The University Of Sheffield.

Phylogeography of littorinid snails

By:

Mauricio Montaño-Rendón

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield

Faculty of Sciences

Department of Animal and Plant Sciences

June 2016

Table of contents

Summary

Interactions between evolutionary forces such as natural selection, genetic drift and gene flow are complex. Natural selection can lead to parallel formation of phenotypes under similar environmental conditions. Phylogenetic relationships can be inferred from the accumulation of genetic variation caused by genetic drift, regardless of phenotypes. Gene flow between populations can sometimes facilitate the formation of species by natural selection.

In this thesis, two groups of marine snails in the diverse subfamily Littorininae were studied. The three rocky-shore species *Littorina saxatilis*, *Littorina arcana* and *Littorina compressa* form one such group, whereas the two sister-species *Littoraria cingulata* and *Littoraria filosa* form the other group. Previous studies of the *L. saxatilis* complex have shown high levels of phenotypic and genetic diversity both at local and broader scales. Previous studies of *L. cingulata* and *L. filosa* have found some of the typical signatures of reinforcement.

Chapters II and III focused on analysing morphological and genetic variation, respectively, within and among species in the *Littorina saxatilis* complex from the British Isles. Geometric morphometrics analyses revealed a diversity of shell shape among species, but especially so within *L. saxatilis*. Shell shape was better explained by environment rather than by geography. Molecular data obtained by high-throughput targeted capture showed the opposite pattern, i.e. genetic variation showed a strong phylogeographic pattern.

Chapter IV focused on testing whether reinforcement had contributed towards speciation between *Littoraria cingulata* and *Littoraria filosa*. Approximate Bayesian computation analysis supported absence of gene flow between the species in sympatry, suggesting that reinforcement did not contribute towards speciation. However, the results need to be validated and more complex models tested.

This thesis highlights the relevance of marine snails in the subfamily Littorininae as model species for addressing a wide range of evolutionary questions. It also provides a wealth of data for many potential follow-up studies.

Chapter I

General introduction

Evolution has many times been considered the unifying theory in biology, casting light on a wide range of subjects from molecular biology to ecology (Futuyma 2013). Darwin (1859) suggested that the ancestry of all kinds of organisms could theoretically be tracked back in time to a common origin of life, which meant that organisms making up the biodiversity observed at any given moment would necessarily be related to each other. Ever since then, there have been countless studies suggesting hypotheses about the phylogeny of all forms of life (e.g. Murphy *et al.* 2001). This task often leads biologists to look deeper into the internal population structure of species, the study of which provides important information about the processes generating and maintaining biological diversity.

The genetic material of organisms offers a vast source of information for addressing the aforementioned questions about the origin of biodiversity, which has been shaped through time by a combination of different evolutionary processes such as mutation, genetic drift, and natural selection, among others. Phylogeographic studies have used this information specifically to try to understand the spatial distribution of different genetic lineages mainly within species (Avise 2009). Contributions from this field constitute integral information to better understand how speciation works, a process whereby new species are formed through the evolution of reproductive isolation between divergent populations (Coyne & Orr 2004).

Butlin *et al*. (2008) have suggested that the traditional way of studying speciation by considering only its spatial component has been unsatisfactory because of its reductionist assumptions. For this reason, they proposed that a more holistic approach should be used that considers “the balance between local adaptation and gene flow, the interaction between components of reproductive isolation, and the genetic basis of differentiation”. The generation of such genetic data – which is necessary to embrace this approach – is becoming increasingly accessible due to technological progress (Davey *et al.* 2011, and see “DNA sequencing” section in this chapter), whereby genotyping

a large number of individuals at millions of positions along the genome is now a common reality. Limitations now lie in the computational power to analyse these data.

Adaptation is an important component in the study of speciation (Butlin *et al.* 2008), and genomic analyses have recently proved to be enlightening as to this matter (Wood *et al.* 2008; Duforet-Frebourg *et al.* 2014). Additionally, morphological studies have an essential role in understanding adaptation when combined with genomic studies. An especially relevant contribution made by such studies comes from the fact that many of them have found strong correlations between organismal form and environment with repeated and independent origins, which might suggest parallel evolution by local adaptation. Some examples include the independent divergence in the length of gill rakers in the three-spine stickleback *Gasterosteus aculeatus* from different lake–stream systems (Lucek *et al.* 2013); the repeated evolution of carapace form in different species of freshwater turtles in the family Emydidae from fast- and slow-flowing aquatic environments (Rivera *et al.* 2014), and the parallel evolution of shell shape in the intertidal snail *Littorina saxatilis* from crab-rich and wave-exposed habitats (reviewed by Johannesson *et al.* 2010).

Local adaptation tends to cause divergence into different populations, while gene flow tends to have the opposite effect of homogenising them. Therefore, studying the balance between these two elements contributes to the understanding of speciation (Butlin *et al.* 2008). A widely used approach to study the subject is by identifying genes or genomic regions that are divergent in the face of gene flow. This approach involves conducting genome scans, which is a method to sample genotypic data across whole genomes. Comparing the degree of differentiation across these genomic regions allows the identification of outlier loci, which are characterised by showing greater differentiation than the neutral background. Examples of identification of outlier loci between divergent populations come from – but are not limited to – grass (Gray *et al.* 2014), bees (Chávez-Galarza *et al.* 2013), stick insects (Soria-Carrasco *et al.* 2014), fish (Perrier *et al.* 2013), and snails (Westram *et al.* 2014), among many others. However, the power to detect local adaptation using genome scans depends on the sampling design and statistical methods employed (Lotterhos & Whitlock 2015). In order to identify such outlier loci, it is

first needed to model the neutral background differentiation, and this can only be achieved once a good historical demographic model for the study populations is available.

With the increasing rate of genomic data production comes, as well, the necessity for appropriate analytical methods that can handle such large datasets. This area has benefited from both theoretical and technical advances. Approximate Bayesian computation (ABC) is a model-based approach that offers a powerful tool to test different demographic models based on coalescent simulations of data with the same attributes as those of the original data (Beaumont 2010). An important attribute of this approach is the parameterisation and estimation of historic demographic features such as population origins, time since divergence, effective population size, bottlenecks, expansions and gene flow, which are all of interest in the fields of population genomics, phylogeography and speciation (e.g. Butlin *et al.* 2014; Pelletier & Carstens 2014; Rittmeyer & Austin 2015).

This study

A much better understanding of how evolution works at the molecular level has followed the technological advances that have facilitated the acquisition of genomic data (Davey *et al.* 2011). This has been particularly important for non-model organisms, for which little or no previous knowledge existed (e.g. Davey & Blaxter 2010). This study makes use of genomic tools to investigate the evolution of two different groups of snails in the subfamily Littorininae. It also makes the most of quantitative methods to analyse shape variation in the context of local adaptation. Thus, the present study further enhances our understanding of local adaptation, phylogeography, and speciation in a subfamily of snails with species that are becoming model systems.

Chapter II is focused on the morphometric analysis of shell shape variation within and among species of intertidal snails in the *Littorina saxatilis* complex, a system renowned for its great shell diversity. The novelty of the study lies in the extensive sampling that was carried out in the British Isles across more locations than any other previous work. It also explores in greater detail the

continuous shell shape variation that exists beyond the discrete phenotypic variation so far described (Conde-Padín *et al.* 2007).

In Chapter III emphasis is placed on the demographic history of *Littorina saxatilis* across the same sites studied in Chapter II by using genomic data to infer the possible routes of colonisation after the Last Glacial Maximum. Apart from the extensive sampling already mentioned, an additional quality of this study is the use of synthetic probes to capture and sequence specific targeted genomic regions. The documentation and analysis of such detailed genetic variation surely serves as a platform to address questions about speciation in a group with repeated and independent evolution of similar ecotypes both at a local and global scale.

Chapter IV, while still being focused on speciation in marine snails, slightly shifts the attention onto the pair of sister species of snails *Littoraria cingulata* and *Littoraria filosa* from the mangrove forests of northern Australia to address the fundamental question of whether reinforcement strengthened the reproductive isolation in this subfamily of snails once they had come into secondary contact after allopatric divergence. This final chapter uses genomic data to compare allopatric and sympatric populations of the two species, especially looking for past gene flow between them.

The study organisms

Sea snails in the subfamily Littorininae (Mollusca: Gastropoda), which comprises 152 recognised species (summarised by Reid *et al.* 2012), are common inhabitants of rocky shores and mangroves throughout the world (Reid 1986, 1996). The production of pelagic egg capsules in many species of this group may have allowed the colonisation of harsh oceanic and continental environments (Reid 1986, 1996). Among the 18 recognised species within the genus *Littorina*, *Littorina saxatilis*, which is distributed along the rocky shores of the North Atlantic (Reid 1996), is the only one to bear crawling offspring instead of laying eggs, which has been suggested to restrict its dispersal (Janson 1983). However, this lack of pelagic larvae combined with highly promiscuous females (Mäkinen *et al.* 2007; Panova *et al.* 2010) may allow drifted females to colonise and establish new populations in remote areas (Johannesson 1988).

This could be possible because promiscuous gravid females drifting and arriving to new areas would produce offspring from various males, which, given the low mobility of the species, could establish and maintain a population, like the one in Rockall, an islet 430 km north-west of Ireland (Johannesson 1988).

The species forms part of the *Littorina saxatilis* complex (Fig. I.1). This complex also includes *Littorina arcana* and *Littorina compressa* (Reid 1996; Reid *et al.* 2012), which have a more restricted distribution and are usually found at sites in the UK where *L. saxatilis* is present too. Phylogenetic relationships among the three species have been studied extensively, but have resulted in a still unresolved trichotomy (Reid *et al.* 2012; Panova *et al.* 2014). *L. arcana* and *L. compressa* females are characterised by a prominent jelly glad whose function is to embed the egg capsules together in a sticky layer that helps them attach to the substrate, such as rock or driftwood (Hannaford-Ellis 1979; Buckland-Nicks & Chia 1990). *L. saxatilis*, on the other hand, has a brood pouch where embryos develop and are born as miniature crawling snails. The shape of their shells has been found to be different in all three species (Conde-Padín *et al.* 2007b). Within *L.* saxatilis, different phenotypes have been described based on shell traits and habitat (summarised by Reid 1996 and Johannesson *et al.* 2010), and many studies support the hypothesis of repeated and independent formation of these ecotypes in different locations (Quesada *et al.* 2007; Mäkinen *et al.* 2008; Butlin *et al.* 2014; Westram *et al.* 2014).

In the UK, the following four ecotypes have been described (Fig. I.1): (1) the M ecotype, the biggest of all, is mainly found among boulders at a lower tidal level where crabs may be common, their shell is thick and the aperture is rather narrow; (2) the H ecotype, which is smaller than M, is found at a higher tidal level where rocks are more exposed to the waves, their shell is thinner and the aperture rather big compare to body size; (3) the *neglecta* ecotype, the smallest of all, lives inside empty barnacle shells at lower tidal levels where the wave action is strong, and (4) the *tenebrosa* ecotype, which is mainly found in brackish coastal lagoons with little or no connection to the sea, is characterised for having a tall spire and thin walls.

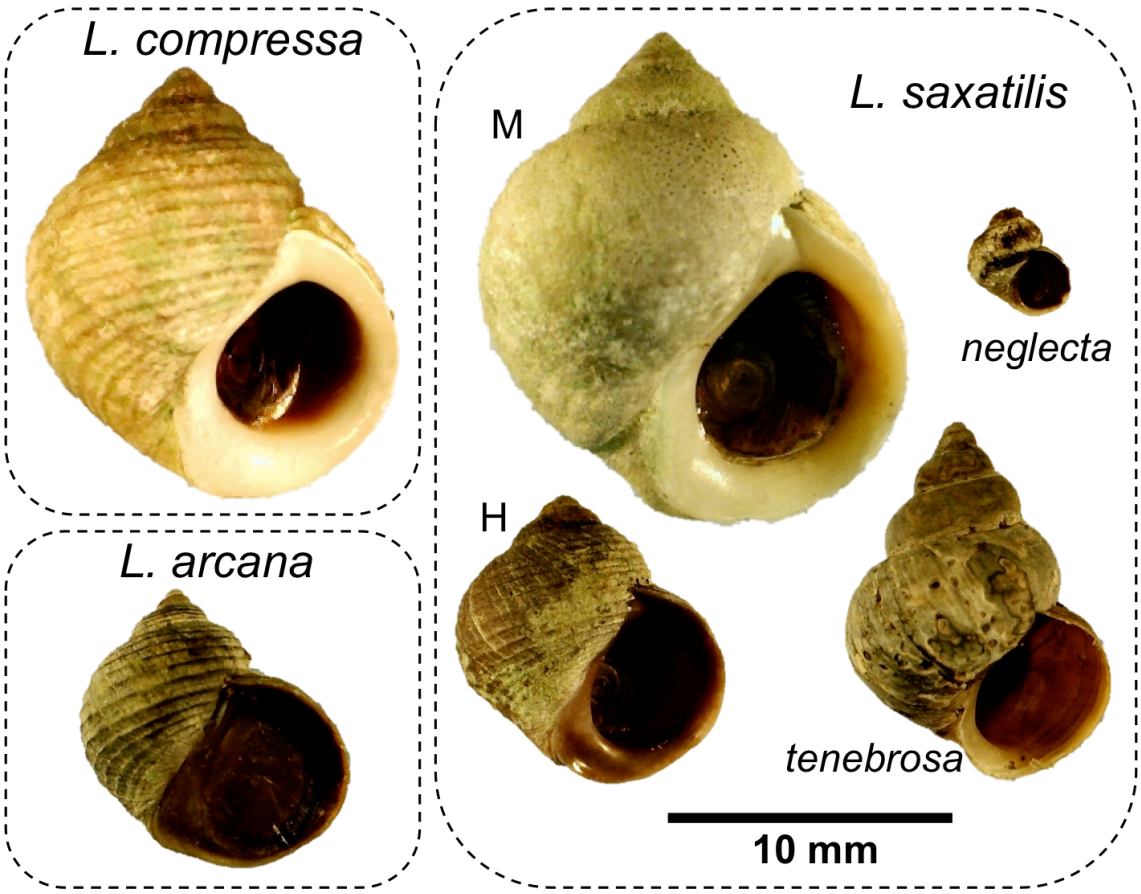


Fig. I.1 Species in the *Littorina saxatilis* complex. *L. compressa* from Oban, Argyll and Bute, Scotland. *L. arcana* from Amble, Northumberland, England. *L. saxatilis* ecotype M from Isle of Mull, Argyll and Bute, Scotland. *L. saxatilis* ecotype H from St. Abbs, Berwickshire, Scotland. *L. saxatilis* ecotype neglecta from Old Peak, North Yorkshire, England. *L. saxatilis* ecotype *tenebrosa* from Loch Aibhnin, Galway, Ireland.

