------------------</td>
<td>----------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>TA1548_57577 (BB915621)</td>
<td><em>Lycopersicon esculentum</em></td>
<td>Ethylene response factor 4 (ERF4) mRNA</td>
<td>AY192370</td>
<td>1e-34(^e)</td>
<td>77</td>
<td>37</td>
<td>ERF4 is induced by drought stress</td>
<td></td>
<td>Fujimoto <em>et al.</em>, 2000; Tournier <em>et al.</em>, 2003; Yang <em>et al.</em>, 2005</td>
<td></td>
</tr>
<tr>
<td>TA3611_57577</td>
<td><em>Glycine max</em></td>
<td>GmFAD2-2a gene for microsomal omega-6 fatty acid desaturase</td>
<td>AB188252</td>
<td>0.0(^d)</td>
<td>86</td>
<td>90</td>
<td></td>
<td></td>
<td>Gimeno <em>et al.</em>, 2009</td>
<td></td>
</tr>
<tr>
<td>TA3695_57577 (BB916074)</td>
<td><em>Populus x Canadensis</em></td>
<td>mRNA for osmosensor histidine-aspartate kinase (hk1 gene)</td>
<td>AJ937747</td>
<td>2e-58(^e)</td>
<td>69</td>
<td>99</td>
<td>HK1 functions as a sensor to osmotic stress</td>
<td></td>
<td>Urao <em>et al.</em>, 1999; Chefdor <em>et al.</em>, 2006</td>
<td></td>
</tr>
<tr>
<td>T. pratense EST reference</td>
<td>Template gene</td>
<td>Species</td>
<td>Function</td>
<td>BLAST ref</td>
<td>E value&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Max ident&lt;sup&gt;b&lt;/sup&gt; (%)</td>
<td>Query coverage&lt;sup&gt;c&lt;/sup&gt; (%)</td>
<td>Direct function in drought stress</td>
<td>Potential function in drought stress</td>
<td>References</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------</td>
<td>---------</td>
<td>----------</td>
<td>-----------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>TA3981_57577</td>
<td><em>Medicago truncatula</em></td>
<td><em>Medicago truncatula</em></td>
<td>Respiratory burst oxidase 1 mRNA</td>
<td>AY821801</td>
<td>0.0&lt;sup&gt;d&lt;/sup&gt;</td>
<td>92</td>
<td>71</td>
<td>Reactive oxygen intermediates (ROIs) produced in response to both abiotic and biotic stresses</td>
<td>Mittler, 2002</td>
<td></td>
</tr>
<tr>
<td>TA555_57577 (BB926319)</td>
<td><em>Medicago truncatula</em></td>
<td><em>Medicago truncatula</em></td>
<td>Aquaporin protein PIP1;1 mRNA</td>
<td>AF386739</td>
<td>0.0&lt;sup&gt;d&lt;/sup&gt;</td>
<td>86</td>
<td>72</td>
<td>Overexpression of aquaporins can increase/decrease sensitivity to drought stress</td>
<td>Aharon et al., 2003; Yu et al., 2005; Aroca et al., 2006</td>
<td></td>
</tr>
<tr>
<td>TA572_57577</td>
<td><em>Citrus sinensis</em></td>
<td><em>Citrus sinensis</em></td>
<td>psaDa mRNA for PSI reaction center subunit II</td>
<td>AF322116</td>
<td>3e-127&lt;sup&gt;e&lt;/sup&gt;</td>
<td>83</td>
<td>51</td>
<td>No known role in drought stress response</td>
<td>Giardi et al., 1996; Kohzuma et al., 2009</td>
<td></td>
</tr>
<tr>
<td>TA613_57577</td>
<td><em>Glycine max</em></td>
<td><em>Glycine max</em></td>
<td>mRNA for peroxisomal ascorbate peroxidase</td>
<td>NM119244</td>
<td>2e-32&lt;sup&gt;d&lt;/sup&gt;</td>
<td>86</td>
<td>23</td>
<td>Antioxidant function in response to stress including drought</td>
<td>Mittler &amp; Zilinskas, 1994; D'Arcy-Lameta et al., 2006; Arai et al., 2008;</td>
<td></td>
</tr>
<tr>
<td>TA989_57577 (BB906196)</td>
<td><em>Trifolium pratense</em></td>
<td><em>Trifolium pratense</em></td>
<td>RNA for putative transcription factor EREBP</td>
<td>AF311221</td>
<td>0.0&lt;sup&gt;d&lt;/sup&gt;</td>
<td>99</td>
<td>95</td>
<td>EREBP family implicated in the regulation of drought and cold tolerance genes</td>
<td>Kizis et al., 2001; Isobe et al., 2003</td>
<td></td>
</tr>
</tbody>
</table>
4.2.2 Leaf material and DNA extraction

Leaf material for *T. pratense* was collected and DNA extracted as detailed in Chapter 3. From the 15 individuals per population analysed in the previous AFLP study, five individuals per population were randomly selected using Microsoft Excel for further analysis.

4.2.3 Primer design

PRIMER3 version 0.4.0 (Rozen & Skaletsky, 2000) was used to design primers for each locus. Primer pair design details are given for each locus in Appendix 5. All primers were ordered from Eurofins MWG Operon and resuspended in 1 x TE buffer to a final volume of 100 pmol/µl, before storage at -20°C. Primers for each locus are given in Appendix 5.

4.2.4 Trial set

One individual was randomly selected from three dispersed populations, Isles of Scilly, the Lake District and Skye for PCR optimisation and sequencing trials of *T. pratense*. To avoid some of the problems associated with ascertainment bias no locus was excluded from further analysis, even in the absence of observed SNPs.

4.2.5 Locus amplification

Following trials varying DNA quantity, Mg$^{2+}$ concentration, thermocycling profiles and PCR adjuvants the PCR mix outlined in Table 4.2 was found to produce optimal and reliable results for all loci. The standard thermocycling profile consisted of an initial denaturation at 94°C for 3 min, followed by 35 cycles consisting of 94°C for 30 s, annealing for 1 min and 72°C for 1 min, with a final extension period of 5 min at 72°C. Annealing temperature was specific to each primer and is outlined in Table 4.3.
Table 4.2 Components of PCR mix to a final volume of 25µl per individual reaction.

<table>
<thead>
<tr>
<th>Reagents</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReddyMix™ PCR Master Mix (2x)</td>
<td>12.5µl</td>
</tr>
<tr>
<td>Forward primer</td>
<td>25pmol</td>
</tr>
<tr>
<td>Reverse primer</td>
<td>25pmol</td>
</tr>
<tr>
<td>DNA</td>
<td>10ng</td>
</tr>
<tr>
<td>BSA</td>
<td>0.8µg</td>
</tr>
<tr>
<td>SDW</td>
<td>To a final volume of 25µl</td>
</tr>
</tbody>
</table>

Amplification products were separated in 1.2% agarose gel (1 x TBE buffer). The desired fragments were excised from the gel and the DNA cleaned using QIAquick Gel Extraction Kit (Qiagen). The protocol outlined by Qiagen (2006) for gel extractions was followed, including the optional steps recommended for DNA that will be subjected to subsequent sequencing. In addition the ethanol wash stage was held for five minutes and repeated twice to ensure less contaminants remained in the DNA. See Appendices 1 and 2 for DNA extraction protocol and stock solution information.

Table 4.3 Annealing temperature for each locus after PCR optimisation: * n/a amplification was unsuccessful.

<table>
<thead>
<tr>
<th>EST reference</th>
<th>T. pratense</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB910055</td>
<td>61</td>
</tr>
<tr>
<td>BB920885</td>
<td>n/a</td>
</tr>
<tr>
<td>BB925882</td>
<td>60</td>
</tr>
<tr>
<td>TA1010_57577</td>
<td>n/a</td>
</tr>
<tr>
<td>TA1548_57577 (BB915621)</td>
<td>61</td>
</tr>
<tr>
<td>TA3078_57577</td>
<td>n/a</td>
</tr>
<tr>
<td>TA3695_57577 (BB916074)</td>
<td>61</td>
</tr>
<tr>
<td>TA555_57577</td>
<td>n/a</td>
</tr>
<tr>
<td>TA989_57577 (BB906196)</td>
<td>60</td>
</tr>
</tbody>
</table>

4.2.6 SEQUENCING

Sequencing reactions were trialled in the Functional Genomics and Proteomics unit at the University of Birmingham, using an ABI 3730 DNA analyser. After trials gel extractions
were sent to the Institute of Biological, Environmental and Rural Sciences (formerly IGER) at the University of Aberystwyth for sequencing on the ABI Prism 3130 sequencer.

4.2.7 **Genetic Diversity Analysis – Programs Used for *T. pratense* Analysis**

**File Preparation**

Sequences were manually edited and aligned using the software PROSEQ version 3 (Filatov, 2002). Heterozygous sites in the unresolved dataset were manually coded using IUPAC (International Union of Pure and Applied Chemistry) notation. Singletons, those polymorphisms appearing in one individual, were removed from the dataset due to the potential for these to have derived from sequencing error. PDRAW32 (AcaClone software, http://www.acaclone.com) was used to image the loci amplified in this study. BLAST (Zhang *et al.*, 2000), ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/), GenScan (Burge & Karlin, 1997) and a translation macro written in Excel were used to obtain putative coding regions for the amplified loci.

**Basic analyses and neutrality tests**

The PHASE algorithm (Stephens *et al.*, 2001; Stephens & Donnelly 2003) was implemented within DNA SP version 5 (Librado & Rozas, 2009) to resolve haplotype phase for the *T. pratense* individuals. Basic statistics, such as the number of variable sites for each locus, nucleotide diversity and haplotype diversity were generated from the resolved sequences using DNA SP. Sites with alignment gaps, as in locus 22 at 383bp (Figure 4.1), and sites containing missing data are excluded from most analyses by the program. Nucleotide diversity (π) was assessed as the average number of nucleotide differences per site between two sequences (Nei, 1987). $F_{ST}$, a measure of population differentiation or genetic distance was also calculated for each locus using DNA SP.
Neutrality tests such as Tajima’s $D$, Fay and Wu’s $H$ and the McDonald and Kreitman test were conducted to establish whether there are any departures from the neutral model of evolution (Kimura, 1983) which could bias further analyses that require loci to be in Hardy-Weinburg equilibrium. Tajima’s $D$ (Tajima, 1989) is used to test whether mutations are selectively neutral by assessing the differences between the number of segregating sites and the average number of nucleotide differences. If the numbers are the same or similar then null hypothesis of neutrality cannot be rejected, however if the two values are significantly different then there is evidence for selection at loci. Positive values, a result of low numbers of low and high frequency SNPs indicate balancing selection or population decline, with negative values resulting from an excess of low frequency mutations, indicating purifying selection and/or population expansion. Fay and Wu’s $H$ (Fay & Wu, 2000) is similarly based on the distribution of alleles within populations, testing the average number of nucleotide differences between pairs of sequences and $\theta_H$ which is an estimator based on the frequency of derived variants. Significance was tested using 10,000 coalescent simulations within DNASp. A sequence from $T. dubium$ was used as an outgroup for this test. Under neutrality this value should be near to 0, while significantly negative values indicate an excess of high frequency variants (non ancestral) , potentially indicative of positive selection.

The above tests can be susceptible to underlying demographic factors, such as the presence of population structure or population growth (Nielsen, 2005). The McDonald and Kreitman test (MKT) was used to further test these loci for evidence of selection as this test is more robust to demographic assumptions (although see Egea et al., 2008). MKT compares synonymous and non-synonymous substitutions within and between species to test for selection, testing the idea that negative selection should decrease the level of non-synonymous mutations while positive selection increase it, with the effect stronger in divergence data than
within species (Nielsen, 2005; Egea et al., 2008). Significant values indicate a departure from neutral evolution.

Isolation by distance

Isolation by distance (IBD) was tested using a Mantel test of matrices of pairwise $F_{ST}$ values transformed to $F_{ST}/(1-F_{ST})$ and log-transformed geographic distances Rousset (1997). One population was removed from the T21 locus analysis (SHT4) as it contains only one individual. Mantel tests were carried out in GENALEX version 6.1 (Peakall & Smouse, 2006), on $F_{ST}$ matrices produced in DNA SP, with significance tested using 999 permutations. A matrix of average $F_{ST}$ values between populations was generated by averaging values from the locus specific matrices.

Haplotype analysis and AMOVA

Networks based on haplotype frequencies were produced using TCS version 1.21 (Clement et al., 2000), using the default parameters. TCS uses the statistical parsimony (SP) method outlined in Templeton et al. (1992), which assesses the maximum number of differences that are the results of a single substitution, and then joins individuals that differ by one substitution, then two, then three etc until the maximum limit is reached or all the haplotypes are included in the network (Posada & Crandall, 2002). Thus, SP emphasizes similarities between haplotypes and gives weight to shorter branches (Joly et al., 2007). The relationship between haplotypes and geographical area was shown by grouping the haplotypes present in each region and using this data to produce a pie chart, displaying the chart on the appropriate region on the UK.