Phylogeographic studies using mitochondrial DNA of snails in the *L. saxatilis* complex have suggested that *L. saxatilis* originated around the British Isles, and that the observed strong population structure in the Northeast Atlantic possibly was the result of colonisation from different glacial refugia (Doellman *et al.* 2011; Panova *et al.* 2011). These findings have served as a strong basis to address questions about the genetic basis of parallel phenotypic divergence (Westram *et al.* 2014) and parallel evolution of local adaptation (Butlin *et al.* 2014) across different locations in Europe that have been colonised at different times. Even more recently, a study combining morphometrics and genotypic data has identified an association between outlier loci and key adaptive phenotypic traits (Hollander *et al.* 2015).

The second system treated is this study, also in the subfamily Littorininae, is the pair of sister species *Littoraria cingulata* and *Littoraria filosa* from the mangrove forests of Australia (Reid 1986, 2001). Among the 37 species within *Littoraria*,

this pair of sister species is the youngest, having diverged in allopatry during the Pleistocene 2.59 Mya (Reid *et al.* 2012), separated by a current estimated mitochondrial genetic distance of 4.15% (Reid *et al.* 2010). These species, whose current geographic ranges overlap by 50% of the smaller range (*L. cingulata*), undergo a planktotrophic development, which means that individuals feed on plankton during their larval stage (Reid 1986). Morphological and genetic data have shown evidence of geographic structure within both species (Reid 1986, 2001; Johnson & Black 1998; Inness-Campbell *et al.* 2003; Stuckey 2003).

A comparison of penial similarity in 40 pairs of sister species across the whole subfamily Littorininae, whose overlap of geographic distributions ranges from none to complete, found that *L. cingulata* and *L. filosa* have an overall penial similarity of 92% (Hollander *et al.* 2013). This finding was part of a more general pattern observed in that study, which found that penis similarity between pairs of sister species was lower for those pairs with range overlap than for those pairs with no range overlap. These findings have led to current investigations about reproductive character displacement – a pattern of variation of a mating trait with greater divergence in sympatry than in allopatry – and assortative mating in the sympatric and allopatric ranges of *L. cingulata* and *L. filosa* as a possible signature of reinforcement. The analysis of penial shape is still in progress. However, preliminary results of the mating preference experiments revealed stronger assortative mating in sympatry than in allopatry for both species, as measured by mating duration, suggesting greater isolation in sympatry (Hollander, Butlin and Reid, unpublished).

Study questions

Chapter II: What are the patterns of shell shape variation between and within species in the *Littorina saxatilis* complex from the British Isles? What are the influences of other species, environment and geography on these patterns within and among sites?

Chapter III: What is the spatial distribution of the genetic variation in the *Littorina saxatilis* complex from the British Isles? Is there population structure? How were these isles colonised?

Chapter VI: Was there gene flow between the sister species *Littoraria cingulata* and *Littoraria filosa* from Australia upon secondary contact? What is the demographic history of their populations?

Methodology

In this section, a brief description of best practice in morphometrics is presented, as well as a short review of alternative DNA sequencing and genotyping methods. Morphometrics was used in this work for examining shell shape variation in the *L. saxatilis* complex. DNA sequencing was used for producing the genotype data in the *L. saxatilis* complex and in the *Littoraria* species.

Morphometrics

The best practice in morphometrics has changed over time as new approaches and tools have been developed. Organismal form can be described in terms of familiar shapes, such as round, horseshoe-shaped, egg-shaped, among others (e.g. Díaz & Rodríguez 1977). However, not all shapes can be described in these terms; and even if they could, the descriptions would be very limited and lacking in precision (Zelditch *et al.* 2012). In this context, a quantitative approach is better suited to study shape because it provides a much more objective measurement of shape. Morphometrics, a branch of mathematical shape analysis, has widely been used for this purpose. However, early morphometric studies used to report shape differences as large tables of numbers that did not facilitate visualisation of the shapes (e.g. Hughes 1972).

Traditional morphometric studies used to consider distance and angle measurements in an attempt to capture shape information (e.g. Mill & Grahame 1995). However, there were problems associated with this approach (discussed by Mitteroecker & Gunz 2009). One of them was that distances usually overlapped and shared end points, making many of the measurements redundant and increasing the error introduced by wrongly placing end points. Another problem was that measurements in an individual lost their spatial relationship when analysed independently, making it difficult to localise with precision where shapes differed. Yet, one of the biggest problems was that shape could not be clearly told apart from size, and they could not therefore be

independently treated in order to study, for example, the relationship between one and the another.

A solution to most of these problems was provided by geometric morphometrics, a method formalised by Bookstein (1991) and adapted from work by Kendall & Kendall (1980) and Strauss & Bookstein (1982). It uses coordinates of landmarks containing not only all the information about distances between all landmarks, but also their spatial location relative to each other. This latter feature of geometric morphometrics, including its ability to describe and graphically represent magnitude, direction and location of shape differences, is among the greatest advantages of the method. However, landmarks, defined as individually homologous anatomical loci (Bookstein 1991), do not solve all of the problems, as they usually fail to capture information on featureless or smooth curvatures. In these cases, other methodologies accounting for outlines should be considered, such as the use of semilandmarks (Bookstein 1997), eigenshape analysis (Lohmann & Schweitzer 1990), and Fourier analysis (Foote 1989).

Another relevant problem in morphometrics arises when analysing a three-dimensional shape from a two-dimensional representation of it (discussed by Zelditch *et al.* 2012). Working with three-dimensional data has the obvious benefit of potentially capturing shape data from all parts of the individual in a single configuration of landmarks, keeping the spatial relationship between all of them. However, one limitation of three-dimensional approaches is data collection, as the equipment to record landmarks (e.g. three-dimensional scanner) is often expensive. Despite the practicalities of collecting three-dimensional data, the mathematical theory and graphical visualization of shape differences follow the same basic principles as those of two-dimensional data. A recent method proposes the use of surface scans as data, rather than landmarks, to analyse shape differences (Pomidor *et al.* 2016).

The distinction between size and shape is a key advance of geometric morphometrics because it allows analysing them separately (e.g. test the effect of size, or environmental variables, on shape). Shape has been defined as “all the geometric information that remains when location, scale and rotational effects are filtered out from an object” (Kendall 1977). After removing these

effects, the differences between landmark configurations are solely explained by shape differences. The definition of size is tightly related to that of shape in that it represents the effect of scale. In this sense, size is defined considering the distance from all landmarks to the centre of the configuration. The square root of the sums of squared distances from each landmark to the centre of a configuration has been termed centroid size and it is widely used in geometric morphometrics for being the only measurement of size that is independent of shape. It is noteworthy that even though size and shape are mathematically independent, this is seldom true for biological systems (Zelditch *et al.* 2012).

Data analysis in geometric morphometrics involves addressing two main questions (Bookstein 1991): whether there are any effects on shape variation, and if so, what those effects are. The first question can be addressed, for example, by regressing shape on size. For this, an important consideration to be made is the fact that analyses of shape in geometric morphometrics have to be multivariate given that the set of coordinates of landmarks within each individual is considered as an individual record. Addressing the second question involves a combination of exploratory analyses and hypothesis testing. Ordination methods such as principal component analysis and canonical variate analysis are used for exploring the main axes of shape variation between individuals and between groups of individuals, respectively. Though not always the case, the expectation of these ordination methods is that individual axes would help to understand the factors that affect shape variation (Zelditch *et al.* 2012).

Hypothesis testing is a much less structured procedure in geometric morphometrics, since it involves tailoring a variety of analyses to specific research questions. At this stage, the question about whether there is an effect on shape takes a further step and becomes a matter of discriminating the effect, and relative contribution, that each factor might have on shape.

DNA sequencing

The description of DNA structure (Watson & Crick 1953) and the advent of DNA sequencing (Sanger *et al.* 1977) have been milestones in practically all disciplines of biological research, greatly advancing our understanding of life. DNA sequencing refers to a process whereby a particular order in which

nucleotides follow each other is determined. Knowing this particular order is particularly relevant to branches of biological knowledge such as medical diagnosis (e.g. Marra *et al.* 2003), biotechnology (e.g. Henne *et al.* 2004), forensic biology (e.g. Ballantyne *et al.* 2010), systematics (e.g. Pyron & Wiens 2011), evolution (e.g. Martin *et al.* 2015), among others. Improvement of DNA sequencing methods has gone hand in hand with technological advances, leading, for example, to whole genomes being sequenced for several species (e.g. Parker *et al.* 2013; Jarvis *et al.* 2014; Kim *et al.* 2014; Wang *et al.* 2014; Zimin *et al.* 2014; Albertin *et al.* 2015; Boothby *et al.* 2015; Zwarycz *et al.* 2015).

Sanger *et al.* (1977) introduced one of the pioneering sequencing methods, marking the start of what would be known later as Sanger sequencing. The principle of the method is based on the inhibitory activity that dideoxynucleotides (ddNTPs) have on DNA polymerase, terminating chain growth upon their incorporation (Atkinson *et al.* 1969). The method involves four separate reactions, one for each of the nitrogenous bases in DNA (i.e. adenine, guanine, thymine, and cytosine). Within each reaction, DNA synthesis proceeds with all four deoxyribonucleotides (dNTPs) and a specific ddNTP. Thus, each reaction produces a set of DNA strands of different lengths, each terminating at a position where a specific ddNTP was incorporated. These products are then separated by electrophoresis in parallel, displaying a banding pattern from which DNA sequence is read. The principle of the method was then automated in instruments known as DNA sequencers (e.g. 3730 DNA Analyzer, Applied Biosystems).

Different strategies that aim at sequencing long stretches of DNA have employed the Sanger method. One such strategy is shotgun sequencing (Anderson 1981; e.g. Rondon *et al.* 2000), which consists of randomly breaking DNA into shorter fragments that are then cloned into vectors, which are in turn used for infecting a bacterial host. DNA is then extracted from the grown cultures and sequenced using the Sanger method. Finally, resulting overlapping reads from different fragments are assembled into a continuous sequence using computer programmes (e.g. Staden 1979). However, despite the benefit of generating relatively long reads (400–9000 bp), the Sanger method is being overcome by the development of much quicker and cheaper sequencing

technologies, generally known as next-generation sequencing (NGS) or high-throughput sequencing.

Shotgun approaches are still used by most NGS technologies. Several authors have reviewed the main features of these technologies and highlighted differences among them (Mardis 2008; Shendure & Ji 2008; Ansorge 2009; Metzker 2010). Some of the most important advantages of these technologies have been the parallelisation of sequencing – making possible to overcome the limited 96-well capacity of the then-available instruments – and the high level of multiplexing, i.e. simultaneous sequencing of different samples in single reactions. Discussing the current state of the art in NGS technologies goes beyond the scope of this review. Instead, the major features of the most common NGS technologies currently used in molecular ecology research are presented here.

Most NGS technologies involve clonal amplification of DNA fragments through polymerase chain reaction (PCR) prior to sequencing in order to amplify their signal and allow their detection. Clonal amplification can be achieved by bridge PCR (Adams & Kron 1997; Kawashima *et al.* 1998) or emulsion PCR (Williams *et al.* 2006). In both cases, short oligonucleotide adapters are ligated onto the DNA fragments, allowing the denatured DNA to bind to the surface where the amplification takes place. Bridge PCR, used in methods developed by the company Illumina, requires adapters on both ends of DNA fragments in order to bind to a flow cell that has primers matching the adapters. Then, clone density on the flow cell increases as amplification proceeds, generating clusters of cloned fragments. On the other hand, emulsion PCR, used in methods developed by Margulies *et al.* (2005), Shendure *et al.* (2005), and the company Ion Torrent Systems, takes place in water-in-oil droplets, where DNA fragments bind to beads that have primers matching the adapters. Then, clone density on each bead increases as amplification proceeds.

Not only do NGS technologies differ with respect to their clonal amplification step, but they also differ regarding the principle whereby actual sequences are determined, namely sequencing-by-synthesis and sequencing-by-ligation. In sequencing-by-synthesis, used in methods developed by Illumina, Ion Torrent Systems, and Margulies *et al.* (2005), the DNA sequence is determined as a

DNA polymerase adds nucleotides – one at a time – to the growing chain. Unlike this principle, in sequencing-by-ligation, used in methods developed by Shendure *et al.* (2005), a DNA ligase is used to ligate oligonucleotides, which are complementary to the DNA template, to the growing chain.

Lastly, detection of newly added elements (i.e. nucleotides or oligonucleotides) to the growing chain differs among NGS technologies. Fluorescence is used in methods developed by Shendure *et al.* (2005) and Illumina, where oligonucleotides and dNTPs, respectively, are fluorescently labelled. After the incorporation of every element to the growing chain, the colour of the light, which is unique to each element, is recorded, thus revealing its identity. In methods developed by Margulies *et al.* (2005), each of the four dNTPs is added one at a time, its incorporation being detected by the light that is produced through a series of reactions involving pyrophosphate – a by-product of DNA synthesis – and a luciferase. In methods developed by Ion Torrent Systems, each of the four dNTPs is also added one at a time; however, detecting its incorporation is not based on the pyrophosphate but on the hydrogen ion – also a by-product of DNA synthesis – that is detected by a pH sensor.

As previously mentioned, all NGS technologies hitherto presented rely on an initial clonal amplification step. In contrast to these, other technologies are capable of sequencing in real time a single DNA molecule. One such a method has been developed by Eid *et al.* (2009), and commercialised by the company Pacific Biosciences. In this method, a single DNA-polymerase complex is fixed to the bottom of a well-like structure, where a sensor detects the light colour emitted by each fluorescently labelled dNTP that is incorporated to the growing chain.

Selecting the optimal NGS technology for a given project is not a straightforward task since all technologies have different advantages and disadvantages, most of which have been summarised by several authors (e.g. Ekblom & Galindo 2011; Glenn 2011; Quail *et al.* 2012). For instance, regarding the clonal amplification step, technologies using emulsion PCR have the advantage that amplification and sequencing take place independently, and so successful clonal amplification can be ensured before the sequencing step. However, emulsion PCR can suffer from inconsistencies similar to those of

random amplified polymorphic DNA (RAPDs), and PCR reagents are expensive. On the other hand, in technologies using bridge PCR, both clonal amplification and sequencing take place in the same location, which increases the efficiency of the method. However, this can also be a disadvantage if the clonal amplification performs poorly since the expensive sequencing reagents would then be wasted.

Many other factors need to be considered when selecting a particular NGS technology, such as costs, turnaround time, library preparation, consistency of the results, number and length of reads, error rates, computational resources for analysis, among others. Up to 2011, technologies developed by Illumina were leading in advantages such as number and percentage of error-free reads, number of reads per run, number of gigabases per run, and cost per gigabase sequenced (Glenn 2011). As costs continue to decrease and new technologies are developed, selecting the optimal sequencing approach will be based on the specific aims of a project, rather than on time and cost constraints.

Genotyping

The genotype of an individual usually refers to the part of its inherited genetic constitution that differs among alike individuals. Genotypes are typically defined in terms of specific genomic regions and, in diploid organisms, they denote the specific allelic combination carried by individuals, whose genotype can be homozygous (identical alleles) or heterozygous (different alleles). Determining an individual's genotype is known as genotyping, a process whereby the individual's alleles at a specific genomic location are identified. DNA sequences of these genomic regions that vary among individuals are commonly referred to as genetic markers, and their discovery – prior or simultaneous to genotyping – can be achieved through different methods, some of which are based on DNA fragment length polymorphism, and some others on determining the actual sequence of nucleotides.

Genetic markers based on DNA fragment length polymorphism include restriction fragment length polymorphism (RFLP, Botstein *et al.* 1980), random amplified polymorphic DNA (RAPD, Williams *et al.* 1990), variable number tandem repeat (VNTR), amplified fragment length polymorphism (AFLP, Vos *et al.* 1995), among others. On the other hand, genetic markers based on DNA

sequencing include single nucleotide polymorphism (SNP) and restriction-site associated DNA (RAD, Baird *et al.* 2008). In certain cases, such as population genetics studies involving large numbers of individuals and loci, the latter type of markers are supplanting those based on DNA length polymorphism because of the increasing cost-effectiveness of modern DNA sequencing technologies (see previous section on DNA sequencing).

Analysis of RFLP markers (Botstein *et al.* 1980) involves enzymatic digestion of DNA, followed by gel electrophoresis of the resulting fragments, and identification of specific alleles based on hybridisation of the restriction fragments with complementary synthetic probes. This analysis is now obsolete because it is slow and inefficient – usually one probe at the time – requiring a large amount of DNA and prior sequence information for probe design. Analysis of RAPD markers (Williams *et al.* 1990), when compared to RFLP analysis, requires no prior sequence information since DNA is randomly amplified using arbitrary primers, although high-quality DNA is required. However, there are issues regarding the reproducibility of this analysis (Penner *et al.* 1993) given that results are often dependant on PCR conditions, e.g. DNA quality and concentration, reagents concentration, and cycling programme.

The aim of analysing VNTR markers is to determine, in each individual, the lengths of particular tandem repeats, which are specific sequences of nucleotides whose repetitions are immediately adjacent to each other (e.g. King *et al.* 1997). Microsatellites constitute one of the two classes of VNTRs, the other being minisatellites, which differs from the former for having longer repeated blocks. VNTR analysis can be performed by RFLP and PCR techniques, the latter being more common since PCR is now a more established technique (e.g. Selkoe & Toonen 2006). In this way, specific primers are used for amplification of the microsatellites and their lengths determined using a DNA sequencer. A disadvantage of microsatellite analysis is that prior sequence information is required, and that the process of primer design and testing is often laborious and costly. Another issue with these markers is the potential occurrence of null alleles (Dakin & Avise 2004), which are the cases when a heterozygote is mistakenly scored as a homozygote because the region to which the primer anneals is mutated in one of the alleles, thus failing to detect it.

Similarly to RFLP analysis, AFLP analysis (Vos *et al.* 1995) involves digestion of DNA, except that in this case a subset of the restriction fragments are selectively amplified using adapters that are ligated to the sticky ends produced by enzymatic cuts. The principle of the method relies on the assumption that what it is a restriction site in one sample might not be in another one due to nucleotide substitution, digestion resulting in different fragmenting patterns. Amplified fragments are then separated using gels or capillary instruments; however, despite their name, genotypes are not scored in terms of length polymorphism but rather as presence/absence polymorphism. Apart from the advantage of not requiring prior sequence information, AFLP's reproducibility, resolution, sensitivity, and throughput are higher than RFLPs, RAPDs, and microsatellites (Mueller & Wolfenbarger 1999).

Genotyping methods using the markers described so far are based on fragment length or presence/absence polymorphism. However, genotyping can also be achieved by identification and allele determination of SNPs (Berger *et al.* 2001), which are polymorphisms at the level of a single nucleotide position caused by substitutions. SNP discovery and SNP genotyping are not always simultaneous processes, and the former does not necessarily results in the latter (e.g. sequencing of pooled individuals, Novaes *et al.* 2008). SNP discovery can be achieved through different methods such as single-strand conformation polymorphism (SSCP, Noll & Collins 1987) and DNA sequencing. The principle of SSCP analysis is based on the different structures formed by single strands of DNA with different composition but of identical lengths; such differences in structure are then detected by differing patterns of migration during gel electrophoresis (e.g. Bertin *et al.* 2005). However, the SSCP analysis is being supplanted by DNA sequencing methods due to higher cost-effectiveness. In DNA sequencing for SNP discovery, sequences from a reduced representation of the genomes of one or more individuals are aligned and screened for SNPs.

Once discovered, SNPs can be genotyped through different methods including hybridisation of the DNA with allele-specific oligonucleotides (ASO, Studencki *et al.* 1985) or with DNA microarrays (e.g. Hacia *et al.* 1999), allele-specific PCR (e.g. Walton *et al.* 1999), DNA sequencing (e.g. Davey *et al.* 2011), among others. In all cases, as with SNP discovery, the most common practice is to genotype only a subset of genomic regions. In hybridisation methods,

determination of SNP identity is based on the principle that only fully complementary DNA will hybridise with the allele-specific synthetic probes. In allele-specific PCR, specific primers that are complementary to the sequence where the SNP lies are used to selectively amplify different alleles. A disadvantage of these last two types of methods is that only those regions that fall within the probe area are genotyped, excluding any potential SNP discovery on the flanking sequences of targeted region. In DNA sequencing methods, the sequences of nucleotides are determined, thus revealing the SNP identity when compared with other such sequences.

A brief review of alternative sequencing methods has been presented in the previous section; here, different approaches of DNA library preparation that are used in combination with NGS technologies are discussed. Library preparation varies widely from one protocol to another (e.g. Baird *et al.* 2008; Gnirke *et al.* 2009; Andolfatto *et al.* 2011; Peterson *et al.* 2012), but the basic principle of library preparation is that DNA – once extracted and purified – is sheared and/or digested with a restriction enzyme in order to reduce fragment size. Adapters are then ligated to the fragments, which are then amplified in order to increase signal and allow detection by the sequencing instruments. A crucial step during this process is fragment selection since this will define the target regions that are to be sequenced. In this step, fragments may be selected based on their size (RADseq and other similar approaches; Baird *et al.* 2008; Andolfatto *et al.* 2011; Peterson *et al.* 2012) and/or specific sequence (targeted capture approach; Gnirke *et al.* 2009).

Originally developed as a genotyping tool in combination with DNA microarrays (Miller *et al.* 2007), RAD tags are the flanking sequences on either side of a restriction site. Since it is also a method for SNP discovery, sequencing of RAD tags is well suited for genotyping organisms with no prior sequence information available. Therefore, selection of DNA fragments during library preparation is based on their size. The basic protocol of RAD library preparation (Baird *et al.* 2008) involves random shearing of DNA after digestion and ligation of the first adapter. Variations of this basic protocol exist, such as that used for double digest RAD sequencing (ddRADseq; Peterson *et al.* 2012) and for multiplexed shotgun sequencing (MSG; Andolfatto *et al.* 2011), which avoid the shearing step – and therefore having to repair blunt ends of fragments – by using a

second restriction enzyme (ddRADseq) or a more frequent cutter (MSG) for further reducing fragment size. This feature of ddRADseq and MSG is very important because it may increase the method's reproducibility by excluding the step of DNA random shearing.

Note that predicting an enzyme's cut frequency is dependant on the length and composition of its recognition site, as well as on the GC content of the genome to digest. Cut frequency can also be affected by DNA methylation patterns since some enzymes are sensitive to these epigenetic modifications. Additionally, occurrence of null alleles should also be considered when sequencing RAD tags (Arnold *et al.* 2013; Gautier *et al.* 2013; Ravinet *et al.* 2015), since a mutated restriction sites will lead to the absence of the two corresponding tags in the library.

As previously mentioned, DNA fragments may also be selected based on their specific sequences during library preparation, as in a targeted capture approach (Gnirke *et al.* 2009). This is only possible for organisms with prior sequence information available since preselected sequences are targeted here (although see Jones & Good 2015). Such pre-selection of genomic regions can be performed at random (e.g. studies of neutral variation; this study) or specifically targeting loci of interest (e.g. follow-up studies of previously identified makers; Westram *et al.* 2016). In both cases, after an initial shearing and adapter ligation, the targeted genomic regions are captured by hybridisation probes, which are stretches of DNA that are complementary to the targeted sequence and that are, in this case, biotinylated, i.e. marked with biotin. This is an important feature of the probes because it allows the retention target-probe complexes by binding the biotin on the probes to streptavidin-coated magnetic beads, which are then isolated using magnets.

An advantage of sequencing preselected genomic regions using NGS technologies is that sequencing efforts are concentrated on target regions and are not wasted on off-target regions. On the other hand, a disadvantage of this approach is that prior sequence information is required in order to design the probes. This, however, will probably not be an obstacle in future given that NGS technologies are becoming more affordable and sequence information can relatively easily be generated for most organisms (Jones & Good 2015), for

example, by sequencing RAD tags. Another benefit of targeted capture sequencing, which is better suited for detecting selection in the genome and for phylogenetic analyses, among others (Jones & Good 2015), is the clear advantage that mapping reads to reference sequences represents over *de novo* assembly of reads into contigs and/or scaffolds prior to SNP genotyping.

References

Adams CP, Kron SJ (1997) Patent: Method for performing amplification of nucleic acid with two primers bound to a single solid support.

Albertin CB, Simakov O, Mitros T *et al.* (2015) The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*, **524**, 220–224.

Anderson S (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research*, **9**, 3015–3027.

Andolfatto P, Davison D, Erezyilmaz D *et al.* (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, **21**, 610–617.

Ansorge WJ (2009) Next-generation DNA sequencing techniques. *New Biotechnology*, **25**, 195–203.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179–3190.

Atkinson MR, Deutscher MP, Kornberg A, Russell AF, Moffatt JG (1969) Enzymatic synthesis of deoxyribonucleic acid. XXXIV. Termination of chain growth by a 2',3'-dideoxyribonucleotide. *Biochemistry*, **8**, 4897–4904.

Avise JC (2009) Phylogeography: retrospect and prospect. *Journal of Biogeography*, **36**, 3–15.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Ballantyne KN, Goedbloed M, Fang R *et al.* (2010) Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *American Journal of Human Genetics*, **87**, 341–353.

Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 379–406.

Berger J, Suzuki T, Senti K-A *et al.* (2001) Genetic mapping with SNP markers in *Drosophila*. *Nature Genetics*, **29**, 475–481.

Bertin I, Zhu JH, Gale MD (2005) SSCP-SNP in pearl millet – a new marker system for comparative genetics. *Theoretical and Applied Genetics*, **110**, 1467–1472.

Bookstein FL (1991) *Morphometric tools for landmark data: geometry and biology*. Cambridge University Press, Cambridge, UK.

Bookstein FL (1997) Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. *Medical Image Analysis*, **1**, 225–243.

Boothby TC, Tenlen JR, Smith FW *et al.* (2015) Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 15976–15981.

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, **32**, 314–331.

Buckland-Nicks J, Chia F-S (1990) Egg capsule formation and hatching in the marine snail *Littorina sitkana*. *Philosophical Transactions of the Royal Society B*, **326**, 159–176.

Butlin RK, Galindo J, Grahame JW (2008) Sympatric, parapatric or allopatric: the most important way to classify speciation? *Philosophical Transactions of the Royal Society B*, **363**, 2997–3007.

Butlin RK, Saura M, Charrier G *et al.* (2014) Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. *Evolution*, **68**, 935–949.

Chávez-Galarza J, Henriques D, Johnston JS *et al.* (2013) Signatures of selection in the Iberian honey bee (*Apis mellifera iberiensis*) revealed by a genome scan analysis of single nucleotide polymorphisms. *Molecular Ecology*, **22**, 5890–5907.