Analysis of molecular variance (AMOVA) was carried out using ARLEQUIN (version 2.0) (Schneider et al., 2000) on each locus to estimate the partition of variation among
regions, among populations within regions and within populations. Population regions were identified as groups defined by clustering analysis of AFLP markers (see Chapter 3).

**Linkage disequilibrium**

DNA SP was used to calculate the degree of linkage disequilibrium (LD) between all informative sites within loci. Tri-allelic positions, such as those in locus 96, are excluded from linkage disequilibrium analysis by DNA SP and therefore these two SNPs (positions 103bp and 181bp) were removed from further analyses. $R^2$ values (Hill & Robertson, 1968) generated in DNA SP were chosen to provide values between 0 and 1 for graphing purposes. Data was transferred to Excel where triangular matrices were formatted to produce heatmaps to graphically display the data, with red squares indicating LD measures near to 1 and green used to indicate values near to 0. A Fisher’s exact test, implemented in DNA SP, was used to test whether associations between two SNP positions were significant.

**Analysis across all five loci**

An input file for analyses was created by coding the base pairs at each of the SNP positions at each locus as numerical codominant data. Using this input file genetic distance between populations was generated in GENALEX and the resulting matrix used to perform a principal coordinates analysis (PCA) within GENALEX. An unrooted dendrogram was constructed from the matrix using UPGMA within the NEIGHBOR program of the PHYLIP package (version 3.67) (Felsenstein, 2004). Model-based clustering analysis implemented in the program STRUCTURE version 2.3.3 (Pritchard et al., 2000) was used to further investigate population structure in the *T. pratense* individuals, without prior information on sampling areas. STRUCTURE identifies $K$ (unknown) populations within the dataset and assigns each population/individual to one or more population/cluster. An admixture model with independent allele frequencies was run fifteen times for each value of $K$ ($K = 2$-$10$), with a
burn-in period of $10^5$ for $50 \times 10^4$ iterations. Runs using the correlated and independent options for allele frequencies showed equivalent results. The number of clusters was determined using the method outlined by Evanno et al. (2005). STRUCTURE plots were visualised using DISTRUCTION version 1.1 (Rosenburg, 2004).
4.3 Results

4.3.1 DNA Extraction and Sequencing

Good quality sequences were obtained for the five amplified loci; T21 (EST reference BB915621), T52 (BB925852), T55 (BB910055), T74 (BB916074) and T96 (BB906196), totalling 2,331bp. All sequence data was checked by eye before subsequent analysis. Sequence lengths and number of individuals sequenced for each locus are given in Table 4.4, with images of the regions sequenced in *T. pratense* given in Figure 4.1.

![Table 4.4 Number and types of polymorphisms found in the five loci for the 70 *T. pratense* individuals assessed in the study. * including one INDEL. ^ Two transitions were present in the coding region, and no transversion. # Only two tri-allelic SNPs were identified in the coding region and were not counted for transition/transversion ratios.](image-url)
SNP diversity in *T. pratense*

---

**a.**

- (A/G/C/A - Ala-Thr - 36)
- A/T(G/T) - Ile-Met - 47
- G/G(A/G) - Gly - 83
- A/T(A/G) - Met-ile - 137
- G/T(DCA) - Ala-Ser - 279
- G(DCA) - Ala-Ser - 281
- InDel - chip - 383

420 bp

---

**b.**

- CT - 87
- A/T - 89

420 bp

---

**c.**

- T/G(A/G) - Ser - 120
- C/A(G/T) - Gln-His - 258
- A/G(A/T) - Ala-Asp - 357
- A/G(T/C) - His-Thr - 362

437 bp
SNP diversity in *T. pratense*

Figure 4.1 Pictorial representations of loci amplified from *Trifolium pratense*. Blue bars indicate areas homologous to sequences found during a BLAST search. Orange bars indicate those areas designated as ORFs using ORF Finder. a) T21, similar to ERF4 b) T52, similar to EOL1 c) T55, similar to sucrose phosphate synthase d) T74, similar to HK1 e) T96, similar to EREBP.

Total SNP frequency within the loci analysed is 1/63 base pairs (polymorphism frequency 1/61 base pairs), and 1/93 base pairs in the coding regions. The average ratio of transitions (purine/purine or pyrimidine/pyrimidine) to transversion (purine to pyrimidine and vice versa) is 1.71 indicating a transition bias in this species. The ratio found here is similar to
the values reported for other plant species, such as 1.72 for *Arabidopsis thaliana* and 1.82 in *Populus tremula* (Martinez-Castilla & Alvarez-Buyella, 2003; Ingvarsson, 2008).

### 4.3.2 Intraspecific Diversity and Neutrality Analysis

General diversity statistics are shown in Table 4.5. These results show a high level of variability in this species, with diversity measures varying between loci, as well as between populations for the same loci. Highest diversity across all individuals was found in the T21, T74 and T96 loci, with the lowest diversity found in the T55 locus. These results are reflected in the haplotypic diversity measures. While locus T21 has a high level of diversity this is restricted to some populations as two populations show little or no diversity for this locus (IOS4 and LKD3). Similarly for locus T55, only six of the 14 populations are polymorphic.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Samples</th>
<th>Length</th>
<th>Variable sites</th>
<th>( \pi_{\text{all}} )</th>
<th>( \pi_{\text{syn}} )</th>
<th>( \pi_{\text{rep}} )</th>
<th>( \pi_{\text{sil}} )</th>
<th>Hd</th>
<th>Tajima's D</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>52</td>
<td>426</td>
<td>6</td>
<td>4.42</td>
<td>7.56</td>
<td>4.62</td>
<td>4.09</td>
<td>0.771</td>
<td>1.375</td>
</tr>
<tr>
<td>IOS4</td>
<td>3</td>
<td>420</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.000</td>
<td>n/a</td>
</tr>
<tr>
<td>IOS5</td>
<td>5</td>
<td>420</td>
<td>4</td>
<td>3.92</td>
<td>4.72</td>
<td>4.62</td>
<td>2.55</td>
<td>0.689</td>
<td>0.626</td>
</tr>
<tr>
<td>LKD1</td>
<td>4</td>
<td>420</td>
<td>2</td>
<td>2.64</td>
<td>0.00</td>
<td>3.97</td>
<td>0.00</td>
<td>0.679</td>
<td>1.621</td>
</tr>
<tr>
<td>LKD3</td>
<td>5</td>
<td>420</td>
<td>2</td>
<td>1.59</td>
<td>0.00</td>
<td>2.39</td>
<td>0.00</td>
<td>0.600</td>
<td>-0.184</td>
</tr>
<tr>
<td>LKD6</td>
<td>5</td>
<td>420</td>
<td>3</td>
<td>2.54</td>
<td>4.73</td>
<td>2.55</td>
<td>2.55</td>
<td>0.356</td>
<td>0.021</td>
</tr>
<tr>
<td>NWS4</td>
<td>5</td>
<td>420</td>
<td>4</td>
<td>3.65</td>
<td>9.42</td>
<td>2.95</td>
<td>5.10</td>
<td>0.511</td>
<td>1.375</td>
</tr>
<tr>
<td>T21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKY3</td>
<td>3</td>
<td>420</td>
<td>4</td>
<td>5.56</td>
<td>7.98</td>
<td>6.22</td>
<td>4.31</td>
<td>0.733</td>
<td>1.799</td>
</tr>
<tr>
<td>SKY5</td>
<td>3</td>
<td>420</td>
<td>5</td>
<td>5.08</td>
<td>11.51</td>
<td>4.54</td>
<td>6.22</td>
<td>0.800</td>
<td>-0.144</td>
</tr>
<tr>
<td>UIS5</td>
<td>5</td>
<td>420</td>
<td>5</td>
<td>4.39</td>
<td>10.90</td>
<td>3.67</td>
<td>5.90</td>
<td>0.511</td>
<td>-0.247</td>
</tr>
<tr>
<td>DEV2</td>
<td>5</td>
<td>420</td>
<td>3</td>
<td>3.39</td>
<td>4.72</td>
<td>3.83</td>
<td>2.55</td>
<td>0.711</td>
<td>1.227</td>
</tr>
<tr>
<td>DEV3</td>
<td>5</td>
<td>420</td>
<td>2</td>
<td>2.12</td>
<td>7.56</td>
<td>4.62</td>
<td>4.09</td>
<td>0.622</td>
<td>0.830</td>
</tr>
<tr>
<td>SHT2</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>SHT4</td>
<td>1</td>
<td>426</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.000</td>
<td>n/a</td>
</tr>
<tr>
<td>SHT6</td>
<td>3</td>
<td>420</td>
<td>3</td>
<td>3.81</td>
<td>7.08</td>
<td>3.83</td>
<td>3.83</td>
<td>0.800</td>
<td>1.124</td>
</tr>
</tbody>
</table>