Conde-Padín P, Grahame JW, Rolán-Alvarez E (2007) Detecting shape differences in species of the *Littorina saxatilis* complex by morphometric analysis. *Journal of Molluscan Studies*, **73**, 147–154.
Coyne JA, Orr HA (2004) *Speciation*. Sinauer Associates, Inc., Massachusetts.
Dakin EE, Avise JC (2004) Microsatellite null alleles in parentage analysis. *Heredity*, **93**, 504–509.
Darwin C (1859) *On the origin of the species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, London.
Davey JW, Blaxter ML (2010) RADseq: next-generation population genetics. *Briefings in Functional Genomics*, **9**, 416–423.
Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.
Díaz H, Rodríguez G (1977) The branchial chamber in terrestrial crabs: a comparative study. *Biological Bulletin*, **153**, 485–504.
Doellman MM, Trussell GC, Grahame JW, Vollmer S V. (2011) Phylogeographic analysis reveals a deep lineage split within North Atlantic *Littorina saxatilis*. *Proceedings of the Royal Society B*, **278**, 3175–3183.
Duforet-Frebourg N, Bazin E, Blum MGB (2014) Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Molecular Biology and Evolution*, **31**, 2483–2495.
Eid J, Fehr A, Gray J *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.
Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.
Foote M (1989) Perimeter-based Fourier analysis: a New Morphometric Method Applied To the trilobite Cranidium. *Journal of Paleontology*, **63**, 880–885.
Futuyma DJ (2013) *Evolution*. Sinauer Associates, Inc., Massachusetts.
Gautier M, Gharbi K, Cezard T *et al.* (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22**, 3165–3178.
Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–769.
Gnirke A, Melnikov A, Maguire J *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, **27**, 182–189.
Gray MM, St. Amand P, Bello NM *et al.* (2014) Ecotypes of an ecologically dominant prairie grass (*Andropogon gerardii*) exhibit genetic divergence across the U.S. Midwest grasslands' environmental gradient. *Molecular Ecology*, **23**, 6011–6028.
Hacia JG, Fan JB, Ryder O *et al.* (1999) Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nature Genetics*, **22**, 164–167.
Hannaford-Ellis CJ (1979) Morphology of the oviparous rough winkle, *Littorina arcana* Hannaford-Ellis, 1978, with notes on the taxonomy of the *L. saxatilis* species complex (Prosobranchia: Littorinidae). *Journal of Conchology*, **30**, 43–56.
Henne A, Brüggemann H, Raasch C *et al.* (2004) The genome sequence of the extreme thermophile *Thermus thermophilus*. *Nature Biotechnology*, **22**, 547–553.
Hollander J, Galindo J, Butlin RK (2015) Selection on outlier loci and their association with adaptive phenotypes in *Littorina saxatilis* contact zones. *Journal of Evolutionary Biology*, **28**, 328–337.
Hollander J, Smadja CM, Butlin RK, Reid DG (2013) Genital divergence in sympatric sister snails. *Journal of Evolutionary Biology*, **26**, 210–215.
Hughes GM (1972) Morphometrics of fish gills. *Respiration Physiology*, **14**, 1–25.
Inness-Campbell J, Stuckey M, Johnson MS (2003) Allozymic investigation of phylogeny of *Littoraria* (Gastropoda: Littorinidae). *Journal of Molluscan Studies*, **69**, 19–26.
Janson K (1983) Selection and migration in two distinct phenotypes of *Littorina saxatilis* in Sweden. *Oecologia*, **59**, 58–61.
Jarvis ED, Mirarab S, Aberer AJ *et al.* (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, **346**, 1320–1331.
Johannesson K (1988) The paradox of Rockall: why is a brooding gastropod (*Littorina saxatilis*) more widespread than one having a planktonic larval dispersal stage (*L. littorea*)? *Marine Biology*, **99**, 507–513.
Johannesson K, Panova M, Kemppainen P *et al.* (2010) Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. *Philosophical Transactions of the Royal Society B*, **365**, 1735–1747.

Johnson MS, Black R (1998) Effects of isolation by distance and geographical discontinuity on genetic subdivision of *Littoraria cingulata*. *Marine Biology*, **132**, 295–303.
Jones MR, Good JM (2016) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, **25**, 185–202.
Kawashima E, Farinelli L, Mayer P (1998) Patent: Method of nucleic acid amplification.
Kendall DG (1977) The diffusion of shape. *Advances in Applied Probability*, **9**, 428–430.
Kendall DG, Kendall WS (1980) Alignments in two-dimensional random sets of points. *Advances in Applied Probability*, **12**, 380–424.
Kim S, Park M, Yeom S-I *et al.* (2014) Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nature Genetics*, **46**, 270–278.
King DG, Soller M, Kashi Y (1997) Evolutionary tuning knobs. *Endeavour*, **21**, 36–40.
Lohmann GP, Schweitzer PN (1990) On eigenshape analysis. In: *Proceedings of the Michigan morphometrics workshop* (eds Rohlf FJ, Bookstein FL), pp. 147–166. University of Michigan Museum of Zoology, Ann Arbor, Michigan.
Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.
Lucek K, Sivasundar A, Roy D, Seehausen O (2013) Repeated and predictable patterns of ecotypic differentiation during a biological invasion: lake-stream divergence in parapatric Swiss stickleback. *Journal of Evolutionary Biology*, **26**, 2691–2709.
Mäkinen T, Panova M, André C (2007) High levels of multiple paternity in *Littorina saxatilis*: hedging the bets? *Journal of Heredity*, **98**, 705–711.
Mäkinen T, Panova M, Johannesson K *et al.* (2008) Genetic differentiation on multiple spatial scales in an ecotype-forming marine snail with limited dispersal: *Littorina saxatilis*. *Biological Journal of the Linnean Society*, **94**, 31–40.
Mardis ER (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, **9**, 387–402.
Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
Marra MA, Jones SJ, Astell CR *et al.* (2003) The genome sequence of the SARS-associated coronavirus. *Science*, **300**, 1399–1404.
Martin CH, Cutler JS, Friel JP *et al.* (2015) Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. *Evolution*, **69**, 1406–1422.
Metzker ML (2010) Sequencing technologies – the next generation. *Nature Reviews Genetics*, **11**, 31–46.
Mill PJ, Grahame J (1995) Shape variation in the rough periwinkle *Littorina saxatilis* on the west and south coasts of Britain. *Hydrobiologia*, **309**, 61–71.
Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.
Mitteroecker P, Gunz P (2009) Advances in geometric morphometrics. *Evolutionary Biology*, **36**, 235–247.
Mueller UG, Wolfenbarger LL (1999) AFLP genotyping and fingerprinting. *Trends in Ecology and Evolution*, **14**, 389–394.
Murphy WJ, Eizirik E, Johnson WE *et al.* (2001) Molecular phylogenetics and the origins of placental mammals. *Nature*, **409**, 614–618.
Noll WW, Collins M (1987) Detection of human DNA polymorphism with a simplified denaturing gradient gel electrophoresis technique. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 3339–3343.
Novaes E, Drost DR, Farmerie WG *et al.* (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics*, **9**, 312.
Panova M, Blakeslee AMH, Miller AW *et al.* (2011) Glacial history of the North Atlantic marine snail, *Littorina saxatilis*, inferred from distribution of mitochondrial DNA lineages. *PLoS ONE*, **6**, e17511.
Panova M, Boström J, Hofving T *et al.* (2010) Extreme female promiscuity in a non-social invertebrate species. *PLoS ONE*, **5**, e9640.
Panova M, Johansson T, Canbäck B *et al.* (2014) Species and gene divergence in *Littorina* snails detected by array comparative genomic hybridization. *BMC Genomics*, **15**, 687.
Parker J, Tsagkogeorga G, Cotton JA *et al.* (2013) Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*, **502**, 228–231.
Pelletier TA, Carstens BC (2014) Model choice for phylogeographic inference using a large set of models. *Molecular Ecology*, **23**, 3028–3043.

Penner GA, Bush A, Wise R *et al.* (1993) Reproducibility of random amplified polymorphic DNA (RAPD) analysis among laboratories. *PCR Methods and Applications*, **2**, 341–345.
Perrier C, Bourret V, Kent MP, Bernatchez L (2013) Parallel and nonparallel genome-wide divergence among replicate population pairs of freshwater and anadromous Atlantic salmon. *Molecular Ecology*, **22**, 5577–5593.
Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.
Pomidor BJ, Makedonska J, Slice DE (2016) A landmark-free method for three-dimensional shape analysis. *PLoS ONE*, **11**, e0150368.
Pyron RA, Wiens JJ (2011) A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. *Molecular Phylogenetics and Evolution*, **61**, 543–583.
Quail M, Smith M, Coupland P *et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.
Quesada H, Posada D, Caballero A, Morán P, Rolán-Alvarez E (2007) Phylogenetic evidence for multiple sympatric ecological diversification in a marine snail. *Evolution*, **61**, 1600–1612.
Ravinet M, Westram A, Johannesson K *et al.* (2015) Shared and nonshared genomic divergence in parallel ecotypes of *Littorina saxatilis* at a local scale. *Molecular Ecology*, **25**, 287–305.
Reid DG (1986) *The littorinid molluscs of mangrove forests in the Indo-Pacific region: the genus* Littoraria. British Museum (Natural History), London.
Reid DG (1996) *Systematics and evolution of* Littorina. The Ray Society, London.
Reid DG (2001) New data on the taxonomy and distribution of the genus *Littoraria* Griffith and Pidgeon, 1834 (Gastropoda: Littorinidae) in Indo-West Pacific mangrove forests. *The Nautilus*, **115**, 115–139.
Reid DG, Dyal P, Williams ST (2010) Global diversification of mangrove fauna: a molecular phylogeny of *Littoraria* (Gastropoda: Littorinidae). *Molecular Phylogenetics and Evolution*, **55**, 185–201.
Reid DG, Dyal P, Williams ST (2012) A global molecular phylogeny of 147 periwinkle species (Gastropoda, Littorininae). *Zoologica Scripta*, **41**, 125–136.
Rittmeyer EN, Austin CC (2015) Combined next-generation sequencing and morphology reveal fine-scale speciation in Crocodile Skinks (Squamata: Scincidae: *Tribolonotus*). *Molecular Ecology*, **24**, 466–483.
Rivera G, Davis JN, Godwin JC, Adams DC (2014) Repeatability of habitat-associated divergence in shell shape of turtles. *Evolutionary Biology*, **41**, 29–37.
Rondon MR, August PR, Bettermann AD *et al.* (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Applied and Environmental Microbiology*, **66**, 2541–2547.
Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463–5467.
Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters*, **9**, 615–629.
Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145.
Shendure J, Porreca GJ, Reppas NB *et al.* (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.
Soria-Carrasco V, Gompert Z, Comeault AA *et al.* (2014) Stick insect genomes reveal natural selection's role in parallel speciation. *Science*, **344**, 738–742.
Staden R (1979) A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, **6**, 2601–2610.
Strauss RE, Bookstein FL (1982) The truss: body form reconstruction in morphometrics. *Systematic Zoology*, **31**, 113–135.
Stuckey M (2003) Genetic divergence along Australia's tropical coastline, as seen in littorine snails: effects of geological history and ocean currents. Ph.D. Thesis, University of Western Australia.
Studencki AB, Conner BJ, Impraim CC, Teplitz RL, Wallace RB (1985) Discrimination among the human beta A, beta S, and beta C-globin genes using allele-specific oligonucleotide hybridization probes. *American Journal of Human Genetics*, **37**, 42–51.
Vos P, Hogers R, Bleeker M *et al.* (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research*, **23**, 4407–4414.

Walton C, Handley JM, Kuvangkadilok C *et al.* (1999) Identification of five species of the *Anopheles dirus* complex from Thailand, using allele-specific polymerase chain reaction. *Medical and Veterinary Entomology*, **13**, 24–32.

Wang W, Haberer G, Gundlach H *et al.* (2014) The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nature Communications*, **5**, 3311.

Watson JD, Crick FHC (1953) Molecular structure of nucleic acids. *Nature*, **171**, 737–738.

Westram AM, Galindo J, Rosenblad MA *et al.* (2014) Do the same genes underlie parallel phenotypic divergence in different *Littorina saxatilis* populations? *Molecular Ecology*, **23**, 4603–4616.

Westram AM, Panova M, Galindo J, Butlin RK (2016) Targeted re-sequencing reveals geographic patterns of differentiation for loci implicated in parallel evolution. *Molecular Ecology*, Accepted Author Manuscript. doi:10.1111/mec.13640.

Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey S V. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acid Research*, **18**, 6531–6535.

Williams R, Peisajovich SG, Miller OJ *et al.* (2006) Amplification of complex gene libraries by emulsion PCR. *Nature Methods*, **3**, 545–550.

Wood HM, Grahame JW, Humphray S, Rogers J, Butlin RK (2008) Sequence differentiation in regions identified by a genome scan for local adaptation. *Molecular Ecology*, **17**, 3123–3135.

Zelditch M, Swiderski D, Sheets H (2012) *Geometric morphometrics for biologists: a primer*. Academic Press, London.

Zimin A, Stevens KA, Crepeau MW *et al.* (2014) Sequencing and assembly of the 22-Gb loblolly pine genome. *Genetics*, **196**, 875–890.

Zwarycz AS, Nossa CW, Putnam NH, Ryan JF (2015) Timing and scope of genomic expansion within Annelida: evidence from homeoboxes in the genome of the earthworm *Eisenia fetida*. *Genome Biology and Evolution*, **8**, 271–281.

Chapter II

Shell shape variation of the intertidal snails in the *Littorina saxatilis* complex from the British Isles

Abstract

Organismal form results from a complex interaction between environments and genomes, and is often determined by adaptation to specific habitats. Rocky shores represent heterogeneous environments with a vast diversity of ecological conditions, making them ideal systems to study morphological and genetic variation among and within species. The snail species of the *Littorina saxatilis* complex (*L. saxatilis*, *L. arcana* and *L. compressa*) constitute an excellent study system for this, given the variation in shell shape that exists among species and most notably within *L. saxatilis* (ecotypic variation). Here, individuals from across the British Isles were analysed under a landmark-based geometric morphometric framework in order to characterise the variation in shell shape. Ordination methods revealed a clear distinction between *L. arcana* and *L. compressa*, which are species most commonly found in exposed and sheltered shores, respectively. Within *L. saxatilis*, no clear distinction was found between ecotypes H and M, which are found in contrasting habitats on the same shore, suggesting perhaps continuous shell shape variation. Ecotype *neglecta*, commonly found living among barnacles, formed a cohesive shape unit, whereas ecotype *tenebrosa*, collected from inland pools, formed two distinct shape groups. The extent to which variation in shape was associated with variation in environment and geography was investigated. Models obtained through analyses of covariance fitted the data well, although specific associations between shape and environment were difficult to interpret. However, there was a trend towards site and local environment having effects on shape, although the nature of the effect was unclear. Further studies should consider assessment of sites prior to collection and a more standardised sampling scheme across sites. Finally, the present study provides a platform for a study on the relationship between morphological and genetic variation, given the availability of DNA for each individual sample.

Introduction

Shape analysis is a fundamental tool that allows the quantification of morphological variation (Mitteroecker & Gunz 2009) and the understanding of its causes (e.g. Gómez Cano *et al.* 2013). Previous studies of the marine snail genus *Littorina* have found an association of shell shape variation with environmental conditions (Carvajal-Rodríguez *et al.* 2005), where ecotypes are found at different shore levels and habitats. There is also evidence that shell shape varies within *Littorina* species from different geographic areas, regardless of their habitat (Conde-Padín *et al.* 2007b). Shell shape variation has also an interspecific component, where different species have been found to have diagnostic shapes (Caley *et al.* 1995). Even though most of the morphological variation is inherited, experimental studies have found that a small proportion of it can be caused by plasticity (Hollander *et al.* 2006b; Conde-Padín *et al.* 2007a, 2009).

In the British Isles, *L. saxatilis* is found from the upper eulittoral area down to the littoral fringe of the intertidal zone, as well as in coastal bodies of brackish water (Reid 1996). The level of exposure to wave action (Janson & Sundberg 1983) and crab predation (Johannesson 1986) have been suggested to be an important factor affecting the shell shape in *L. saxatilis*, which might be even more pronounced in the British Isles due to a wider intertidal range compared to other regions where the species is found (Conde-Padín *et al.* 2007b). Different ecotypes of *L. saxatilis* have been described based on shell shape (reviewed by Reid 1996), and received different names in different geographic areas. As previously mentioned in Chapter I, at least four ecotypes have been described in the UK (Fig. I.1). The divergent polymorphism observed between ecotypes H and M is maintained by selection in the face of gene flow, but there is a further reduction of gene flow due to the spatial separation of habitats where the two ecotypes occur, and probably also to habitat choice (Grahame *et al.* 2006).

The shell morphology of this highly variable group of sea snails has been extensively studied using methods of multivariate analyses on distance measurements, such as width, length and depth (Janson & Sundberg 1983; Johannesson 1986; Caley *et al.* 1995; Mill & Grahame 1995; Johannesson & Johannesson 1996; Cruz *et al.* 2001). However, some of the problems with

reducing the shape of an individual to a set of distance measurements are that those variables are often not independent from each other, shape changes are difficult to locate due to a loss of spatial information, and size cannot clearly be separated from shape (Bookstein 1991; Stone 1998; Zelditch *et al.* 2012).

The use of geometric morphometrics as a tool for shape analysis has allowed these problems to be overcome by representing individual shapes as a set of landmarks on a coordinate system (Rohlf & Marcus 1993; Zelditch *et al.* 2012), which allows the mathematical separation of shape and size (Kendall 1977). The application of this methodology to *L. saxatilis* have revealed differences in shell shape between the H and M ecotypes in the UK (Conde-Padín *et al.* 2007b), as well as in their equivalent pairs of ecotypes in Sweden (Hollander *et al.* 2005, 2006b; Hollander & Butlin 2010) and in Spain (Carvajal-Rodríguez *et al.* 2005; Conde-Padín *et al.* 2007a). Most of these previous studies have focused on sampling individuals from environments where the well-defined ecotypes are typically found, perhaps failing to capture the overall shape variation that may exist in a given locality. These previous studies have also focused on a handful of intensively studied localities, leaving a wide area of the species' distribution underrepresented in the literature.

This chapter reports a study of shell shape in the *Littorina saxatilis* complex sampled along transects across various sites around the British Isles using a geometric morphometric approach. The specific aims of this study were: (1) to characterise shell shape variation within and between species, and (2) to investigate the role of other species, environment and geographic distribution as potential causes of shape variation within and among sites. Based on the principle that a distinct ecotype of *L. saxatilis* would be expected wherever the species occupies a different microenvironment (Johannesson *et al.* 2010), patterns of morphological variation were expected to be associated with particular microenvironments within sites, but also among sites, regardless of geographic distribution.

Methods

Sampling

Individuals of the *Littorina saxatilis* complex were collected from 31 sites around the British Isles (Fig. II.1, Table II.1). Some of the sites were selected based on previous sampling (Reid 1996; Small & Gosling 2000a; Conde-Padín *et al.* 2007b), some others through personal communication with John Grahame (University of Leeds), and some others by examination of satellite photographs looking for potentially suitable habitat. At most sites, one transect was set perpendicular to the shoreline and running inland, and divided into different numbers of equidistant sections to capture ecotypic variation. Transects at different sites were divided into different numbers of sections according to the extent of the area that could be sampled, e.g. fewer sections were sampled where the distance between the sea and the upper high water level was shorter. At each section within a given transect, snails were collected within an area of about 1 square metre. No transect was set at sites where snails were found only at a single shore level or in brackish water. Snails were preserved at –80 ºC prior to photographing and dissection.

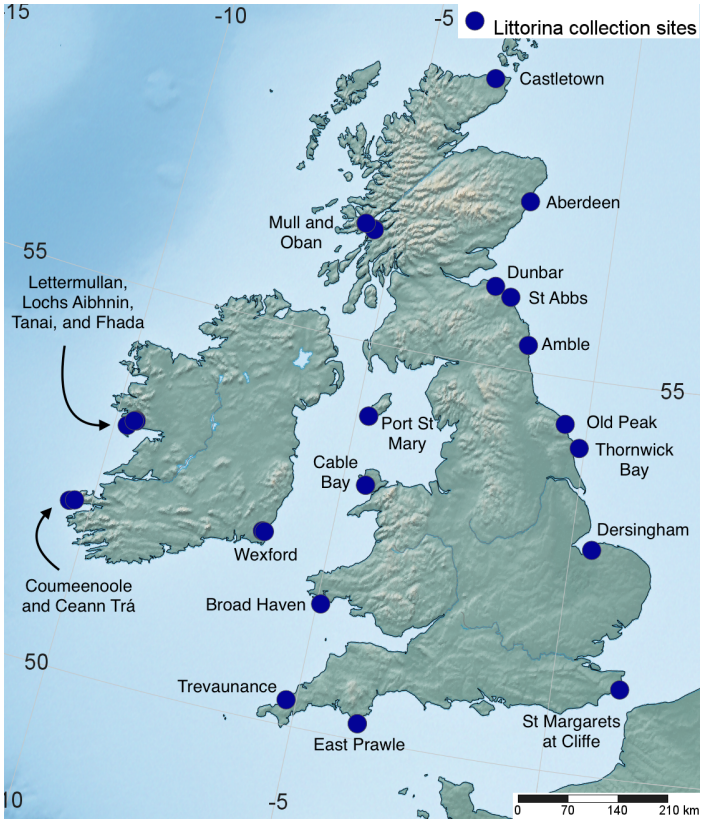


Fig. II.1 Map of sites sampled around the British Isles

Table II.1 Collection sites of *Littorina* individuals used in this study

Site	Site code	Latitude (°)	Longitude (°)	*L. saxatilis*			*L. arcana*	*L. compressa*
				H and *M*	*neglecta*	*tenebrosa*		
Thornwick Bay	THB	54.13267	–0.11503	71				
Cable Bay I	CAB	53.20724	–4.50024	15			4	
Cable Bay II	CAB	53.20703	–4.50458	41			8	
Broad Haven	BHE	51.60891	–4.91878	5				21*
Oban I	OB1	56.42207	–5.48392	7				12
Oban II	OB2	56.41133	–5.48025	2	4			3
Isle of Mull I	IM1	56.46981	–5.70344	20				
Isle of Mull II	IM2	56.47068	–5.69599	10				8
Trevaunance	TRE	50.32211	–5.19972	1			14	
East Prawle I	PEP	50.20544	–3.71664	20			4	
East Prawle II	PEP	50.20530	–3.71686	6				
St Margarets at Cliffe	SMC	51.14785	1.38415	22				
Castletown	CAS	58.59832	–3.38172	9				
Aberdeen	ABE	57.14018	–2.05124	14				
Dunbar	DUN	56.00360	–2.51174	6				
St Abbs	STA	55.89968	–2.13004	8			6	
Amble	AMB	55.33215	–1.56292	12	1		5	1
Old Peak	OLD	54.41036	–0.49196	6	2		1	
Port Saint Mary I	PS1	54.07602	–4.73618	24				6
Port Saint Mary II	PS2	54.07695	–4.73746	16	4			
Lettermullan	LET	53.23096	–9.72898	15			5	
Loch Aibhnin I	AIB	53.32606	–9.57610			4		
Loch Aibhnin II	AIB	53.31895	–9.57471	1		3	1	
Loch Tanai	AIB	53.31477	–9.57206			5		
Loch Fhada	FHA	53.31041	–9.61229	10				1
Coumeenoole	COU	52.10880	–10.46297	1			8	
Ceann Trá	CEA	52.13205	–10.36071	15			2	3
Wexford I	WEX	52.35618	–6.42006	4				
Wexford II	WEX	52.35683	–6.41952			1		
Wexford III	WEX	52.35373	–6.37763			3		
Dersingham	DER	52.86750	0.44738			3		

* Species identity of individuals from Broad Haven with a jelly gland could not be determined

Selection of individuals

Currently, there is no quantitative method based on external characters that allows the unambiguous discrimination between the closely related species *L. saxatilis*, *L. arcana* and *L. compressa*. Only after dissection is it possible to identify female *L. saxatilis* by the presence of a brood pouch containing eggs and/or embryos. Females of both *L. arcana* and *L. compressa* are recognised

by the presence of a jelly gland, but despite the existence of descriptions about their morphological differences (thorough review by Reid 1996) or methods to discriminate between species (Conde-Padín *et al.* 2007b), assignment of individuals to a particular species is difficult. These two species with jelly gland were identified based on morphology and habitat whenever possible, following descriptions by Reid (1996). Here, *L. compressa* was visually distinguished from *L. arcana* for having flattened spiral ridges – which were wider than the grooves between them – and by a thicker wall, as well as for having been collected at a rather lower level on less exposed shores. The four main ecotypes of *L. saxatilis* were included in this study; however, it was not possible to distinguish unambiguously between the M and H ecotypes, and only the *neglecta* and *tenebrosa* ecotypes were identified as such prior to morphometric analyses. Only adult females were used in this study, although males should also be used in future studies.

In total, 504 individuals were used for the morphometric analyses, of which 391 were *L. saxatilis*, 58 were *L. arcana*, and 34 were *L. compressa* (21 individuals had a jelly gland, but species identity could not be determined). Within *L. saxatilis*, 361 individuals were a mixture of M and H ecotypes with intermediate forms, 11 individuals were *neglecta*, and 19 individuals were *tenebrosa* (Table II.1).

Image acquisition

Shells were photographed with the aperture facing the camera's lens to capture functionally important features (Carvajal-Rodríguez *et al.* 2005) (Fig. II.2a). In order to ensure the same standard position of all shells, these were orientated following three alignments, one for each dimension, following the traditional orientation used in most research groups (Carvajal-Rodríguez *et al.* 2005; Walker & Grahame 2011). For this purpose, each individual was temporarily fixed to the centre of a microscope slide with a pressure-sensitive adhesive in order to restrict its movement. The first alignment (Fig. II.2a) was performed by parallel rotation of the shell relative to the slide plane, orientating its short edge parallel to the columella. The second alignment (Fig. II.2b) was performed by perpendicular rotation of the shell relative to the slide and parallel to its short edge, also orientating this short edge parallel to the columella. The third

alignment (Fig. II.2c) was performed by perpendicular rotation of the shell relative to the slide and parallel to its long edge, orientating this long edge parallel to the longest section perpendicular to the columella. Finally, all three alignments were verified as many times as necessary to ensure the correct orientation of shells in all dimensions. All three alignments, especially the second and third, were necessary to make certain that the position of homologous anatomical features relative to the camera's lens was the same across all shells. This way, the distortion associated with a two-dimensional representation of a three-dimensional object was minimised, ensuring thus that points identified later at landmarks did actually represent homologous anatomical features.