Table 4.5 Summary polymorphism statistics for all five loci; \(^a\) Nucleotide diversity in all sites; \(^b\) Nucleotide diversity in synonymous sites; \(^c\) Nucleotide diversity in replacement sites; \(^d\) Nucleotide diversity in silent sites; \(^e\) haplotype diversity; \(^1\) no SNPs present; \(^2\) Data from two or more individuals required to perform analyses; \(^*\)values have been multiplied by 10\(^3\).
<table>
<thead>
<tr>
<th>Locus</th>
<th>Samples</th>
<th>Length</th>
<th>Variable sites</th>
<th>$\pi_{\text{all}}$</th>
<th>$\pi_{\text{syn}}$</th>
<th>$\pi_{\text{rep}}$</th>
<th>$\pi_{\text{sil}}$</th>
<th>$\text{Hd}^c$</th>
<th>Tajima's D</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>62</td>
<td>547</td>
<td>9</td>
<td>2.01</td>
<td>2.22</td>
<td>0.00</td>
<td>3.41</td>
<td>0.620</td>
<td>-0.822</td>
</tr>
<tr>
<td>IOS4</td>
<td>4</td>
<td>547</td>
<td>3</td>
<td>2.02</td>
<td>0.00</td>
<td>0.00</td>
<td>3.43</td>
<td>0.607</td>
<td>-0.177</td>
</tr>
<tr>
<td>IOS5</td>
<td>4</td>
<td>547</td>
<td>3</td>
<td>1.70</td>
<td>0.00</td>
<td>0.00</td>
<td>2.87</td>
<td>0.643</td>
<td>-0.812</td>
</tr>
<tr>
<td>LKD1</td>
<td>5</td>
<td>547</td>
<td>5</td>
<td>2.97</td>
<td>0.00</td>
<td>0.00</td>
<td>5.02</td>
<td>0.733</td>
<td>-0.329</td>
</tr>
<tr>
<td>LKD3</td>
<td>5</td>
<td>547</td>
<td>1</td>
<td>0.37</td>
<td>2.85</td>
<td>0.00</td>
<td>0.66</td>
<td>0.200</td>
<td>-1.112</td>
</tr>
<tr>
<td>LKD6</td>
<td>5</td>
<td>547</td>
<td>3</td>
<td>1.67</td>
<td>0.00</td>
<td>0.00</td>
<td>2.82</td>
<td>0.511</td>
<td>-0.507</td>
</tr>
<tr>
<td>NWS4</td>
<td>5</td>
<td>547</td>
<td>4</td>
<td>1.46</td>
<td>0.00</td>
<td>0.00</td>
<td>2.48</td>
<td>0.378</td>
<td>-1.667</td>
</tr>
<tr>
<td>T52</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKY3</td>
<td>5</td>
<td>547</td>
<td>3</td>
<td>1.95</td>
<td>5.07</td>
<td>0.00</td>
<td>3.30</td>
<td>0.622</td>
<td>0.021</td>
</tr>
<tr>
<td>SKY5</td>
<td>5</td>
<td>547</td>
<td>5</td>
<td>2.11</td>
<td>5.07</td>
<td>0.00</td>
<td>3.58</td>
<td>0.644</td>
<td>-1.388</td>
</tr>
<tr>
<td>UIS5</td>
<td>5</td>
<td>547</td>
<td>3</td>
<td>2.32</td>
<td>2.22</td>
<td>0.00</td>
<td>3.41</td>
<td>0.600</td>
<td>-0.822</td>
</tr>
<tr>
<td>DEV2</td>
<td>5</td>
<td>547</td>
<td>4</td>
<td>1.95</td>
<td>6.65</td>
<td>0.00</td>
<td>3.30</td>
<td>0.711</td>
<td>-0.943</td>
</tr>
<tr>
<td>DEV3</td>
<td>2</td>
<td>547</td>
<td>1</td>
<td>0.91</td>
<td>7.13</td>
<td>0.00</td>
<td>1.55</td>
<td>0.500</td>
<td>-0.612</td>
</tr>
<tr>
<td>SHT2</td>
<td>4</td>
<td>547</td>
<td>1</td>
<td>0.46</td>
<td>0.00</td>
<td>0.00</td>
<td>0.77</td>
<td>0.250</td>
<td>-1.055</td>
</tr>
<tr>
<td>SHT4</td>
<td>5</td>
<td>547</td>
<td>1</td>
<td>0.98</td>
<td>0.00</td>
<td>0.00</td>
<td>1.65</td>
<td>0.533</td>
<td>1.303</td>
</tr>
<tr>
<td>SHT6</td>
<td>3</td>
<td>547</td>
<td>1</td>
<td>0.61</td>
<td>0.00</td>
<td>0.00</td>
<td>1.03</td>
<td>0.333</td>
<td>-0.933</td>
</tr>
<tr>
<td>All</td>
<td>67</td>
<td>437</td>
<td>4</td>
<td>0.66</td>
<td>0.63</td>
<td>0.67</td>
<td>0.00</td>
<td>0.209</td>
<td>-1.143</td>
</tr>
<tr>
<td>IOS5</td>
<td>5</td>
<td>437</td>
<td>2</td>
<td>1.68</td>
<td>5.72</td>
<td>0.59</td>
<td>0.00</td>
<td>0.600</td>
<td>0.120</td>
</tr>
<tr>
<td>LKD1</td>
<td>5</td>
<td>437</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.000</td>
<td>n/a</td>
</tr>
<tr>
<td>LKD3</td>
<td>5</td>
<td>437</td>
<td>2</td>
<td>1.63</td>
<td>0.00</td>
<td>2.08</td>
<td>0.00</td>
<td>0.356</td>
<td>0.019</td>
</tr>
<tr>
<td>LKD6</td>
<td>5</td>
<td>437</td>
<td>2</td>
<td>0.92</td>
<td>0.00</td>
<td>1.17</td>
<td>0.00</td>
<td>0.200</td>
<td>-1.401</td>
</tr>
<tr>
<td>NWS4</td>
<td>4</td>
<td>437</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.000</td>
<td>n/a</td>
</tr>
<tr>
<td>T55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKY3</td>
<td>4</td>
<td>437</td>
<td>1</td>
<td>0.98</td>
<td>0.00</td>
<td>1.25</td>
<td>0.00</td>
<td>0.429</td>
<td>0.334</td>
</tr>
<tr>
<td>SKY5</td>
<td>5</td>
<td>437</td>
<td>1</td>
<td>0.81</td>
<td>0.00</td>
<td>1.04</td>
<td>0.00</td>
<td>0.356</td>
<td>0.015</td>
</tr>
<tr>
<td>UIS5</td>
<td>4</td>
<td>437</td>
<td>3</td>
<td>2.37</td>
<td>0.00</td>
<td>3.03</td>
<td>0.00</td>
<td>0.679</td>
<td>-0.431</td>
</tr>
<tr>
<td>DEV2</td>
<td>5</td>
<td>437</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.000</td>
<td>n/a</td>
</tr>
<tr>
<td>DEV3</td>
<td>5</td>
<td>437</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.000</td>
<td>n/a</td>
</tr>
<tr>
<td>SHT2</td>
<td>5</td>
<td>437</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.000</td>
<td>n/a</td>
</tr>
<tr>
<td>SHT4</td>
<td>5</td>
<td>437</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.000</td>
<td>n/a</td>
</tr>
<tr>
<td>SHT6</td>
<td>5</td>
<td>437</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.000</td>
<td>n/a</td>
</tr>
<tr>
<td>All</td>
<td>65</td>
<td>492</td>
<td>11</td>
<td>4.68</td>
<td>16.20</td>
<td>2.57</td>
<td>8.51</td>
<td>0.732</td>
<td>0.348</td>
</tr>
<tr>
<td>IOS4</td>
<td>5</td>
<td>492</td>
<td>5</td>
<td>3.97</td>
<td>14.03</td>
<td>2.11</td>
<td>7.37</td>
<td>0.778</td>
<td>0.427</td>
</tr>
<tr>
<td>IOS5</td>
<td>4</td>
<td>492</td>
<td>4</td>
<td>3.12</td>
<td>12.06</td>
<td>1.36</td>
<td>6.33</td>
<td>0.607</td>
<td>-0.020</td>
</tr>
<tr>
<td>LKD1</td>
<td>4</td>
<td>492</td>
<td>5</td>
<td>3.63</td>
<td>12.07</td>
<td>2.15</td>
<td>6.34</td>
<td>0.464</td>
<td>-0.335</td>
</tr>
<tr>
<td>LKD3</td>
<td>5</td>
<td>492</td>
<td>6</td>
<td>5.78</td>
<td>17.40</td>
<td>3.94</td>
<td>9.14</td>
<td>0.711</td>
<td>1.411</td>
</tr>
<tr>
<td>LKD6</td>
<td>5</td>
<td>492</td>
<td>9</td>
<td>6.96</td>
<td>19.38</td>
<td>5.20</td>
<td>10.17</td>
<td>0.822</td>
<td>0.332</td>
</tr>
<tr>
<td>NWS4</td>
<td>5</td>
<td>492</td>
<td>4</td>
<td>3.66</td>
<td>13.79</td>
<td>1.69</td>
<td>7.24</td>
<td>0.644</td>
<td>1.048</td>
</tr>
<tr>
<td>T74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKY3</td>
<td>5</td>
<td>492</td>
<td>5</td>
<td>4.20</td>
<td>17.43</td>
<td>1.48</td>
<td>9.15</td>
<td>0.778</td>
<td>0.679</td>
</tr>
<tr>
<td>SKY5</td>
<td>5</td>
<td>492</td>
<td>4</td>
<td>4.20</td>
<td>16.69</td>
<td>1.69</td>
<td>8.77</td>
<td>0.733</td>
<td>1.772</td>
</tr>
<tr>
<td>UIS5</td>
<td>4</td>
<td>492</td>
<td>5</td>
<td>3.63</td>
<td>9.32</td>
<td>2.94</td>
<td>4.90</td>
<td>0.464</td>
<td>-0.335</td>
</tr>
<tr>
<td>DEV2</td>
<td>5</td>
<td>492</td>
<td>6</td>
<td>5.15</td>
<td>12.32</td>
<td>4.43</td>
<td>6.48</td>
<td>0.800</td>
<td>0.804</td>
</tr>
</tbody>
</table>
Tests of neutrality were conducted on the data and the results given in Table 4.6, with some loci giving positive (indicating balancing selection or population decline) and some giving negative (indicating purifying selection or population expansion) Tajima’s $D$ values.

Table 4.6 Neutrality tests in 5 *Trifolium pratense* loci. $D$ = Tajimas D, $H$ =Fay and Wu’s $H$ (normalisation). $d_n$ Number of synonymous nucleotide substitutions per synonymous site, $d_s$ number of non-synonymous sites per non-synonymous sites, $d_N$ and $d_S$ Jukes and cantor correction.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Samples</th>
<th>Length</th>
<th>Variable sites</th>
<th>$\pi$ all*</th>
<th>$\pi$ synb*</th>
<th>$\pi$ rep*</th>
<th>$\pi$ sil*</th>
<th>$H_d$</th>
<th>Tajima’s D</th>
</tr>
</thead>
<tbody>
<tr>
<td>T74</td>
<td>DEV3</td>
<td>4</td>
<td>492</td>
<td>7</td>
<td>4.14</td>
<td>11.26</td>
<td>3.17</td>
<td>5.92</td>
<td>0.643</td>
</tr>
<tr>
<td></td>
<td>SHT2</td>
<td>5</td>
<td>492</td>
<td>1</td>
<td>0.72</td>
<td>3.86</td>
<td>0.00</td>
<td>2.03</td>
<td>0.356</td>
</tr>
<tr>
<td></td>
<td>SHT4</td>
<td>4</td>
<td>492</td>
<td>4</td>
<td>2.03</td>
<td>8.17</td>
<td>0.79</td>
<td>4.29</td>
<td>0.250</td>
</tr>
<tr>
<td></td>
<td>SHT6</td>
<td>5</td>
<td>492</td>
<td>1</td>
<td>0.41</td>
<td>2.17</td>
<td>0.00</td>
<td>1.14</td>
<td>0.200</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>62</td>
<td>429</td>
<td>7</td>
<td>3.74</td>
<td>10.53</td>
<td>2.68</td>
<td>4.34</td>
<td>0.801</td>
</tr>
<tr>
<td></td>
<td>IOS4</td>
<td>4</td>
<td>429</td>
<td>3</td>
<td>3.08</td>
<td>5.81</td>
<td>0.00</td>
<td>4.79</td>
<td>0.679</td>
</tr>
<tr>
<td></td>
<td>IOS5</td>
<td>5</td>
<td>429</td>
<td>5</td>
<td>3.94</td>
<td>4.65</td>
<td>0.00</td>
<td>6.12</td>
<td>0.933</td>
</tr>
<tr>
<td></td>
<td>LKD1</td>
<td>5</td>
<td>429</td>
<td>4</td>
<td>3.83</td>
<td>8.83</td>
<td>3.65</td>
<td>3.95</td>
<td>0.778</td>
</tr>
<tr>
<td></td>
<td>LKD3</td>
<td>4</td>
<td>429</td>
<td>2</td>
<td>2.25</td>
<td>10.03</td>
<td>3.52</td>
<td>1.55</td>
<td>0.607</td>
</tr>
<tr>
<td></td>
<td>LKD6</td>
<td>5</td>
<td>429</td>
<td>4</td>
<td>3.42</td>
<td>4.67</td>
<td>3.50</td>
<td>3.38</td>
<td>0.889</td>
</tr>
<tr>
<td></td>
<td>NWS4</td>
<td>5</td>
<td>429</td>
<td>1</td>
<td>1.30</td>
<td>0.00</td>
<td>3.65</td>
<td>0.00</td>
<td>0.556</td>
</tr>
<tr>
<td>T96</td>
<td>SKY3</td>
<td>2</td>
<td>429</td>
<td>3</td>
<td>3.50</td>
<td>11.72</td>
<td>3.28</td>
<td>3.63</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>SKY5</td>
<td>4</td>
<td>429</td>
<td>3</td>
<td>4.08</td>
<td>22.44</td>
<td>1.64</td>
<td>5.44</td>
<td>0.679</td>
</tr>
<tr>
<td></td>
<td>UIS5</td>
<td>5</td>
<td>429</td>
<td>4</td>
<td>2.49</td>
<td>16.54</td>
<td>0.00</td>
<td>3.86</td>
<td>0.356</td>
</tr>
<tr>
<td></td>
<td>DEV2</td>
<td>5</td>
<td>429</td>
<td>4</td>
<td>3.99</td>
<td>12.44</td>
<td>2.33</td>
<td>4.91</td>
<td>0.822</td>
</tr>
<tr>
<td></td>
<td>DEV3</td>
<td>5</td>
<td>429</td>
<td>4</td>
<td>2.85</td>
<td>4.66</td>
<td>3.06</td>
<td>2.74</td>
<td>0.800</td>
</tr>
<tr>
<td></td>
<td>SHT2</td>
<td>5</td>
<td>429</td>
<td>3</td>
<td>2.75</td>
<td>0.00</td>
<td>3.06</td>
<td>2.58</td>
<td>0.689</td>
</tr>
<tr>
<td></td>
<td>SHT4</td>
<td>4</td>
<td>429</td>
<td>2</td>
<td>1.83</td>
<td>5.83</td>
<td>3.52</td>
<td>0.91</td>
<td>0.679</td>
</tr>
<tr>
<td></td>
<td>SHT6</td>
<td>4</td>
<td>429</td>
<td>3</td>
<td>2.58</td>
<td>0.00</td>
<td>1.64</td>
<td>3.11</td>
<td>0.607</td>
</tr>
</tbody>
</table>

However, these results give no significant evidence of a departure from neutral expectations, except in two loci, T52 and T55 where Fay and Wu’s $H$ indicated that these loci...
are putatively under positive selection. The McDonald and Kreitman test, a test for neutrality that is not as susceptible to bias due to population structure as Fay and Wu’s $H$ (Nielson, 2005), showed no significant evidence of selection at any of the loci (tables not shown).

4.3.3 Genetic Differentiation

Values of $F_{ST}$ for each locus are given in Figure 4.2, varying from 0.134 in locus T55 to 0.225 in locus T74.

![Figure 4.2](image)

Figure 4.2 $F_{ST}$ values for five loci assessed in the study.

4.3.4 Isolation by Distance

A significant relationship between geographic and genetic distance was found in locus T52 and T74 (Figure 4.3 b. and Figure 4.3 d. respectively) as well as when grouping data for all loci. However, while these relationships are significant, only a small proportion of variation in genetic distance in these loci is described by geographic distance, as reflected in the $R^2$ values.
Figure 4.3 Mantel tests of isolation by distance plotted using transformed $F_{ST}$ (Rousset, 1997) against log transformed distance values; a. T21 b. T52 c. T55 d. T74 e. T96 f. all loci.
4.3.5 Analysis of Molecular Variation

AMOVA results revealed that the largest part of the variation is found within populations for all loci (Table 4.7).