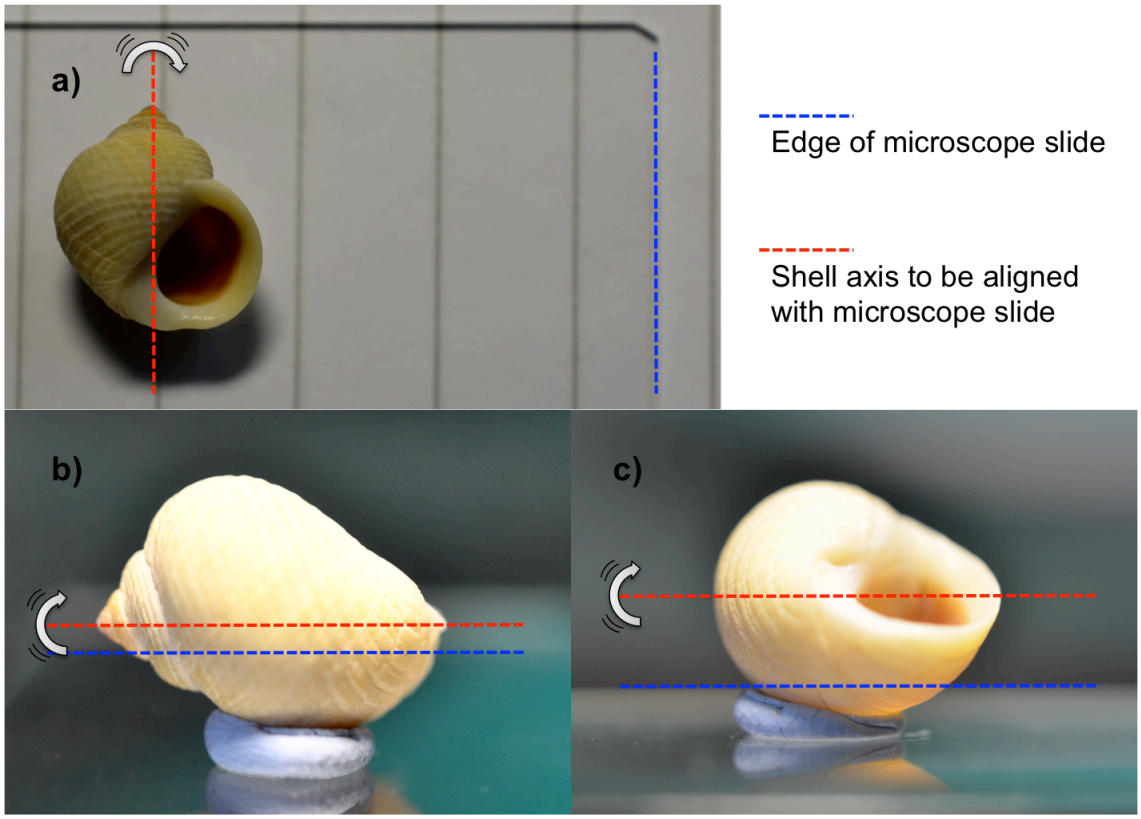


Fig. II.2 Orientation of shells for shape analyses

Shells were individually placed under the camera, with the slide perpendicular to the objective lens. Focus was adjusted to the perimeter of shells. Images were saved as JPEG files with dimensions of 640 x 480 pixels. Graph paper in the background was used as a size reference. In order to minimise distortion effects, shells were photographed occupying the same field of view at approximately the same distance from the lens (Zelditch *et al.* 2012). Photographs were taken using a digital microscope with an integrated camera

(Veho®, 400x USB microscope) operated by image capture software (Veho®, MicroCapture version 1.0, www.veho-world.com). In order to assess distortion associated with the camera, distances of one millimetre in the centre and on the edge of the field of view were measured on photographs and compared, but no appreciable distortion was observed.

Selection of landmarks

Fifteen points were designated as landmarks (LM) (Fig. II.3), some of which have previously been described (Carvajal-Rodríguez *et al.* 2005; Hollander *et al.* 2005, 2006a; Conde-Padín *et al.* 2007a; b). LM1 was at the apex of the shell. LM2 was on the right border of the profile of the shell at the end of the upper suture of the last whorl. LM3 was on the right border of the profile of the shell at the end of the lower suture of the last whorl. LM4 was at the end of the suture of the last whorl. LM13 was on the left border of the profile of the shell at the end of the upper suture of the last whorl.

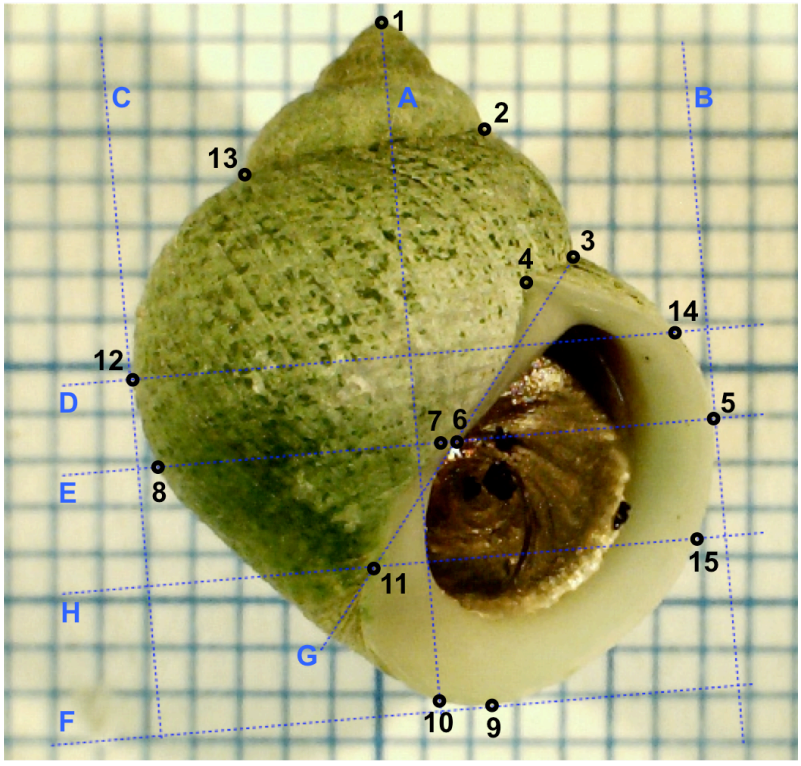


Fig. II.3 Landmarks and lines drawn on shells

The unambiguous location of the remaining 10 landmarks required the drawing of eight straight lines. Line A originated at the apex and was tangent to the inner border of the lip closest to the columella. Line B was parallel to line A and

tangent to the right border of the profile of the shell. Line C was parallel to line A and tangent to the left border of the profile of the shell. Line D was perpendicular to line A, intersecting the contact point between line C and the left border of the profile of the shell. Line E was perpendicular to line A, intersecting the contact point between line B and the right border of the profile of the shell. Line F was perpendicular to line A and tangent to the lower border of the profile of the shell. Line G originated at LM3 and was tangent to the inner border of the lip closest to the columella. Line H was perpendicular to line A, intersecting the contact point of line G and the outer border of the lip closest to the columella.

LM5 was on the intersection of lines B and E. LM6 was at the intersection of line E and the inner border of the lip closest to the columella. LM7 was at the intersection of line E and the outer border of the lip closest to the columella. LM8 was at the intersection of line E and the left border of the profile of the shell. LM9 was at the contact point between line F and the lower border of the profile of the shell. LM10 was at the intersection of line A and the lower border of the profile of the shell. LM11 was at the intersection of lines G and H. LM12 was at the intersection of lines C and D. LM14 was at the intersection of line D and the right border of the profile of the shell. LM 15 was at the intersection of line H and the right border of the profile of the shell.

According to the classification of landmarks proposed by Bookstein (1991), only LM1 and 4 fall within Type I category because their location corresponds to homologous anatomical points that can be identified in all specimens irrespective of their orientation, and because they are surrounded by tissue in all directions. LM2, 3 and 13 are Type II landmarks because they are locally defined, but their position depends on the orientation of the shell, and they are not completely surrounded by tissue. LM5, 6, 7, 8, 9, 10, 11, 12, 14 and 15 are Type III landmarks because they are defined by distant structures and their position is therefore geometrically identified. Despite the fact that Type II and Type III landmarks do not necessarily represent homologous parts across all shells from an ontogenetic point of view, they have proved to be useful for analysing shape variation in this species in an objective and repeatable way (Carvajal-Rodríguez *et al.* 2005; Hollander *et al.* 2005; Conde-Padín *et al.* 2007a; b).

Digitisation of landmarks

Pixelmator (version 3.3, Pixelmator Team) was used for drawing the eight straight lines on each photograph. The 15 landmarks were digitised as *x* and *y* coordinates using tpsDig (version 2.16; written by F. J. Rohlf; available at: http://life.bio.sunysb.edu/morph/), using the graph paper on the background as reference to set the scale. Configurations of landmarks were saved into text files, in tps format, which contain the following information about each individual shell: number of landmarks, *x* and *y* coordinates of each landmark, photograph filename, unique identification label, and scale. Prior to landmark digitisation, tpsUtil (version 1.44; written by F. J. Rohlf; available at: http://life.bio.sunysb.edu/morph/) was used for building the tps files.

Shell size and Procrustes superimposition

The maximum dimension of a shell was considered to be its height, i.e. the Euclidean metric distance between LM1 and LM9. Shell size was also measured using the centroid size, as defined by Bookstein (1991). Shell height and centroid size were plotted against each other to clarify the biological meaning of centroid size, which increases with the number of landmarks. A regression line was drawn using a linear model. Calculations of shell height and graphs were produced in R (R Core Team 2014), while centroid size was calculated in MorphoJ (Klingenberg 2011)..

Procrustes superimposition, a method described by Rohlf & Slice (1990), was used to remove the effects of location, size and rotation. The result was a set of configurations of landmarks whose differences are exclusively explained by shape. Procrustes superimposition was performed in MorphoJ (Klingenberg 2011).

Description of shape

Principal component analysis (PCA) was used for describing the diversity of shell shape among individuals within the entire sample (n = 504), i.e. all individuals belonging to the three species. Separate PCAs were also performed within all *L. saxatilis*, within *L. saxatilis* excluding *neglecta* and *tenebrosa*, and within *L. arcana*, *L. compressa* and the rest of the individuals with a jelly gland. The expectation with the last two PCAs was that they would show differences

between H and M ecotypes of *L. saxatilis* and whether unidentified individuals with a jelly gland overlapped more with *L. arcana* or *L. compressa*, respectively. In every case, the extreme deformations along any given principal component were depicted as soft wireframe graphs to graphically appreciate the shell shape changes. Scree and scatter plots were produced in R (R Core Team 2014), and PCAs and soft wireframes graphs were produced in MorphoJ (Klingenberg 2011). Only those principal components (PC) that individually explained more than 5% of the variance were considered for analysis, since PCs that explain less than this proportion are unlikely to be biologically meaningful (Zelditch *et al.* 2012). In order to know whether the combination of size and shape could more clearly separate species and/or ecotypes where shape alone did not, the relationship between size and shape across the entire sample was explored by comparing the variation along the leading PCs with shell height.

Canonical variate analysis (CVA) was used for describing the differences among groups of individuals that could be *a priori* sorted into mutually exclusive sets. This study considered five such groups: (i) *L. saxatilis* H and M ecotypes, (ii) *L. saxatilis* ecotype *neglecta*, (iii) *L. saxatilis* ecotype *tenebrosa*, (iv) *L. arcana*, and (v) *L. compressa*. However, since individuals scored as *neglecta* and *tenebrosa* had *a priori* been identified as such solely based on shell shape, these were excluded from the CVA. In contrast, since individuals scored as either *L. saxatilis* H and M ecotypes, *L. arcana* and *L. compressa* were identified as such based on female reproductive anatomy (*L. saxatilis* from the other two species) or visual inspection of shell shape combined with habitat (i.e. *L. arcana* and *L. compressa*), these three groups were included in the CVA in order to explore quantitative shape differences between them. Mahalanobis distances between the three group means were calculated, and a permutation test of 10,000 rounds performed on them. The deformations of shape associated with canonical variates (CVs) were computed by regression of shape onto the CV scores, and then depicted as a soft wireframe graphs. The scatter plot was produced in R (R Core Team 2014), and CVA and soft wireframes graphs were produced in MorphoJ (Klingenberg 2011).

Discriminant function analyses were performed in order to further investigate the possibility of distinguishing *L. saxatilis* H and M ecotypes, *L. arcana* and *L.*

compressa based on shell shape alone. Mahalanobis distances and shape deformation were computed as in the CVA. The rate of correct individual assignment to groups was reanalysed using a cross-validation test of 10,000 rounds, which estimates the actual error rate associated with the circularity of the method (Efron & Tibshirani 1995). Discriminant function analysis, cross-validation and wireframes graphs were produced in MorphoJ (Klingenberg 2011), while frequency graphs were produced in R (R Core Team 2014).

Covariation of shape with other variables

Statistical modelling was used to investigate the extent to which shell shape variation (response variable) was associated with variation in a number of environmental and geographical variables (explanatory variables). In brief, response variables were the PC scores from each of the PCs that were selected in the shape analysis performed over the whole sample. Explanatory variables included a mixture of both continuous and categorical variables (details below). Analysis of covariance (ANCOVA) was used as the statistical method given the specific combination of response and explanatory variable types in this study. A maximal model was initially fitted that contained all explanatory variables except their interactions, and it was then simplified by stepwise deletion in order to determine the minimal adequate model, which would include those terms (if any) whose variation is associated with the variation in shape.

Given that statistical modelling is most reliable when the data set contains no missing information, only those 431 individuals from the morphometric analyses with complete information for all explanatory variables were included in the ANCOVA. Collection site was used as a factor with 21 levels. Sites were grouped into five geographical regions (NE Britain, W Britain, S Britain, Isle of Man, and Ireland) according to the three islands sampled. Sites in Britain were further subdivided into three groups based on the colonisation history suggested by Doellman *et al.* (2011) and considering the extensive sandbank found between Thornwick Bay and Dersingham (Table II.2). Bedrock geology was used as a factor with five levels (mudstones and sandstones, limestone, metamorphic, igneous, and conglomerate), based on information from The British Geological Survey and The Geological Survey of Ireland. The structure

where a snail was sitting at collection time was used as a factor with four levels (barnacle, crevice, smooth, and under rock). The last categorical variable used was reproductive structure with two levels (brood pouch and jelly gland). Species and ecotype were not included as explanatory variables because some of their values were inferred from the morphometric analyses, so they could not be used in turn to predict shape. Latitude and centroid size were used as continuous explanatory variables.

Table II.2 Geographical regions used in the ANCOVA. Grouping criteria are presented in main text.