Table 4.7 AMOVA for the five loci assessed in the study. d.f., degrees of freedom; SS, sums of squares; Var, variance components; and %, percentage of variation.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Source of variation</th>
<th>Among regions</th>
<th>Among populations within regions</th>
<th>Within populations</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>d.f.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T21</td>
<td>SS</td>
<td>22.63</td>
<td>9</td>
<td>104.64</td>
<td>146.96</td>
</tr>
<tr>
<td></td>
<td>Var</td>
<td>0.36</td>
<td>0.12</td>
<td>1.16</td>
<td>1.64</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>21.96</td>
<td>7.19</td>
<td>70.85</td>
<td></td>
</tr>
<tr>
<td>T52</td>
<td>SS</td>
<td>4.72</td>
<td>14.68</td>
<td>48.38</td>
<td>67.77</td>
</tr>
<tr>
<td></td>
<td>Var</td>
<td>0.03</td>
<td>0.10</td>
<td>0.44</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>4.52</td>
<td>17.95</td>
<td>77.53</td>
<td></td>
</tr>
<tr>
<td>T55</td>
<td>SS</td>
<td>1.30</td>
<td>2.89</td>
<td>15.03</td>
<td>19.21</td>
</tr>
<tr>
<td></td>
<td>Var</td>
<td>0.01</td>
<td>0.01</td>
<td>0.13</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>6.32</td>
<td>9.63</td>
<td>84.05</td>
<td></td>
</tr>
<tr>
<td>T74</td>
<td>SS</td>
<td>24.44</td>
<td>17.90</td>
<td>106.10</td>
<td>148.43</td>
</tr>
<tr>
<td></td>
<td>Var</td>
<td>0.26</td>
<td>0.08</td>
<td>0.91</td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>20.77</td>
<td>6.15</td>
<td>73.08</td>
<td></td>
</tr>
<tr>
<td>T96</td>
<td>SS</td>
<td>10.50</td>
<td>17.89</td>
<td>70.40</td>
<td>98.79</td>
</tr>
<tr>
<td></td>
<td>Var</td>
<td>0.10</td>
<td>0.11</td>
<td>0.64</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>11.33</td>
<td>13.19</td>
<td>75.47</td>
<td></td>
</tr>
</tbody>
</table>

This pattern is similar to that found using AFLP markers for these populations in Chapter 3, where 84% of variation in these markers was found within populations along with other published studies of variation in *T. pratense* (Hagen & Hamrick 1998; Mosjidis et al., 2004). Both locus T21 and locus T74, in comparison to the other loci in this study, show high variation among regions, 21.96% and 20.77% respectively.
4.3.6 **Geographic Haplotype Diversity**

Haplotypes generated from DNAsp were grouped by regions and used to generate the maps shown in Figure 4.4 - Figure 4.8 to further evaluate the spatial genetic diversity in these loci. It should be noted that population numbers are not equivalent between regions, with the Outer Hebrides consisting of one population, the Isles of Scilly and Devon two populations, and the Lake District, North West Scotland and Shetland consisting of three populations. The haplotypic analysis for the T21 locus indicates that some haplotypes only occur in specific geographical regions (haplotypes T21g, T21h, T21i and T21j), while others have a much wider distribution. However haplotypes T21c and T21d are restricted to the north of the range. The T52 locus (Figure 4.5) shows little geographic segregation, although certain haplotypes T52e, T52f, T52h and T52j, are unique to single regions. T55 only consists of five haplotypes, of which haplotype T55a is found across the UK and T55b and T55c are restricted to the north of the region. T74 shows no clear patterns although T74g is found only in the south, in both Devon and the Isles of Scilly. T96 shows a more geographically segregated pattern, with only four of the 12 haplotypes found in Scotland, and the rest restricted to southern and central UK. The haplotype networks determine which polymorphisms define the differences between the haplotypes, with some polymorphisms shown to be restricted to geographic differences. For example, the haplotype network shown in Figure 4.6 for T55 indicates that the C/T polymorphism is only present in North West Scotland.
Figure 4.4 T21 haplotype analysis. NOTE: The colours in a do not correspond to those in image b. a Haplotype map, representing the number and proportion of haplotypes present at each of the six geographic regions studied. b Haplotype network estimated using statistical parsimony in TCS (Clement et al., 2000). Node size represents number of haplotypes. Letters within nodes correspond to the haplotypes shown in a. DNA changes between haplotypes are indicated on the branches between the nodes. Coloured circles next to each node indicate which populations contain the given haplotype.
Figure 4.5 T52 haplotype analysis. NOTE: The colours in a do not correspond to those in image b. a Haplotype map, representing the number and proportion of haplotypes present at each of the six geographic regions studied. b Haplotype network estimated using statistical parsimony in TCS (Clement et al., 2000). Node size represents number of haplotypes. Letters within nodes correspond to the haplotypes shown in a. DNA changes between haplotypes are indicated on the branches between the nodes. Coloured circles next to each node indicate which populations contain the given haplotype.
Figure 4.6 T55 haplotype analysis. NOTE: The colours in a do not correspond to those in image b. a Haplotype map, representing the number and proportion of haplotypes present at each of the six geographic "regions" studied. b Haplotype network estimated using statistical parsimony in TCS (Clement et al., 2000). Node size represents number of haplotypes. Letters within nodes correspond to the haplotypes shown in a. DNA changes between haplotypes are indicated on the branches between the nodes. Coloured circles next to each node indicate which populations contain the given haplotype.
Figure 4.7 T74 haplotype analysis. NOTE: The colours in a do not correspond to those in image b. a Haplotype map, representing the number and proportion of haplotypes present at each of the six geographic ‘regions’ studied. b Haplotype network estimated using statistical parsimony in TCS (Clement et al., 2000). Node size represents number of haplotypes. Letters within nodes correspond to the haplotypes shown in a. DNA changes between haplotypes are indicated on the branches between the nodes. Coloured circles next to each node indicate which populations contain the given haplotype.
Figure 4.8 T96 haplotype analysis. NOTE: The colours in a do not correspond to those in image b. a Haplotype map, representing the number and proportion of haplotypes present at each of the six geographic „regions“ studied. b Haplotype network estimated using statistical parsimony in TCS (Clement et al., 2000). Node size represents number of haplotypes. Letters within nodes correspond to the haplotypes shown in a. DNA changes between haplotypes are indicated on the branches between the nodes. Coloured circles next to each node indicate which populations contain the given haplotype.
4.3.7 **Linkage Disequilibrium**

Linkage disequilibrium (LD) tests showed that there is a varying level of LD between the SNPs in these loci (Figure 4.9). As a result of this analysis all pairs of polymorphisms that showed significant LD were removed from analyses involving all five loci. In total 15 loci were removed, namely SNPs at 47bp, 83bp and 279bp for the T21 locus, 414bp, 440bp and 525bp for the T52 locus, 337bp for the T55 locus, 287bp, 311bp, 359bp, 392bp, 447bp and 475bp for the T74 locus and 272bp and 389bp for the T96 locus.

![Linkage Disequilibrium Heatmaps](image)

Figure 4.9 Linkage disequilibrium heatmaps for five loci assessed in *T. pratense*. Colours represent $r^2$ values on a scale from 0 to 1 with values nearer to 0 shown as green and those nearer to 1 as red; a T21; b T55; c T96; d T74; e T52.
4.3.8 **ANALYSIS INVOLVING ALL LOCI:**

UPGMA analysis (Figure 4.10) of *T. pratense* populations revealed some population structure, similar to that found using the more numerous AFLP markers (Chapter 3).
Figure 4.10 Population structure of *T. pratense* populations; a UPGMA unrooted dendrogram based on the genetic distance matrix generated in GENALEX. Scale bar represents genetic distance measure; b PCA diagram showing the position of populations in relation to the measured genetic differences in the five loci.

Shetland forms the most distinct cluster, as per the AFLP analysis, although this analysis includes a population from the Outer Hebrides (UIS5). The rest of the populations form a much closer cluster, with the Isles of Scilly forming one grouping, followed by a close grouping of populations from mainland England (Devon and the Lake District) and those from North West Scotland (SKY3, SKY5 and NWS4).

a.

![UPGMA unrooted dendrogram based on genetic distance matrix](image1)

b.

![PCA diagram showing position of populations](image2)
Figure 4.11 STRUCTURE results imaged in DISTRACT for K=3; a Graph of individual membership coefficients for each of the three genetic populations determined by STRUCTURE; b Graph of the population membership coefficients for each of the three genetic populations. Population membership coefficients generated by averaging across individuals within a population.

To further explore the population grouping a PCA was conducted on the matrix of genetic distances generated in GENALEX, (Figure 4.10b) where the two axes capture 81.08% of the total variation, 61.78% for axis 1 and 19.31% for axis 2. In this plot it can be seen that while UJS5 is genetically distant to the main grouping of populations, it is also distant from the populations from Shetland, separated by the variation explained by axis 2.

Results of the STRUCTURE analyses for K=3 are shown in Figure 4.11, as determined by the method devised by Evanno et al. (2005). The graphs show that there is some distinction between three areas of the UK on the basis of this data; southern and central UK (including the Isles of Scilly, south England and LKD), north west Scotland (north west Scotland, Skye and the Outer Hebrides) and Shetland. The southern and central UK grouping shows a varying membership to cluster one (shown in Figure 4.11 in green) and two (purple) and a very low membership to cluster three (grey), excluding LKD6 which shows a slightly higher level of membership to cluster three. The Shetland grouping by contrast shows the highest memberships to clusters one and three with very low membership to cluster two. The north west of Scotland grouping in more evenly apportioned between the three genetic clusters.
4.4 Discussion

*Patterns of polymorphism and population structure in T. pratense using GTMs*

Observed nucleotide diversity across all five loci (3.102 x 10^{-3}) was in general lower than other published values, for example 5.48 x 10^{-3} in *Pinus pinaster* (11 drought stress loci, 20 to 30 individuals from 24 populations across Europe, Eveno et al., 2008) and 10.0 x 10^{-3} in wild annual outcrossing *Zea mays* (23 immunity loci, 8 to 18 individuals from six populations across Mexico, Moeller & Tiffin, 2008). Labate et al. (2009) in a study on *Solanum lycopersicum* landraces found a lower value (1.3 x 10^{-3}) but they noted that studies on wild self-incompatible species within the same genus have a higher diversity 11.0 x 10^{-3} and 12.9 x 10^{-3}. It is known that diversity should theoretically be higher in wild, self-incompatible species such as *T. pratense*, and therefore the low nucleotide diversity demonstrated by these five loci does not fit the expected pattern. One possible explanation for this may be due to the particular loci chosen, with a lower level of diversity indicating a higher level of conservation in these loci across the UK. Another reason may be that singletons were removed from analysis, thus reducing overall diversity estimates.

However, raw numbers of SNPs, (1/93bp in coding regions, 70 individuals) are higher than some values published, such as 1/158bp (coding and non-coding regions, 31 individuals) in *Solanum lycopersicum* (Labate et al., 2009), 1/124bp in *Zea mays* (Ching et al., 2002) and 1/504 in *Glycine max* (Van et al., 2004), but much lower than that identified in wild *Zea mays* (1/18bp, 84 individuals) (Moeller & Tiffin, 2008). Taken on its own this high level of SNPs should indicate a relatively high level of diversity in this species. However in light of the nucleotide diversity estimates being lower than that reported for other species this indicates higher numbers of low frequency variants than in the other species discussed here.
This analysis, using GTMs, further underlined the differentiation between these population groups of *T. pratense* previously found using AFLPs. Haplotype analyses indicate that the differences between regions vary according to loci. T96 for instance indicates a higher diversity at this locus in the south and centre of the UK, with more northern populations showing more limited levels of haplotype diversity. T55 is a less diverse locus, with the lower diversity reflected by fewer haplotypes. However those haplotypes that differ from the majority in this locus are generally unique to geographic regions. It is important to note here that the haplotypes presented were generated from PHASE resolved heterozygous sequences and have not been experimentally confirmed. Therefore the complexities of some of the networks and maps may be attributed to errors during phase allocation and as such this area would benefit from further experimentation.

While population level analyses on the individual loci produced differing patterns of variation, both between populations and between loci, analysis involving all loci (21 polymorphisms) showed that the pattern of variation is linked to geographic area, confirmed by significant but small isolation by distance across all loci. Bayesian cluster analysis split the populations into three genetic groupings, with each geographic region (England, North West Scotland and Shetland) showing a different proportion of membership to the three groups. The UPGMA tree and PCA plot further highlight the distinction of the Shetland grouping from the rest of the populations studied here. In conservation terms these results show the geographically isolated *T. pratense* populations in Shetland are genetically distinct. This distinction shown in both STRUCTURE and the UPGMA analysis would suggest that this population should be treated as a separate management unit for conservation of the genetic diversity of this species.
Signatures of selection

While it is possible to describe the variation in species across a landscape, a more difficult but perhaps more relevant question is to determine the processes that define the patterns of variation. Differentiation between regional types can be due to either restricted migration, or differential adaptation of populations to local conditions (Latta, 2006). Isolation, through restricted migration, will allow random genetic drift to differentiate populations in the absence of migration and/or selection. Conversely patterns of selection for ecologically relevant traits should be expected to be more evident between populations than patterns observed for neutrally evolving loci, known as „local adaptation“ (Lewontin & Krakauer, 1973). Therefore the difference between markers subject to selection and neutral markers would thus be expected to vary. The difference in patterns of variation between neutral and non-neutral markers under similar patterns of inheritance can be evaluated by comparing $F_{ST}$ values; using such different markers can give some evidence of a departure from neutral expectations (Merilä & Crnokrak, 2001; van Tienderen et al. 2002; Latta, 2006). When comparing the $F_{ST}$ values obtained in this analysis with those obtained from AFLP markers (0.1045, see Chapter 3), $F_{ST}$ values for these gene-targeted markers are consistently higher, although the observed difference between values is small. When $F_{ST}$ is higher in gene-targeted markers (those markers within or flanking genes of interest) than in neutral areas (assuming that AFLPs represent neutral markers) this is indicative of divergent selection and local adaptation for the gene (van Tienderen et al., 2002).