Ireland	Isles of Man	S Britain	W Britain	NE Britain
Lettermullan	Port St Mary I	St Margarets at Cliffe	Isle of Mull I	Castletown
Ceann Trá	Port St Mary II	East Prawle	Isle of Mull II	Aberdeen
		Trevaunance	Oban I	Dunbar
			Oban II	St Abbs
			Cable Bay I	Amble
			Cable Bay II	Old Peak
			Broad Haven	Thornwick Bay

As previously mentioned in the sampling section, snails were collected along a line that was perpendicular to the shore. Given that the number of collecting points along a transect varied across sites, the positions of snails along the transect were transformed into a scale from 0 to 1, where 0 represented the position closest to the sea and 1 represented the position furthest from the sea. The purpose of this was to have comparable scales across sites representing the low–high shore axis. The resulting continuous variable was used as an explanatory variable. Thus, sites where no transect was set – having only one sampled position – were not included in this analysis since they lacked this information (mainly *L. saxatilis* ecotype *tenebrosa*).

Most intertidal species are distributed into specific zones according to environmental conditions (Ballantine 1961). Thus, presence/absence of 20 taxa was recorded for every collection point within transects in order to capture environmental variation. Taxa were visually identified to the most exclusive taxonomical level possible on the field. These taxa included anemones (*Actinia*), limpets (*Patella*), sea snails (*Littorina littorea*, *Littorina obtusata*, *Melarhaphe neritoides*, *Nucella lapillus*, *Osilinus lineatus*, and *Gibbula*), mussels (*Mytilus edulis*), barnacles (*Elminius modestus* and other species), crabs, brown algae (*Ascophyllum nodosum*, *Fucus serratus*, *Fucus spiralis*,

Fucus vesiculosus, and *Pelvetia canaliculata*), green algae (*Ulva*), and supralittoral lichens (*Lichina pygmaea* and other species).

These presence/absence data were transformed into continuous variables by multidimensional scaling (MDS) using the R package *prabclus* (Christian Henning, University College London). First, a distance matrix was computed among all collection points, and then rescaled into data with up to *k* dimensions (*k* ranging from 1 to n–1, where n was the number of variables in the original data, i.e. n = 20). After examination of the goodness of the fit for each number of dimensions, the MDS with five dimensions was retained, which together explained 65% of the variation in the presence/absence data.

The maximal model did not include interactions between explanatory variables due to computational limitations of fitting such a large number of interactions between all variables used in this study. Moreover, simple models should be preferred over complex ones according to the principle of parsimony. A maximal model was independently fitted to each of the four PCs selected in the geometric morphometric analysis performed over the whole sample. Each of these models included all explanatory variables here summarised: geographical region, site, bedrock geology, latitude, sitting structure, relative position on the shore, dimensions 1 to 5 of the MDS analysis, reproductive structure and centroid size. Model simplification was performed through stepwise deletion based on Akaike's information criterion as implemented in the R package *MASS* (Venables & Ripley 2002). Models were further simplified by aggregation of non-significant factor levels (mainly within site and sitting structure; p-value < 0.05) in a stepwise *a posteriori* procedure in order to reach the minimal adequate model. Finally, models were checked for constancy of variance, normal distribution of errors and the presence of highly influential points.

Results

Distribution of shell size

The relationship between shell height and centroid size was positive and linear (r^2 = 0.99) (Fig. II.4). The smallest individuals were *L. saxatilis* ecotype *neglecta*, as expected. The widest size range was observed in *L. saxatilis* H and M, with the smaller individuals overlapping with *L. arcana* and the bigger

individuals overlapping with *L. compressa*. Very little size overlap was observed between *L. arcana* and *L. compressa*. Individuals with a jelly gland whose species was not determined were similar in size to the smaller *L. arcana* individuals, suggesting that they might belong to this species. *L. saxatilis* ecotype *tenebrosa* from Ireland and most *L. saxatilis* ecotype *neglecta* were below the regression line. For *L. saxatilis* ecotype *tenebrosa* from Ireland, this means that these individuals had a greater shell height relative to the centroid size in comparison to all other species/ecotypes of similar centroid size, which could relate to the tall spire that characterises this ecotype. In the opposite case, most *L. compressa* and many *L. arcana* individuals were above the regression line. This means that they had shorter shell height relative to centroid size in comparison to all other species/ecotypes of similar centroid size, which could relate to their rather vertically compressed shells.

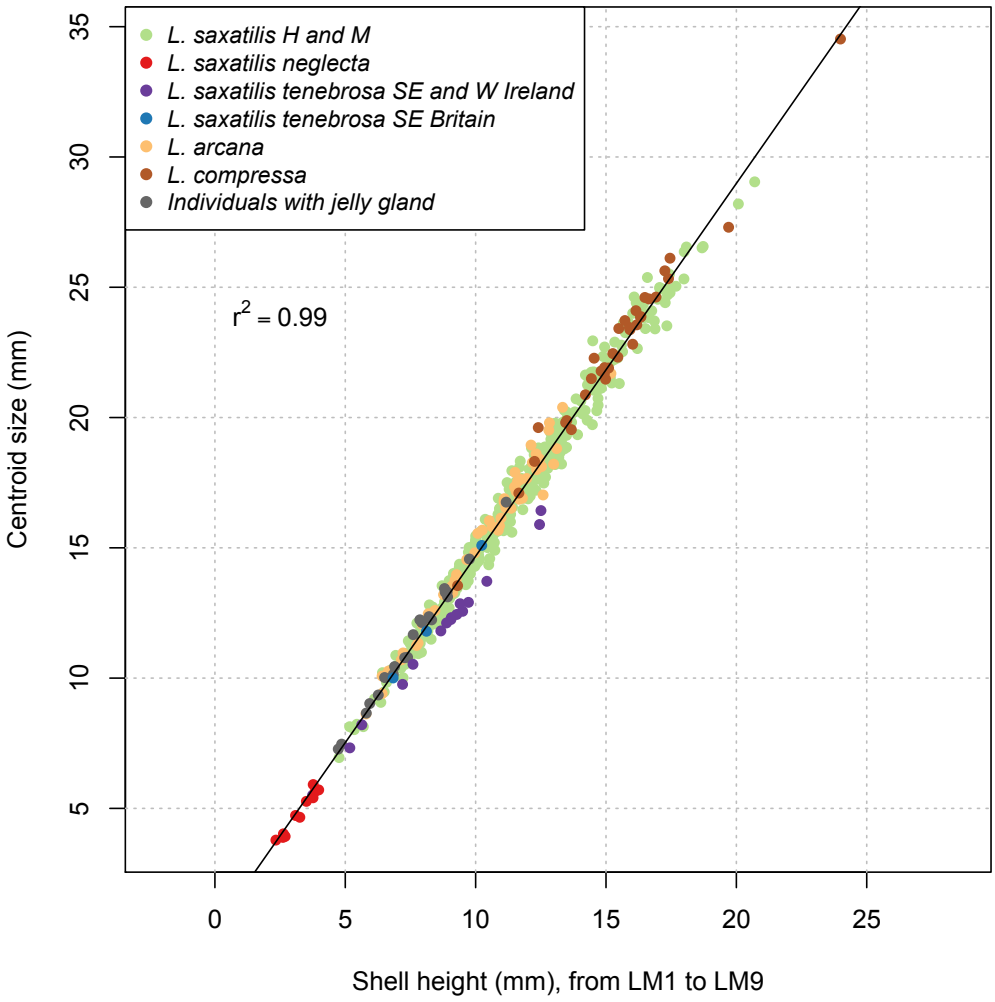


Fig. II.4 Relationship between shell height and centroid size; individual dots represent each of the 504 individuals included in this study

Description of shape diversity

The initial PCA of the whole dataset that included the three species revealed that each of the first four principal components (PC) individually explained more than 5% of the variance, accounting for a total of 84% of the variance in the sample (Fig. II.5a). The variation along PC1 (42%) distinguished *L. saxatilis* ecotypes *neglecta* and *tenebrosa*, and – to a lesser degree – *L. compressa* and unidentified individuals with a jelly gland (Fig. II.5b and Fig. II.5c), with differences associated with the shell's globosity and the aperture shape (Fig. II.6, PC1). PC1 also showed a clear distinction between *L. saxatilis* ecotype *tenebrosa* from Ireland and from Britain, with the latter being more similar to *L. saxatilis* H and M ecotypes. The variation along PC2 (25%) partially distinguished *L. saxatilis* from *L. arcana*, with *L. compressa* being more similar to *L. saxatilis* and unidentified individuals with a jelly gland more similar to *L. arcana* (Fig. II.5b and Fig. II.5d). *L. saxatilis* ecotypes *neglecta* and *tenebrosa* showed a large variation along PC2. Individuals towards the *L. saxatilis* extreme of PC2 had a taller spire, a narrower and longer aperture, and a less rounded border in the area opposite to the apex (Fig. II.6, PC2). The variation along PC3 (12%) showed a separation between a group including *L. arcana*, unidentified individuals with a jelly gland, and *L. saxatilis* ecotypes *neglecta* and *tenebrosa* from Britain, and another group comprising *L. saxatilis* ecotype *tenebrosa* from Ireland (Fig. II.5c and Fig. II.5d), with individuals in the latter group having a taller spire and a smaller aperture relative to the body size (Fig. II.6, PC3). The variation along PC4 (6%) showed no obvious clustering pattern of individuals, although the shape differences were rather similar to those observed along PC3 (Fig. II.6, PC4).

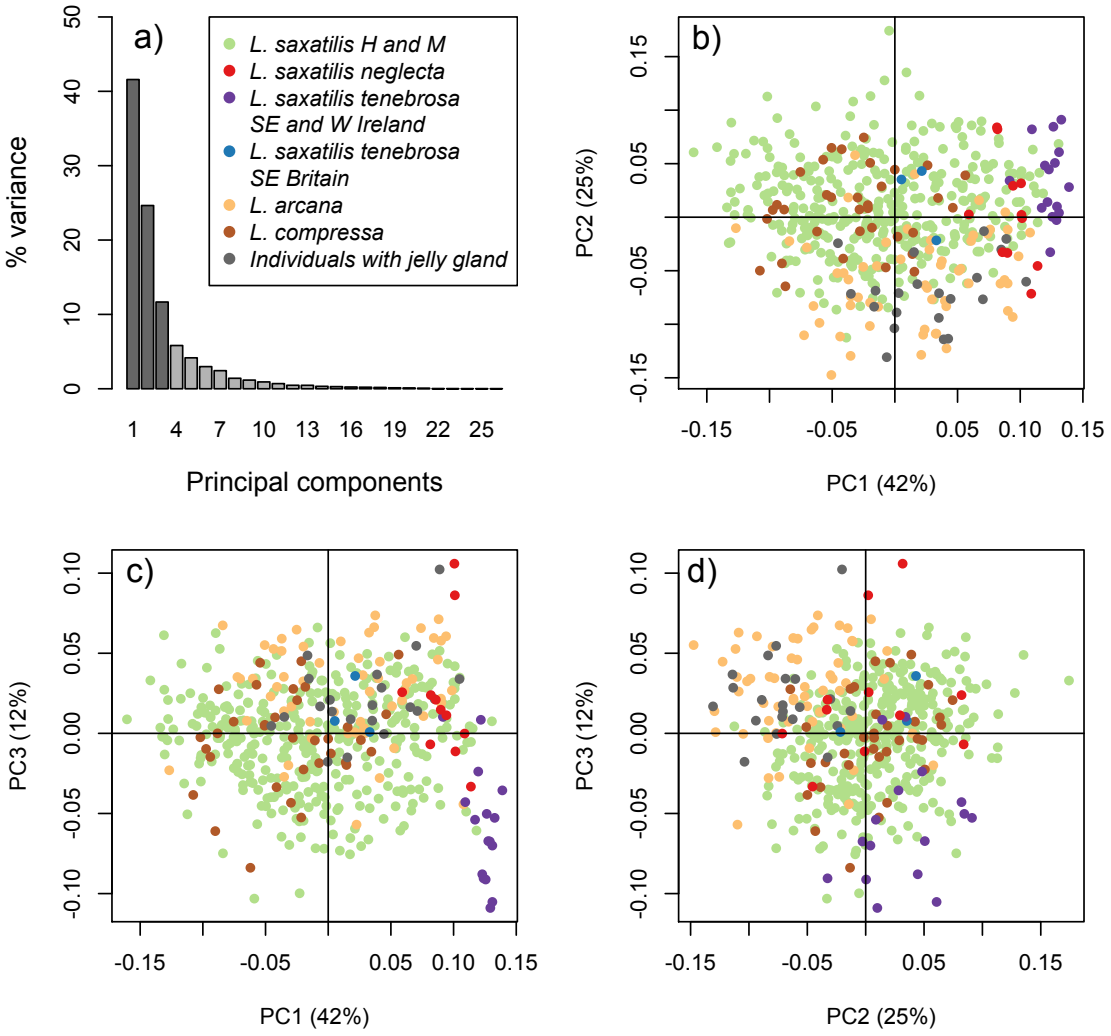


Fig. II.5 Principal component analysis of all 504 individuals. (a) Scree plot of the variance described by each PC. Scatter plots showing scores on (b) PC1 and PC2, (c) PC1 and PC3, and (d) PC2 and PC3; dots represent single individuals.