However, classical tests of selection comparing these polymorphic markers to neutral expectations found no significant evidence of selection at these loci. While two loci (T52 and T55) showed significant values for Fay and Wu’s $H$, this value was not reflected in the other tests of neutrality. Nielson (2005) notes that population subdivision can lead to the rejection of the neutral model with a high probability and as such it can be difficult to interpret.
significant results. However, for all these tests, the small number of accessions tested and the small segment of sequence analysed from the total coding region does mean that these tests are not absolute, and thus this section would benefit from further testing.

**Suitability of GTMs as markers for conservation questions**

SNP markers are thought to provide an improvement in many ways over conventional markers, through their relative abundance throughout the genome and the less complex mode of evolution (Morin et al., 2004). The potential to use markers that thus provide more and better quality information for subsequent analysis has provided the momentum behind SNP diversity studies, in addition to new and more improved technologies for SNP discovery and genotyping. Further, the implicit assumption that neutral markers can be related to functionally important variation does not always hold in a number of cases (see Reed & Frankham, 2001; Hedrick, 2002). Thus the use of SNPs in functional regions of the genome provides the possibility to answer fundamental questions of conservation genetics. While reliably determining the signatures of selection in natural populations is known to be a challenge (Ouborg et al., 2010a), it has been documented in a number of studies in plant species (for example *Hordeum spontaneum*, Cronin et al. 2007; *Picea glauca*, Namroud et al. 2008; *Zea mays* ssp. *parviglumis*, Moeller & Tiffin, 2008; *Arabidopsis thaliana*, Clark et al., 2007).

The most convincing evidence of selection signatures from DNA level analysis come from two types of studies, those based on genome wide assessments of SNP diversity (*Drosophila* sp., Clark et al., 2007b; *Homo sapiens*, Nielsen et al., 2005; *Pinus pinaster*, Eveno et al., 2008; although see Hermisson, 2009) and those using the candidate gene approach alongside a reference set of loci (Moeller & Tiffin, 2008). These resource intensive methods underline the difficulty in applying these methods to non-model species, where
sequencing efforts have not reached the same levels as those in model species. While this study showed no significant evidence of selection in any of the loci studied, the higher level of differentiation between populations using GTMs compared to that found using AFLPs suggests that these loci could be further studied for signatures of selection. Indeed comparative studies with reference loci, those that are not under the same set of selection pressures as the drought-related genes studied here, could prove beneficial to understanding the selection pressures at these loci. However this possibility would require further sequencing efforts in this or closely related species.

Alongside the difficulties in finding the evidence of selection in ecologically important traits, problems also arise in using this data for biodiversity assessments. Thus while there are huge advances being made to generate SNP data such as 454 sequencing (Roche), subsequent statistical approaches may be limiting. While statistical methods are available to calculate diversity, heterozygosity and population subdivision in markers that are at Hardy-Weinburg equilibrium, markers that have been affected by selection violate one of the main underlying assumptions of these methods. However Narum et al. (2008) suggest that the effects on SNP frequencies from selection may not be a significant problem for most population genetic analyses. In a paper on the feasibility of using these types of markers, Van Tienderen et al. (2002) note the lack of available statistical approaches to study diversity in markers that are putatively markers of functional variation and emphasize the need for these to be developed further.

The results of this study prove an interesting insight into the use of SNPs as molecular markers to answer conservation questions in non-model species. While it has been possible to successfully use GTMs to assess the spatial pattern of genetic diversity in this species, a lower resolution of populations was found compared to that found in AFLPs (Chapter 3). This lower information content in SNP markers underlines the necessity of generating large numbers of
markers to successfully assess population differentiation questions for conservation. Additionally the difficulties in successfully ascribing signatures of selection to these loci made it difficult to fulfil the potential for these markers in determining whether this genetic variation is reflecting the selective environment. Ouborg et al. (2010a) suggest that, due to the large resources required for a conservation genomics approach, conservationists should focus on a smaller number of model conservation species, rather than using this approach in any threatened species. This method would thus focus on answering the wider questions not currently answered in conservation genetics, but is a difficult concept for conservationists currently tasked with saving individual species. Indeed for a biodiversity study in non-model species, it suggests that as yet, while these methods are still in their infancy, other more traditional methods of estimating genetic diversity may be more appropriate. Thus, for conservation, the choice of marker is still very much dependant on the conservation question being asked.
Chapter 5. Are Widespread *Trifolium* Species Really of Least Concern?

In order to find more efficient sampling and conservation strategies it is imperative that studies evaluate the diversity in species at the genetic level in order to fulfil the UK’s objectives, namely to achieve a significant reduction in the of the current rate of biodiversity loss and to promote the conservation of genetic diversity. As such three different methods have been evaluated throughout this assessment; genetic diversity assessments using two different types of genetic marker, SNPs and AFLPs, and using information already available to determine conservation priorities.

Basing a conservation strategy on geographic and ecogeographic information alone would suggest that all three species are of „least concern” (IUCN, 2010), and as such require little conservation effort in relation to other species that are perhaps more at risk because of smaller population sizes. However assessing species according to such criteria does not implicitly take into account either the „value” of a particular species for future breeding programs or any genetic diversity information.

The genetic diversity assessments on these three species given in the previous chapters bring new insights into this conservation strategy, with much wider implications for conservation than that shown by the more traditional methods of conservation planning. The issues relating to the potential for introgression and gene flow between cultivated and wild species highlight a major threat to both *T. repens* and *T. pratense*, which should be incorporated into future conservation plans, with the recommendation being for areas to be set aside for genetic reserves, isolated from crops. As such these populations within reserves would be protected from the import of outside seed to shelter them against future threats from an increasing level of gene flow. Further, the fact that the majority of variation is within, rather than among
populations in all three species, (as supported by both the AFLP analysis and SNP study), suggests that the number of genetic reserves may not need to be numerous, as conserving just a few populations may encompass sufficient genetic diversity. However, each species would require reserves sited in different areas with the most important areas shown to be St Kilda for *T. repens*, the Isles of Scilly for *T. dubium* and Shetland for *T. pratense*.

While this study was not implicitly able to study adaptive diversity it is likely that further investigation using the SNP markers produced over wider areas and more individuals may provide some evidence of selection and adaptation. Indeed the strengths of the differences between individuals from different regions based on so few markers in the SNP study suggests that differential selection is acting on or has acted upon the selected loci. By comparing the AFLP results with the SNP results this study provides some insight into the ability of AFLP markers to detect underlying patterns of adaptation, and to act as a proxy measure of adaptive diversity.

5.1 WIDER IMPLICATIONS FOR WIDESPREAD SPECIES CONSERVATION

Two major assumptions are often made in conservation planning for both collecting missions and reserve placement in the absence of prior information, namely that genetic distance is associated with geographic distance and that neutral markers can be used as a proxy measure of genetic diversity as discussed in Chapter 1. There is some evidence that both of the above assumptions hold true for the three species in this study. However, this does not negate the need for genetic assessments in reserve placement as the distribution of genetic diversity across the sites varied markedly between the target species. Indeed, while geographic distance provided the major source of variation within each species when analysed at the population level, the majority of variation in all species was held within populations. This suggests that for species where there is thought to be high genetic diversity within populations, collecting
missions should focus on a larger number of individuals per population, rather than a larger number of target sites.

Additionally this work raises questions for researchers using Red List Assessment data alone to determine conservation planning. It is clear that while widespread species may exist as large numbers of plants, there may be extensive threats to their genetic diversity that are being overlooked. The study in *T. repens* for example shows that, where wild species are grown alongside a conspecific crop or where there is the potential for the widespread movement of seed there may be implications in terms of genetic swamping and the homogenisation of the gene pool. Indeed, to define a species as „least concern’ is somewhat dangerous where there is little information of the genetic instability underlying the large numbers of individuals. This is particularly relevant in light of the need to use the genetic diversity for future breeding programs in response to a widely changing environment. As such a recommendation based on this work should be the need for Red List Assessments to include criteria that incorporates genetic diversity assessment from the outset.

5.2 A COMPARISON OF THE EFFICIENCY OF AFLP AND SNP MARKERS IN *TRIFOLIUM PRATENSE*

The information in the previous chapters outlines genetic assessments in *T. pratense* using two different types of molecular markers, AFLPs and SNPs. While the difficulties in SNP identification and assessment compared to other types of molecular marker system have been discussed in previous chapters, the relative efficiency of each marker system in *T. pratense* has not been examined. Both marker systems identify similar main themes in the partitioning of diversity, defining the main source of variation as that existing within populations, and in identifying Shetland as a region more distinct from the rest of the regions assessed in this study. However, both marker systems show subtle differences in the patterns of diversity
below the division of Shetland. In addition, both marker systems use very different numbers of loci to achieve similar results. In assessing the relative efficiencies of the two marker systems to elucidate underlying patterns of diversity across the UK, genetic distance amongst populations was compared for the two marker types using a Pearson correlation (Garcia et al., 2004), shown in Figure 5.1. While the patterns of genetic distance distribution between populations remained similar between SNPs and AFLPs, SNPs provided much larger estimates of genetic distance.

Figure 5.1. Pearson correlation coefficient ($r$) estimates between genetic distances (GD) obtained using amplified fragment length polymorphism (AFLP) and single nucleotide polymorphism (SNP) markers in *Trifolium pratense*.

$r = 0.433$

$P = <0.001$
These results indicate that the different marker systems provide consistent information on genetic diversity within *T. pratense* and produced highly comparable genetic distance estimates, despite the differences between the marker systems. The main outliers in Figure 5.1, (i.e. those that show a higher genetic distance between populations using SNPs over AFLPs), are those that define distances between Shetland and the Outer Hebrides, and Shetland and the Isle of Scilly. This indicates that either the SNP markers assessed in this study are finding variation not shown by AFLP markers, or that the larger numbers of markers used in the AFLP study mask some of the underlying diversity at potentially useful loci.

Polymorphic information content (PIC), a measure commonly used to measure polymorphism at marker loci, was assessed using the methods outlined in Stajner *et al.* (2009), where corresponds to presence of a band/allele and  corresponds to the absence of a band/presence of an alternate allele:

Both marker systems provided relatively similar levels of information, but with all average PIC values less than 0.2 (Figure 5.2). These relatively low values (e.g. Stajner *et al.*, 2009; Varshney *et al.*, 2010), indicate that either the markers assessed in this study show low polymorphism, or the individuals studied are not diverse. It is important to note here however that when using this method, it is impossible for PIC values in AFLPs or in biallelic SNPs (those used for assessments in this study) to be over 0.5.

While noting the dissimilarity in genetic distance calculations and in the number of individuals studied between the two marker systems (Chapter 3 and Chapter 4), differences in the genetic distance observed between populations using SNP may also be in part due to the selection of SNP loci that are likely to differ between geographic regions, compared to the
random set of loci assessed when using AFLP. Indeed the much smaller number of SNP loci used to elucidate similar patterns to those generated using AFLP loci, suggests this may be the case. Using SNP markers in the way outlined in Chapter 4 presents an exciting opportunity for future genetic diversity studies, providing the ability to actively select for polymorphism between populations enabling patterns of genetic diversity to be identified using a relatively small set of markers.

Figure 5.2. Distribution of polymorphism information content (PIC) data for *T. dubium* (a.), *T. pratense* (b. and d.) and *T. repens* (c.). The data was obtained using amplified fragment length polymorphism (AFLP) (figures a-c) and single nucleotide polymorphism (SNP) (d). SD = standard deviation.
5.3 Limitations

While the use of AFLPs can be recommended for work in species where there is little previous genetic information, their use in this study provided many problems. The existence of polyploidy in *T. dubium* and *T. repens* and the difficulty of clean DNA extraction in *T. pratense* meant that sequence traces were difficult to analyse and required considerable time and effort to clean traces in order to identify ambiguous peaks and remove traces/individuals that fell below the required standard. Whitlock *et al.* (2008) address this challenge by introducing a computer generated and unambiguous method to analyse sequence traces from AFLP, and future work could adopt this program to quantify error in AFLP traces. The dominant nature of AFLP reduced the significance of many of the assessments, as well as the increased likelihood of homoplasy when using this marker. Microsatellites provide additional information on gene flow that is difficult to ascertain from AFLP markers. As such, on the basis of using AFLP in this study, where there is the potential to use microsatellites generated in related species, they would be recommended for use due to the additional information on gene flow and the codominant information generated.

Although the target species were chosen with a view of the available expertise and previous work, assessments of the two polyploid species, *T. dubium* and *T. repens* increased the uncertainty of the results and with hindsight, this work would have been improved by using species known to be diploid. However, with the prevalence of polyploidy in plant species it provides an interesting insight into the difficulties in polyploid analysis and potentially why most conservation genetic studies are conducted in either diploid, or previously well studied species.
5.4 Future work

1. Given further resources and more time, an interesting avenue of research would be to continue the SNP diversity assessments with the other two species of *Trifolium* assessed with AFLP markers, *T. dubium* and *T. repens*, using the loci amplified in *T. pratense*. An assessment such as this would help to determine the suitability of this method in polyploid species for conservation. While this was attempted in *T. dubium*, polyploidy made the identification of haplotypes difficult. *T. dubium* has an allotetraploid origin (Ansari et al., 2008), but is known to be highly inbreeding (*s* = 0.97) (Dhar et al., 2006) and there is some evidence from work conducted in inbred lines of allotetraploid *Zea mays* that the difficulties associated with SNP discovery and haplotype assignment can be alleviated in the presence of inbreeding (Gaut & Doebley, 1997; Ching et al., 2002). After adapting the PCR protocol to use primers designed in *T. pratense* in *T. dubium* (see Appendix 5), it was possible to use differences in sequence traces to infer heterozygotes from signal intensity (Brumfield et al., 2003). Allele dosage was then estimated from electropherogram peak heights using a method adapted from Nybom et al. (2004) which was designed for microsatellite allele counting. Peak heights replaced peak areas to adapt this method to produce allele dosages for sequence traces obtained from *T. dubium*, using ratios of average peak heights to correct the theoretical ratios for the two alleles. However, determining allele dosage in allotetraploid individuals only solved part of the ploidy issue, with analysis in allopolyploid species requiring variation to be assigned to one or both isoloci, similar to analysis with microsatellite loci. The consistent high levels of fixed heterozygosity for the vast majority of polymorphic positions in *T. dubium* meant that separating sequence data into both haplotypes and isoloci proved unfeasible. Thus it is recommended that further analysis of *T. dubium* would require using the progenitor
comparison approach, which would involve further sequencing of the parental genotypes, *T. campestre* and *T. micranthum* (Ansari, 2008). This small study into using SNPs in *T. dubium* highlights the difficulties involved when using SNP data to analyse polyploid species and the need for more programs to be made available to assist in the analysis, as has been made available for more commonly used genetic markers.

2. Another interesting direction for this work would be to include the analysis of populations and cultivars from a wider geographic range, to ascertain whether the broad trends observed in this work apply on a larger scale. While landraces were chosen based on their history in the UK, more landraces and cultivars would be beneficial in the study of *T. repens* to clarify the results observed, and to further assess the effects of gene flow in this species.

3. Foden (2008) define five traits that make species particularly susceptible to climate change; specialised habitat range, narrow environmental tolerance, poor ability to disperse, dependence on environmental cues for dispersal/breeding and dependence on interactions with other species. In light of this and following discussion with a researcher who has assessed climate change in UK species (J. Preston, pers comm. 2009), an assessment of climate change at the species level was deemed less important in *T. repens* and *T. pratense* due to their abundant and generalist nature across the UK. However, with the identification of the north of the UK as the potential northern limit of *T. dubium*, an assessment of the impacts of climate change in this species would be important to determine how climate change may affect the spread of this species in the UK and how those areas identified for future conservation work may be impacted.

4. Finally, for any conservation work it is vital that the information is disseminated across the conservation network to inform future policy. It will be vital to inform those protected area managers within the regions identified as containing some of the most
genetically diverse populations to promote the potential for management of these sites as genetic reserves.
REFERENCES


Allard RW, Jain SK, Workman PL (1968) Genetics of Inbreeding Populations. Advances in Genetics Incorporating Molecular Genetic Medicine, 14, 55-&.


Bennett SJ (2000) Genetic variation of five species of *Trifolium* L. from south-west Turkey. Genetic Resources and Crop Evolution, 47, 81-91.


207


Frankham R (1997) Do island populations have less genetic variation than mainland populations? Heredity, 78, 311-327.


Hardy OJ, Vekemans X (2001) Patterns of allozyme variation in diploid and tetraploid Centaurea jacea at different spatial scales. Evolution, 55, 943-954.

Hardy OJ, Vekemans X (2002) SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes, 2, 618-620.


Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theoretical and Applied Genetics, 38, 226-231.


Kimura M, Crow JF (1964) Number of alleles that can be maintained in finite population. Genetics, 49, 725-738.


Labate JA, Robertson LD, Baldo AM (2009) Multilocus sequence data reveal extensive
departures from equilibrium in domesticated tomato (Solanum lycopersicum L.). Heredity,
103, 257-267.

Lange O, Schifino-Wittmann MT (2000) Isozyme variation in wild and cultivated species of

Latta RG (2006) Integrating patterns across multiple genetic markers to infer spatial
processes. Landscape Ecology, 21, 809-820.

Lawless KA, Drayton MC, Hand MC, Ponting RC, Cogan NOI, et al. (2009) Interpretation of
SNP Haplotype Complexity in White Clover (Trifolium repens L.), an Outbreeding
Allotetraploid Species. In: Molecular Breeding of Forage and Turf (eds. Yamada T,

Lawrence MJ (1996) Number of incompatibility alleles in clover and other species. Heredity,
76, 610-615.

Size When Collecting Germplasm. Euphytica, 84, 89-99.

Size When Collecting Seed of Cross-Pollinating Species and the Information That Can Be

in Quercus petraea (Matt.) Liebl. III. Patterns of variation identified by geostatistical
analyses. Heredity, 80, 464-473.

Legatt RA, Iwama GK (2003) Occurrence of polyploidy in the fishes. Reviews in Fish Biology
and Fisheries, 13, 237-246.


Legendre P, Fortin MJ (1989) Spatial pattern and ecological analysis. Vegetation, 80, 107-
138.

Conservation Biology, 9, 753-760.

Naturalist, 104, 343-353.

Levin DA (1993) Local speciation in plants - the rule not the exception. Systematic Botany,
18, 197-208.


Moreira PA, Oliveira DA (2011) Leaf age affects the quality of DNA extracted from Dimorphandra mollis (Fabaceae), a tropical tree species from the Cerrado region of Brazil. Genetics and Molecular Research, 10, 353-358.


Qiagen (2006) QIAquick spin Handbook for QIA quick PCR Purification Kit, QIAquick Nucleotide Removal Kit, QIAquick Gel Extraction Kit


Schneider K, Weisshaar B, Borcherdt DC, Salamini F (2001) SNP frequency and allelic haplotype structure of *Beta vulgaris* expressed genes. Molecular Breeding, 8, 63-74.


Scotland GRO (1991) Census: Small Area Statistics (Scotland) [computer file]. ESRC/JISC Census Programme, Census Dissemination Unit, Mimas (University of Manchester)/Centre for Interaction Data Estimation and Research (University of Leeds).


Stanton C (1996) Skye and Lochalsh landscape assessment. Landscape Character Series No. 71


References


Wright S (1931) Evolution in Mendelian populations. Genetics, 16, 97-159.


Appendix 1. CTAB DNA EXTRACTION PROTOCOL

1. Mill 20mg dried leaf material in milling machine to a fine powder.
2. Add milled leaf material to 700 µl pre-heated extraction buffer\(^a\).
3. Incubate at 65°C for 30 min. Mix by inversion after 5 min.
4. Add 700 µl of chloroform:isoamyl (24:1, v/v) and mix by inversion for 5 min.
5. Centrifuge at 13000 rpm for 5 min.
6. Remove supernatant and place in a new tube.
7. Add 2 µl of RNase A (10µg/µl), vortex and incubate at 37°C for 30 min.
8. Add 500 µl of ice cold isopropanol and mix by inversion. Incubate samples at -20°C overnight.
9. Centrifuge at 13 000 rpm for 5 min.
10. Remove supernatant and allow pellets to dry.
11. Add 300 µl of 70% ethanol.
12. Centrifuge at 13 000 rpm for 5 min.
13. Remove supernatant and allow pellets to dry.
14. Resuspend pellet in 120 µl TE buffer\(^a\).

\(^a\)Reagents:

- Extraction buffer: 4% CTAB, 100mM Tris-HCL pH 8.0, 1.4M NaCl, 20mM EDTA, 0.1% mercaptoethanol. Mercaptoethanol added immediately prior to use.
- TE buffer: 10mM TRIS, 1.0 mM EDTA, pH 8.0.
Appendix 2: COMPOSITION OF STOCK SOLUTIONS

Extraction buffer 4% CTAB (TRIS 100mM, NaCl 1.4M, EDTA 20mM, 110mM)

For 500ml of extraction buffer:

<table>
<thead>
<tr>
<th>Component</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRIS</td>
<td>6.057g</td>
</tr>
<tr>
<td>NaCl</td>
<td>40.908g</td>
</tr>
<tr>
<td>Na₂EDTA</td>
<td>3.722g</td>
</tr>
<tr>
<td>CTAB</td>
<td>20.00g</td>
</tr>
</tbody>
</table>

Made up to 500ml with sterile distilled water and autoclaved.

TE buffer (10mM Tris, 1mM EDTA) pH 8.0

For 500ml of TE buffer:

<table>
<thead>
<tr>
<th>Component</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRIS</td>
<td>0.6057g</td>
</tr>
<tr>
<td>Na₂EDTA</td>
<td>0.1861g</td>
</tr>
</tbody>
</table>

Made up to 500ml with sterile distilled water, pH 8.0 with HCl and autoclaved.

TBE buffer (89mM Tris, 89mM Boric acid, 2mM EDTA)

For 1 litre of TBE buffer:

<table>
<thead>
<tr>
<th>Component</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRIS</td>
<td>10.8g</td>
</tr>
<tr>
<td>Boric acid</td>
<td>5.5g</td>
</tr>
<tr>
<td>Na₂EDTA</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Made up to 1l with sterile distilled water.
### Appendix 3. DISTANCE MATRICES

**TABLE A3.1.** *T. dubium* population distance matrix. Colours represent the 5 distance classes.

<table>
<thead>
<tr>
<th></th>
<th>DEV2</th>
<th>DEV3</th>
<th>IOS1</th>
<th>IOS2</th>
<th>IOS3</th>
<th>LKD2</th>
<th>LKD5</th>
<th>SKY1</th>
<th>SKY3</th>
<th>SKY5</th>
<th>UIS7</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEV2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEV3</td>
<td>0</td>
<td>40.603</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOS1</td>
<td>206.294</td>
<td>239.561</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOS2</td>
<td>207.819</td>
<td>241.61</td>
<td>5.409</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOS3</td>
<td>210.302</td>
<td>243.631</td>
<td>4.089</td>
<td>5.078</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LKD2</td>
<td>413.209</td>
<td>381.743</td>
<td>505.949</td>
<td>510.965</td>
<td>509.052</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LKD5</td>
<td>424.277</td>
<td>390.442</td>
<td>531.098</td>
<td>535.98</td>
<td>534.359</td>
<td>40.733</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKY1</td>
<td>769.028</td>
<td>744.282</td>
<td>798.372</td>
<td>803.781</td>
<td>800.24</td>
<td>380.876</td>
<td>393.919</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKY3</td>
<td>754.697</td>
<td>729.703</td>
<td>786.126</td>
<td>791.534</td>
<td>788.038</td>
<td>365.454</td>
<td>378.36</td>
<td>15.569</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKY5</td>
<td>787.028</td>
<td>762.581</td>
<td>813.861</td>
<td>819.269</td>
<td>815.675</td>
<td>400.188</td>
<td>413.386</td>
<td>19.478</td>
<td>35.047</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>UIS7</td>
<td>807.915</td>
<td>785.799</td>
<td>820.054</td>
<td>825.442</td>
<td>821.599</td>
<td>435.707</td>
<td>452.971</td>
<td>73.428</td>
<td>86.298</td>
<td>59.193</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distance classes (km)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;40.60</td>
<td></td>
</tr>
<tr>
<td>≤378.36</td>
<td></td>
</tr>
<tr>
<td>≤509.05</td>
<td></td>
</tr>
<tr>
<td>≤787.03</td>
<td></td>
</tr>
<tr>
<td>≤825.52</td>
<td></td>
</tr>
</tbody>
</table>
**Table A3.2.** *T. pratense* population distance matrix. Colours represent the 8 distance classes.

<table>
<thead>
<tr>
<th></th>
<th>DEV2</th>
<th>DEV3</th>
<th>IOS4</th>
<th>IOS5</th>
<th>LKD1</th>
<th>LKD3</th>
<th>LKD6</th>
<th>NWS4</th>
<th>SHT2</th>
<th>SHT4</th>
<th>SHT6</th>
<th>SKY3</th>
<th>SKY5</th>
<th>UIS5</th>
<th>UIS6</th>
<th>UIS8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40.60</td>
<td></td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>210.25</td>
<td>243.58</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>208.28</td>
<td>242.25</td>
<td></td>
<td>6.53</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>413.25</td>
<td>509.45</td>
<td>513.06</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>422.77</td>
<td>528.53</td>
<td>531.92</td>
<td>28.18</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>424.21</td>
<td>534.26</td>
<td>537.55</td>
<td>39.75</td>
<td>11.86</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>775.32</td>
<td>815.33</td>
<td>820.58</td>
<td>379.15</td>
<td>383.08</td>
<td>388.34</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1096.4</td>
<td>1182.0</td>
<td>1186.5</td>
<td>685.01</td>
<td>673.68</td>
<td>672.29</td>
<td>407.72</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1122.9</td>
<td>1206.7</td>
<td>1211.2</td>
<td>711.14</td>
<td>700.17</td>
<td>698.94</td>
<td>426.17</td>
<td>28.11</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1063.2</td>
<td>1150.0</td>
<td>1154.4</td>
<td>652.06</td>
<td>640.54</td>
<td>639.08</td>
<td>381.76</td>
<td>33.39</td>
<td>61.11</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>754.70</td>
<td>729.70</td>
<td>788.04</td>
<td>793.39</td>
<td>365.88</td>
<td>372.23</td>
<td>378.40</td>
<td>34.88</td>
<td>442.42</td>
<td>460.65</td>
<td>416.60</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>787.03</td>
<td>815.68</td>
<td>821.13</td>
<td>400.63</td>
<td>407.21</td>
<td>413.42</td>
<td>44.02</td>
<td>433.05</td>
<td>449.62</td>
<td>409.00</td>
<td>35.05</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>860.52</td>
<td>836.14</td>
<td>886.11</td>
<td>891.65</td>
<td>472.70</td>
<td>478.23</td>
<td>483.98</td>
<td>97.50</td>
<td>397.81</td>
<td>410.78</td>
<td>377.84</td>
<td>107.20</td>
<td>73.56</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>849.71</td>
<td>825.93</td>
<td>871.94</td>
<td>877.52</td>
<td>465.75</td>
<td>472.30</td>
<td>478.44</td>
<td>96.63</td>
<td>416.33</td>
<td>429.45</td>
<td>396.12</td>
<td>100.07</td>
<td>65.14</td>
<td>18.76</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>773.30</td>
<td>752.26</td>
<td>781.97</td>
<td>787.72</td>
<td>411.19</td>
<td>422.83</td>
<td>430.79</td>
<td>119.04</td>
<td>506.70</td>
<td>521.82</td>
<td>484.02</td>
<td>91.32</td>
<td>78.82</td>
<td>117.07</td>
<td>99.08</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

**Distance classes (km)**

<table>
<thead>
<tr>
<th>Distance classes (km)</th>
<th>0.00</th>
<th>40.60</th>
<th>210.25</th>
<th>208.28</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤&lt;39.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤208.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤397.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤429.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤521.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤752.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤860.52</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤1211.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table A3.3. *T. repens* population distance matrix. Colours represent the 7 distance classes.

|          | IOS1 | IOS2 | IOS3 | IOS4 | IOS5 | IOS6 | RYE1 | LKD1 | LKD2 | LKD3 | LKD4 | LKD5 | NWS1 | NWS2 | NWS3 | SKY1 | SKY2 | SKY3 | STK1 | STK2 | STK3 | STK4 | STK5 | SHT1 | SHT2 | SHT3 | SHT4 | SHT5 |
|----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| 0        |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 5        |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 4        |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 237      | 239  | 241  | 0    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 276      | 278  | 280  | 38   | 0    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 387      | 389  | 378  | 390  | 379  | 426  | 46   | 12   | 0    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 387      | 389  | 378  | 390  | 379  | 426  | 46   | 12   | 0    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 387      | 389  | 378  | 390  | 379  | 426  | 46   | 12   | 0    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 387      | 389  | 378  | 390  | 379  | 426  | 46   | 12   | 0    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 387      | 389  | 378  | 390  | 379  | 426  | 46   | 12   | 0    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 387      | 389  | 378  | 390  | 379  | 426  | 46   | 12   | 0    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 387      | 389  | 378  | 390  | 379  | 426  | 46   | 12   | 0    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Distance classes (km) | ≤39 | ≤477 | ≤1209 | ≤165 | ≤749 | ≤373 | ≤864 |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |

245
Appendix 4: TIGR TRANSCRIPT ASSEMBLIES DATABASE

TIGR plant transcript assemblies (TA) are developed from existing EST sequences and sequences available in the NCBI Genbank nucleotide database. TAs are produced using the TIGCL tool (Pertea et al., 2003) consisting of sequence clustering, performed using a modified version of Megablast (Zhang et al., 2000), and assembling using CAP3 (Huang & Madan, 1999). Criteria for assembly include a minimum 50bp match, which must contain at least 95% identity in the aligned region (Childs et al., 2007). Plant transcript assembly diagrams for the loci used in this study are detailed below, with all images taken from the TIGR plant transcript assemblies database (http://plantta.jcvi.org).

Specific ESTs for further analysis were selected from each TA dependent on preliminary amplification assessments and prior information on repeat sequences within the locus (see below).

A4.1 TA1010_57577 TRANSCRIPT ASSEMBLY

One putative microsatellite site is detailed in the TA database, from 76-91 bp so BB925453, BB924456 and BB924106 were excluded from further analysis to avoid the problems associated with the sequencing of heterozygous microsatellite regions.

![Figure A4.1. TA1010_57577 transcript assembly](image-url)

Figure A4.1. TA1010_57577 transcript assembly
A4.2  **TA1548_57577 TRANSCRIPT ASSEMBLY**

Two putative microsatellites are described in the TA database, from 96-107bp and 790-799bp. BB924422 was selected to avoid amplifying these regions to avoid the problems associated with the sequencing of heterozygous microsatellite regions.

![Figure A4.2. TA1548_57577 transcript assembly](image)

A4.3  **TA3078_57577 TRANSCRIPT ASSEMBLY**

No putative microsatellites are described for this locus in the TA database.

![Figure A4.3. TA3078_57577 transcript assembly](image)

A4.4  **TA3695_57577 TRANSCRIPT ASSEMBLY**

No putative microsatellites are described for this locus in the TA database.

![Figure A4.4. TA3695_57577 transcript assembly](image)
A4.5  **TA555_57577 TRANSCRIPT ASSEMBLY**

Two putative microsatellite sites are detailed in the TA database, from 995-1012bp and 1088-1099bp. Consequently BB916867, BB912536, BB914314, BB910324, BB922925, BB908687, BB903117 and BB914174 were excluded from further analysis to avoid the problems associated with the sequencing of heterozygous microsatellite regions.

![Diagram of TA555_57577 transcript assembly](image)

**Figure A4.5. TA555_57577 transcript assembly**

A4.6  **TA989_57577 TRANSCRIPT ASSEMBLY**

No putative microsatellites are described for this locus in the TA database.

248
Figure A4.6. TA989_57577 transcript assembly

REFERENCES


Appendix 5: PRIMER CHOICE

Primer pairs were designed using Primer3 v0.4.0 (Rozen & Skaletsky, 2000). Default settings were used with the exception of product size, which was maximised where possible.

BB910055 T. PRATENSE PRIMER DESIGN

The primer pair used to amplify the selected loci is outlined below in Figure A5.1, with the characteristics summarised in Table A5.1.

<table>
<thead>
<tr>
<th>Start</th>
<th>Length</th>
<th>$t_m$</th>
<th>GC%</th>
<th>A</th>
<th>3'</th>
<th>Sequence 5' – 3'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left primer</td>
<td>20</td>
<td>20</td>
<td>60.34</td>
<td>50</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Right primer</td>
<td>504</td>
<td>20</td>
<td>60.19</td>
<td>40</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

Pair complementarity: 4 3

Total sequence size: **580 bp**  
Product size: 485 bp

Figure A5.1: Primer pair design for BB910055. Bases highlighted in grey denote primer sites.

Table A5.1: Primer3 output for BB910055.  

- Melting temperature of oligonucleotide:  
- Percentage of G and C nucleotides in sequence:  
- Self-complementarity score based on self annealing and secondary structures:  
- Self-complementarity score based on tendency to form primer dimer with itself.

BB910055 T. DUBIUM PRIMER DESIGN

Sequence data from species within the Leguminosae family were aligned to reveal conserved nucleotide regions (Figure A5.2). T. pratense primer pairs (see above) were checked against...
the alignment to ensure these primer sequences were not found in large regions of unconserved nucleotides. In this case the primer pair outlined above for *T. pratense* was successful when used in *T. dubium*.

```
BB910055_Trifolium_pratense
Medicago truncatula_AC144540
Medicago sativa_AF322116
Vicia faba_Z56278
Glycine max_EU039964
Lotus japonicus_AP004498

---

BB910055_Trifolium_pratense
Medicago truncatula_AC144540
Medicago sativa_AF322116
Vicia faba_Z56278
Glycine max_EU039964
Lotus japonicus_AP004498

ATGAG-CAGACTGTCTATGGCAAGACATGGTCAAGATGATATTT
ATGACAGACTGTCTATGGCAAGACATGGTCAAGATGATATTT
AACAGACTGTCTATGGCAAGACATGGTCAAGATGATATTT
CTGAGACTGTCTATGGCAAGACATGGTCAAGATGATATTT
ATGACAGACTGTCTATGGCAAGACATGGTCAAGATGATATTT

BB910055_Trifolium_pratense
Medicago truncatula_AC144540
Medicago sativa_AF322116
Vicia faba_Z56278
Glycine max_EU039964
Lotus japonicus_AP004498

* *** ****** ******** ****** **** ** ** **

BB910055_Trifolium_pratense
Medicago truncatula_AC144540
Medicago sativa_AF322116
Vicia faba_Z56278
Glycine max_EU039964
Lotus japonicus_AP004498

* *** ****** ******** ****** **** ** ** **

BB910055_Trifolium_pratense
Medicago truncatula_AC144540
Medicago sativa_AF322116
Vicia faba_Z56278
Glycine max_EU039964
Lotus japonicus_AP004498

GAATCAGAAGAATCACCTGGTGATTCATTGAGAGATATACATGACTTATC
GAATCAGAAGAATCACCTGGTGATTCATTGAGAGATATACATGACTTATC
GAATCAGAAGAATCACCTGGTGATTCATTGAGAGATATACATGACTTATC
GAATCAGAAGAATCACCTGGTGATTCATTGAGAGATATACATGACTTATC
GAATCAGAAGAATCACCTGGTGATTCATTGAGAGATATACATGACTTATC

BB910055_Trifolium_pratense
Medicago truncatula_AC144540
Medicago sativa_AF322116
Vicia faba_Z56278
Glycine max_EU039964
Lotus japonicus_AP004498

* *** ****** ******** ****** **** ** ** **

BB910055_Trifolium_pratense
Medicago truncatula_AC144540
Medicago sativa_AF322116
Vicia faba_Z56278
Glycine max_EU039964
Lotus japonicus_AP004498

* *** ****** ******** ****** **** ** ** **

BB910055_Trifolium_pratense
Medicago truncatula_AC144540
Medicago sativa_AF322116
Vicia faba_Z56278
Glycine max_EU039964
Lotus japonicus_AP004498

* *** ****** ******** ****** **** ** ** **

BB910055_Trifolium_pratense
Medicago truncatula_AC144540
Medicago sativa_AF322116
Vicia faba_Z56278
Glycine max_EU039964
Lotus japonicus_AP004498

ATGATAATCTTTCGATCCAAATGGGAATCCCGCATGGAAATGCAACCATG
ATGATAATCTTTCGATCCAAATGGGAATCCCGCATGGAAATGCAACCATG
"
Figure A5.2: Multiple alignment of BB910055 EST with potential homologous sequences from related species (within the Leguminosae). Sequences are aligned in order of input. Grey highlighted bands denote primer sequences used for further analysis. Sequences were obtained using BLAST v2.2.21 (Zhang et al., 2000) and aligned using ClustalW v2.0.10 (Larkin et al., 2007).
BB925852 *T. pratense* PRIMER DESIGN

The primer pair used to amplify the selected loci is outlined below in Figure A5.3, with the characteristics summarized in Table A5.2.

1 GAAAGCAAAACCAACGCTCGGCTATAGAGAAGAGGTCGAGTACCGTGATCGTGATCT
61 TACAAAGGCAGATCTGGAGATGGTGACTAGACTAGATCCACTTCGCGTGTATCCTTATAG
121 ATATCGAGCCGAGATTTGTGAGAAGCAACCATAGAGAAACGAAGCCATTGCTGAGCTATC
181 TAGAGCAATTGCATTTAAAGCCTGATTGGCACACCTTCTACATCTCAGGCTATGAGCT
241 TAGAGCAATTGCATTTAAAGCCTGATTGGCACACCTTCTACATCTCAGGCTATGAGCT
301 AAACCACCAAGAAATGTTGGAACTTCACACTCGTTAATAGCCATGAAAGCCTGAGTTGA
361 GTTTGCATGCTATATGAATGAAGACGACAACATTTGTACACTCATCAGGCTATGAGCT
421 AAATTAAATTGTAAATGAAAGATGATAGCTATGAAACTGATAGTTAGAGCAGCAGTTATAC
481 TCTGTGTATGCAATGATCATATTTCCTTTCTTCTAAAAGGGGATGCCCAAATTTTTGTAAT
541 TTATATCTCTGATGTAATGGATGATGAGCT

Figure A5.3: Primer pair design for BB925852. Bases highlighted in grey denote primer sites.

Table A5.2: PRIMER3 output for BB925852. *a* Melting temperature of oligonucleotide; *b* Percentage of G and C nucleotides in sequence; *c* Self-complementarity score based on self annealing and secondary structures; *d* Self-complementarity score based on tendency to form primer dimer with itself.

<table>
<thead>
<tr>
<th></th>
<th>Start</th>
<th>Length</th>
<th>$t_m^a$</th>
<th>GC%$^b$</th>
<th>A$^c$</th>
<th>3$^d$</th>
<th>Sequence 5$’$ – 3$’$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Left primer</strong></td>
<td>20</td>
<td>20</td>
<td>60.36</td>
<td>55</td>
<td>3</td>
<td>2</td>
<td>TCGGCGTATAGAGAAGGAGTC</td>
</tr>
<tr>
<td><strong>Right primer</strong></td>
<td>532</td>
<td>20</td>
<td>60.49</td>
<td>40</td>
<td>4</td>
<td>2</td>
<td>ATTTGGCATTCCCTTTTTTA</td>
</tr>
</tbody>
</table>

Pair complementarity 4 2

Total sequence size: **572 bp**

Product size: 513 bp

BB925852 *T. dubium* PRIMER DESIGN

As only somewhat similar sequences were found when using BLAST, alignments produced contained little conserved sequence. Hence, the primer pair outlined above for *T. pratense* was trialled with *T. dubium* using lower annealing temperatures, which proved successful in amplifying the desired sequence.
TA1548_57577 *T. PRATENSE* PRIMER DESIGN (USING BB915621)

The primer pair used to amplify the selected loci is outlined below in Figure A5.4, with the characteristics summarized in Table A5.3.

Figure A5.4: Primer pair design for BB915621. Bases highlighted in grey denote primer sites.

Table A5.3: PRIMER3 output for BB915621. $^a$ Melting temperature of oligonucleotide: $^b$ Percentage of G and C nucleotides in sequence: $^c$ Self-complementarity score based on self annealing and secondary structures: $^d$ Self-complementarity score based on tendency to form primer dimer with itself.

<table>
<thead>
<tr>
<th></th>
<th>Start</th>
<th>Length</th>
<th>$t_m^a$</th>
<th>GC$^b$</th>
<th>A$^c$</th>
<th>3$^d$</th>
<th>Sequence 5’ – 3’</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Left primer</strong></td>
<td>13</td>
<td>23</td>
<td>59.87</td>
<td>34.78</td>
<td>3</td>
<td>3</td>
<td>TTCACTTCATTGCTTTTCATCAGA</td>
</tr>
<tr>
<td><strong>Right primer</strong></td>
<td>481</td>
<td>20</td>
<td>60.14</td>
<td>50</td>
<td>6</td>
<td>3</td>
<td>TTGCTGCAGAGATTTCGTGAC</td>
</tr>
<tr>
<td>Pair complementarity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4 2</td>
</tr>
</tbody>
</table>

Total sequence size: **580 bp**

Product size: 469 bp

**TA1548_57577 T. DUBIUM PRIMER DESIGN (USING BB915621)**

As only somewhat similar sequences were found when using BLAST, alignments produced contained little conserved sequence. Hence, the primer pair outlined above for *T. pratense* was trialled with *T. dubium* using lower annealing temperatures.
TA3695_57577 *T. pratense* primer design (using BB916074)

The primer pair used to amplify the selected loci is outlined below in Figure A5.5, with the characteristics summarized in Table A5.4.

```
1    ATGAATCATCTTTGCTTTGTAAGAGGTTTGAAGTACATGATATACTGCTTTTCTTGC
61   AAGCTCGTCTTTACGTTATTGGACGAGCTCATGGTTTTGTAACCAATGGAATTTCCTCT
121  CGCGTACTTGAAGAATCTCGTCTCTCCAAATATCTGCTCAAGAAGC
181  TAGATCAATGTCAAGCAATTCAAATATAGGGTTTTCATATCCAGAAGCAGC
241  TTTAGTTCCAGACCCCGATGCGACTGAAATATGCGAAGTACATGGTTATGAGTTG
301  AGTTTTGACATGAAATAGCTTTTCTTCTCAAATTTAATCATATGCTTCTCTTCC
361  CATCGTACCACGCGCTTTTATTTGATAGTAATTCATCTTACTCATGTTGCC
421  ATGTAGTGCAAGAAGTACACTAAATTAGTTGTTAGCAAAATCGACTGCAATGTTG
481  TGTGCTGACACGAGTGGAGTACAGCTCAACAAATGCTTTGACACCTTCTCC
541  CTTTTGACACCAATCTTCTTTCTCCACCCATCT
```

Figure A5.5: Primer pair design for BB916074. Bases highlighted in grey denote primer sites.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Start</th>
<th>Length</th>
<th>t_m</th>
<th>GC%</th>
<th>A</th>
<th>3'</th>
<th>Sequence 5' – 3'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left primer</td>
<td>47</td>
<td>20</td>
<td>60.66</td>
<td>50</td>
<td>4</td>
<td>2</td>
<td>TGTCCTTTTCTTGCAGAAGCTC</td>
</tr>
<tr>
<td>Right primer</td>
<td>519</td>
<td>20</td>
<td>59.32</td>
<td>55</td>
<td>4</td>
<td>3</td>
<td>CGATTGTACCTGGCGACTCAC</td>
</tr>
</tbody>
</table>

Pair complementarity

Total sequence size: **572 bp**  
**Product size**: 473 bp

TA3695_57577 *T. dubium* primer design (using BB916074)

As only somewhat similar sequences were found when using BLAST, alignments produced contained little conserved sequence. Hence, the primer pair outlined above for *T. pratense* was trialled with *T. dubium* using lower annealing temperatures, which proved successful in amplifying the desired region.
TA989_57577 T. pratrese Primer Design (using BB906196)

The primer pair used to amplify the selected loci is outlined below in Figure A5.6, with the characteristics summarized in Table A5.5.

1 AATCCAGAAAACATCCTTATCCTTATAAACATCAAACATAAAAATCCTTAAAAAGGTTTTT
61 GACAAGACAAATGATACCTAATAAAAACCTATAAATAATTTATTTTGGCCACAGAGAAT
121 AAAACTCCACAGCTTAAATACAGGTTAATCTTGTCTTTGTCTTGATTTCCATCAACAATA
181 TAAAAATAAAACCAACATCCTGAACAACAAAAACAAATTTAATTACCACCTTTGTAACAAACAA
241 AATGAACCAAACCGAATGTTAAACTGAAAGTTACCTTCTAGTTATTTACATAACAT
301 GTGAGGCCAAAGTTGTTCAAGAAAACCTCCGCTGAGTTGAAGCCAGGTCATCAAAAGCTCC
361 AAAGGTTTCTTGTTTCTCAGCATCCTGAGTTACATCAGCACCGAGCAAAGACGCAAGTG
421 AAGCATCAXCTCCAGTTATCATCAGAATGAGTTCTCTCAGGAGAAGTCCAGCTGTGATAT
481 CTGAAAGCTTTCCAGAGAGTGTTTTCCTGAGAGTTTGACATAGTTAGCATATCCTGAG
541 AGTTGGTAGACTGCACTGCATCTGAGGGTTT1

Figure A5.6: Primer pair design for BB906196. Bases highlighted in grey denote primer sites.

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Start</th>
<th>Length</th>
<th>t_m</th>
<th>GC%</th>
<th>A</th>
<th>3' d</th>
<th>Sequence 5' – 3'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left primer</td>
<td>45</td>
<td>25</td>
<td>59.96</td>
<td>32</td>
<td>8</td>
<td>3</td>
<td>CCTTAAAAAGGTTTGGACAAGACA</td>
</tr>
<tr>
<td>Right primer</td>
<td>522</td>
<td>20</td>
<td>60.28</td>
<td>45</td>
<td>5</td>
<td>2</td>
<td>TGCAAGCTGAATCTGCAAGAG</td>
</tr>
</tbody>
</table>

Pair complementarity 4 2

Total sequence size: **570 bp**  Product size: 478 bp

TA989_57577 T. dubium Primer Design (using BB906196)

Sequence data from species within the Leguminosae family were aligned to reveal conserved nucleotide regions (Figure A5.7). T. pratense primer pairs (see above) were checked against the alignment to ensure these primer sequences were not found in large areas of unconserved nucleotides. In this case the primer pair outlined above for T. pratense was successful when used in T. dubium.
**Primer choice**

<table>
<thead>
<tr>
<th>Accession</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB906196_Trifolium_pratense</td>
<td>AATCCAGAACCATCCTTATCTCTTATAAACCATCAAACATAAAAAATCTTTAA</td>
</tr>
<tr>
<td>Trifolium_pratense_AB236754</td>
<td>--CCATAAAACAT-------AACCTTTAATACATCACTAACTAAATAATCTATCAA</td>
</tr>
<tr>
<td>Medicago_sativa_EF462215</td>
<td>---------------------------------CACTATAGGTTTTTTTTTGACCATCTGAT</td>
</tr>
<tr>
<td>Galega_orientalis_FJ223566</td>
<td></td>
</tr>
<tr>
<td>BB906196_Trifolium_pratense</td>
<td>AAAAGTTTTTGGACTGACAATATATGATACCTAAATAAAAAATCTTTATAAATCTTTAA</td>
</tr>
<tr>
<td>Trifolium_pratense_AB236754</td>
<td>-----------------------------AATAAAAACCTATAAATAATT</td>
</tr>
<tr>
<td>Medicago_truncatula_AC151460</td>
<td>AA-GGGTTTTTGACAAGACAAAAGG--ATTACCTAAGACTATAAAATATT</td>
</tr>
<tr>
<td>Medicago_sativa_EF462215</td>
<td>-----CCATAAACATA---AACCTTATAAACATCAAACTTAAAAATCCTCAA</td>
</tr>
<tr>
<td>Galega_orientalis_FJ223566</td>
<td></td>
</tr>
<tr>
<td>BB906196_Trifolium_pratense</td>
<td>CTGTGAAAACAAAACAAA----------------ATTTAATTACCTCACTCGATTAA</td>
</tr>
<tr>
<td>Trifolium_pratense_AB236754</td>
<td>ATCTGCGAACAAACAAATAGATTAAAAACAAAATAAATCTAAACAAATACAAATACAAATACAA</td>
</tr>
<tr>
<td>Medicago_truncatula_AC151460</td>
<td>ATCCCGGAAATATCTAAATCTAAATCTAAATCTAAATCTAAATCTAAATCTAAATCTAAATCTAA</td>
</tr>
<tr>
<td>Medicago_sativa_EF462215</td>
<td>ATCCTGAAACAATATCTAAATCTAAATCTAAATCTAAATCTAAATCTAAATCTAAATCTAAATCTAA</td>
</tr>
<tr>
<td>Galega_orientalis_FJ223566</td>
<td></td>
</tr>
<tr>
<td>BB906196_Trifolium_pratense</td>
<td>TATTTTTGCCACAAGAGAATAAATATTGTTTTTGACAAGACACAAATACAAATACAAATACAAATACAA</td>
</tr>
<tr>
<td>Trifolium_pratense_AB236754</td>
<td>TATTTTTGCCACAAGAGAATAAATATTGTTTTTGACAAGACACAAATACAAATACAAATACAAATACAA</td>
</tr>
<tr>
<td>Medicago_truncatula_AC151460</td>
<td>TATTTTTGCCACAAGAGAATAAATATTGTTTTTGACAAGACACAAATACAAATACAAATACAAATACAA</td>
</tr>
<tr>
<td>Medicago_sativa_EF462215</td>
<td>TATTTTTGCCACAAGAGAATAAATATTGTTTTTGACAAGACACAAATACAAATACAAATACAAATACAA</td>
</tr>
<tr>
<td>Galega_orientalis_FJ223566</td>
<td></td>
</tr>
</tbody>
</table>

**A5**

---

257
Figure A5.7: Multiple alignment of BB906196 EST with potential homologous sequences from related species (within Leguminosae). Sequences are aligned in order of input. Grey highlighted bands denote primer sequences used for further analysis. Sequences were obtained using BLAST version 2.2.21 (Zhang et al., 2000) and aligned using ClustalW version 2.0.10 (Larkin et al., 2007).

REFERENCES