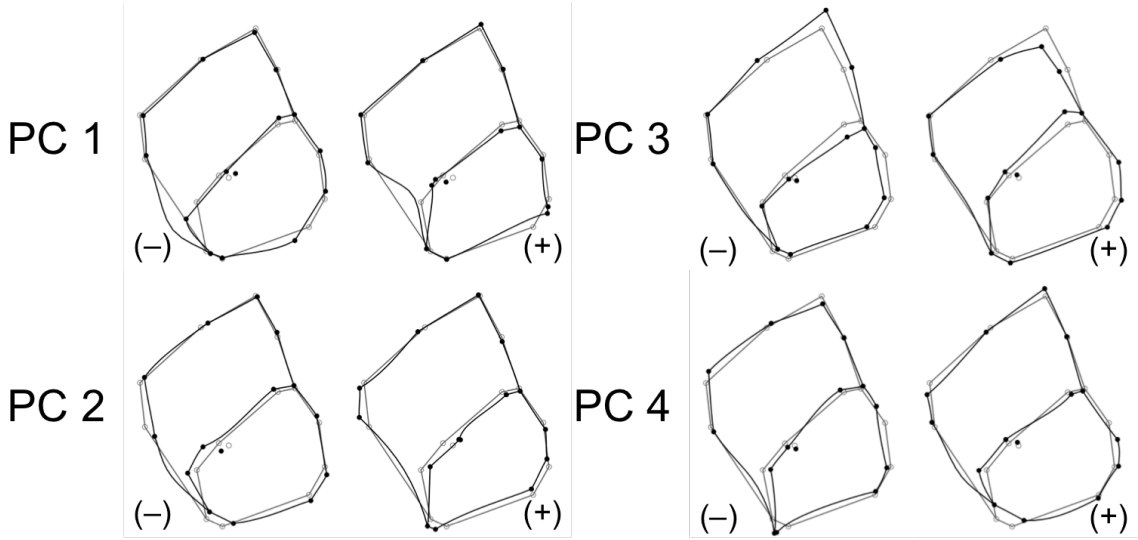


Fig. II.6 Patterns of shape variation along the first four PCs on all 504 individuals. Soft wireframe graphs show the most extreme negative and positive deformations of landmarks (black) with respect to the average configuration (grey). Lines connect some landmarks to facilitate appreciation of shape differences.

Considering the results from the PCA of all 504 individuals, significant linear relationships (p-value < 0.01) were observed between all four principal components and size (Fig. II.7). However, the fraction of total variation in shape that was explained by size was much higher for PC1 and PC3 (r^2 = 0.17 and 0.13, respectively) (Fig. II.7). These results showed that individuals tended to cluster by species/ecotype and size. For instance, *L. saxatilis* ecotype *neglecta*, despite having similar scores along PC1 to other groups (i.e. *L. saxatilis* ecotype *tenebrosa* from Ireland, some *L. saxatilis* H and M, and some *L. arcana*), was clearly distinguished by shell height. One interesting observation was that the wide range of shell size within *L. saxatilis* H and M overlaps with *L. arcana* at one of its extremes, while it does so with *L. compressa* at the other extreme. The unidentified individuals with a jelly gland overlap both in size and shape with the smaller *L. arcana*.

A separate PCA of the *L. saxatilis* individuals revealed that the first four principal components (PC) individually explained more than 5% of the variance, accounting for a total of 84% of the variance in the sample (Fig. II.8a). The variation along PC1 (47%) showed a slight separation between a group comprising *L. saxatilis* ecotypes H, M and *tenebrosa* from Britain, and another group comprising *L. saxatilis* ecotypes *neglecta* and *tenebrosa* from Ireland (Fig. II.8b and Fig. II.8c), with individuals placed towards the latter group having a taller spire and a narrower aperture (Fig. II.9, PC1). The variation along PC2 (21%) showed no obvious clustering pattern of individuals, (Fig. II.8b and Fig. II.8d), although the shape differences suggest that individuals with a wider inner lip have also a smaller and more rounded aperture, as well as a taller spire (Fig. II.9, PC2). The variation along PC3 (11%) showed a separation between *L. saxatilis* ecotypes *neglecta* and *tenebrosa* from Ireland (Fig. II.8c and Fig. II.8d), with individuals towards the *tenebrosa* extreme having clearly a taller spire, a thinner inner lip, and a smaller aperture relative to the body size (Fig. II.9, PC3). The variation along PC4 (6%) showed no obvious clustering pattern of individuals, although the shapes differences observed were associated with individuals having either a wider or a narrower aperture (Fig. II.9, PC4).

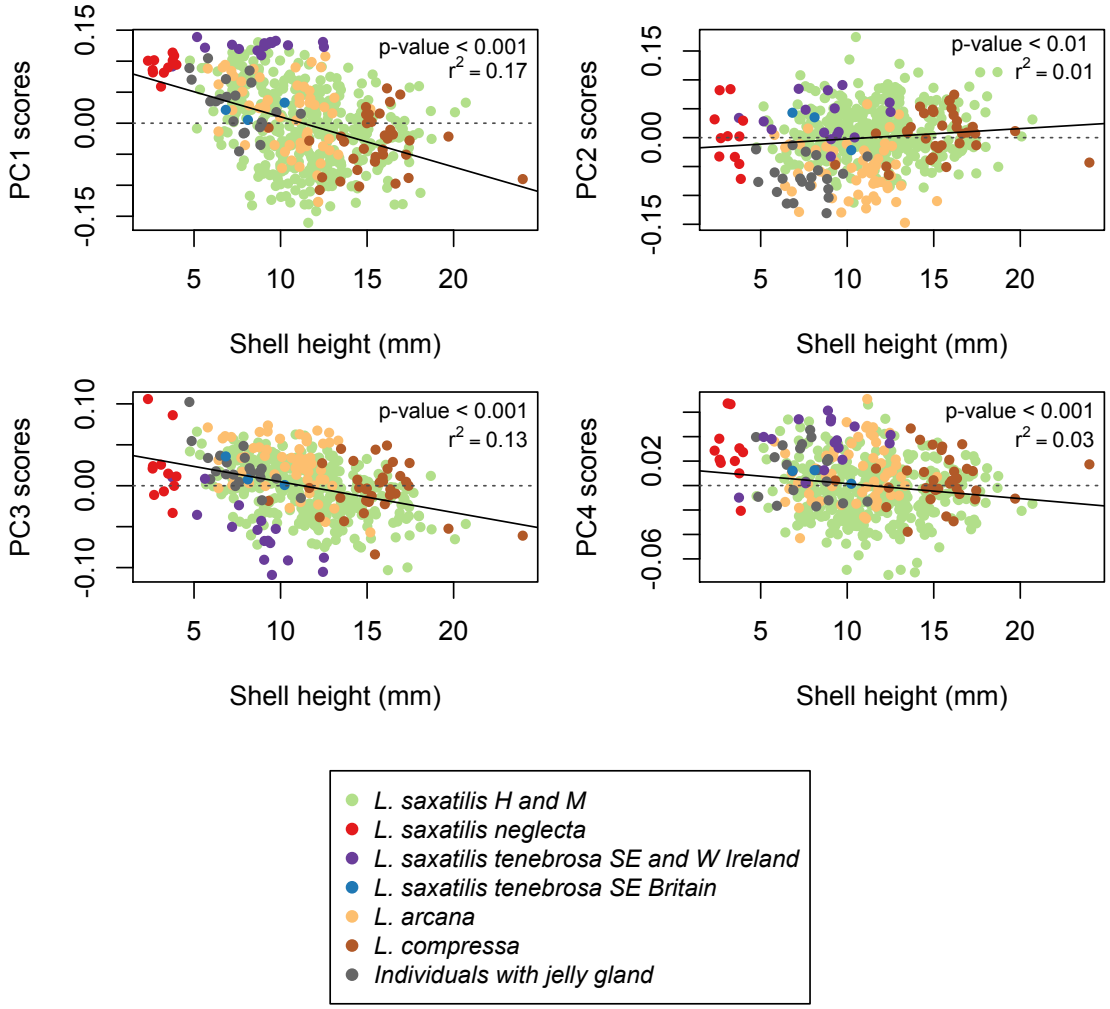


Fig. II.7 Relationship between the first four principal components and shell height across all 504 individuals included in this study. Dots represent single individuals. Scree plot shows the percentage of variation explained by each principal component.

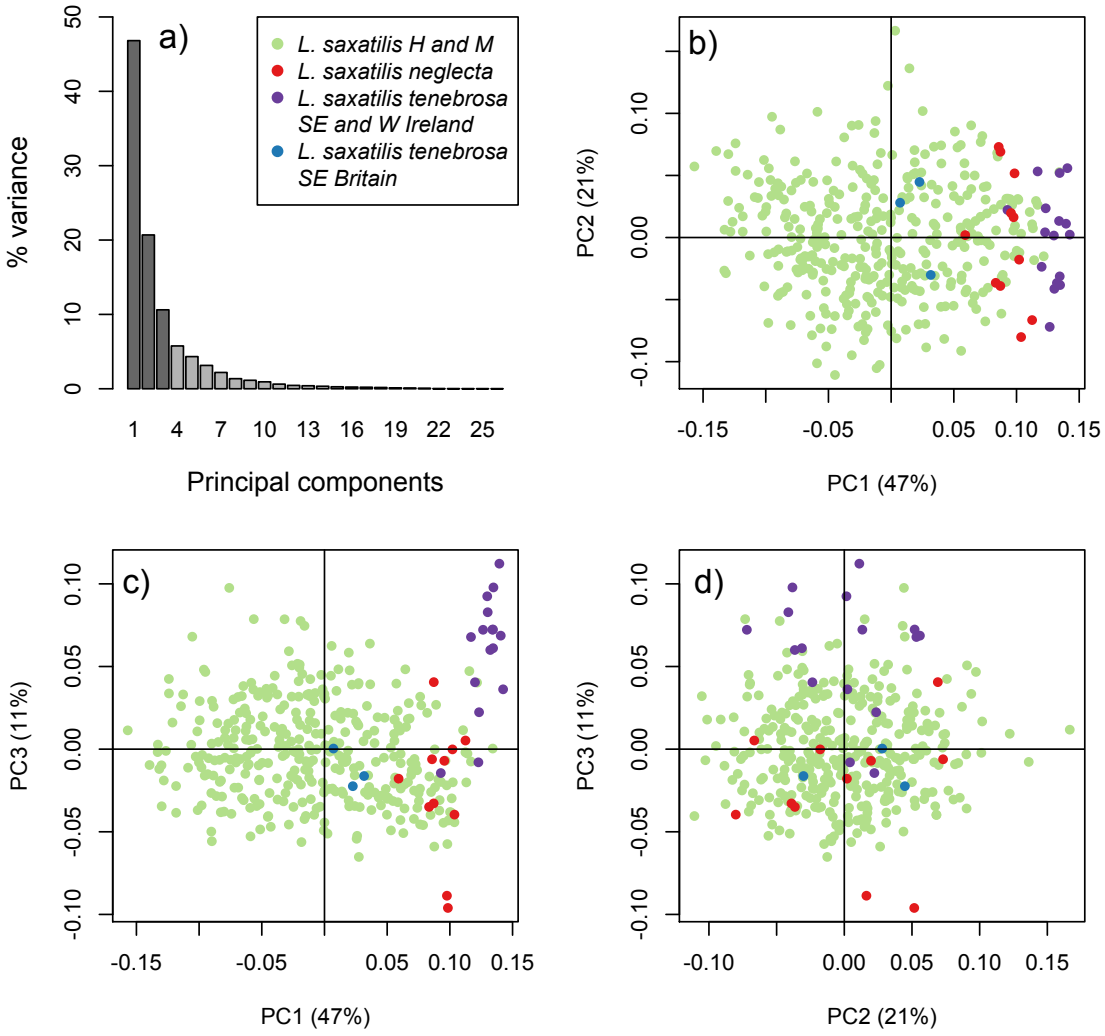


Fig. II.8 Principal component analysis of the 391 *L. saxatilis* individuals. (a) Scree plot of the variance described by each PC. Scatter plots showing scores on (b) PC1 and PC2, (c) PC1 and PC3, and (d) PC2 and PC3; dots represent single individuals.

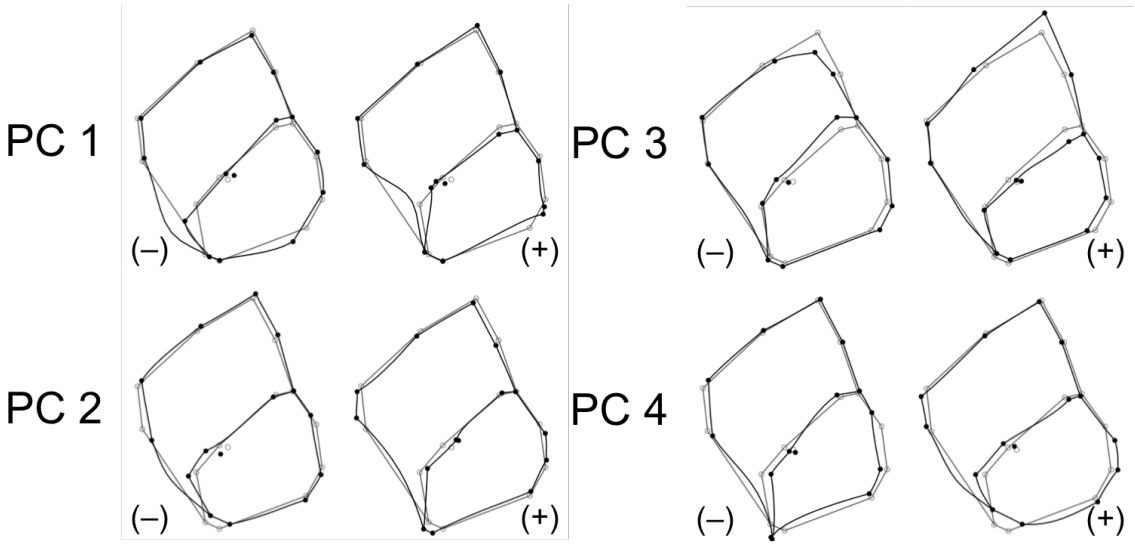


Fig. II.9 Patterns of shape variation along the first four PCs on the 391 *L. saxatilis* individuals. Soft wireframe graphs show the most extreme negative and positive deformations of landmarks (black) with respect to the average configuration (grey). Lines connect some landmarks to facilitate appreciation of shape differences.

A further PCA of the subset of *L. saxatilis* excluding *neglecta* and *tenebrosa* ecotypes led to a very similar result (Fig. II.10, Fig. II.11 and Fig. II.12) to that obtained when including all *L. saxatilis* individuals. The scatter plots were coloured by site (Fig. II.10b – Fig. II.10d) and position on the shore (Fig. II.11b – Fig. II.11d), showing no obvious distinction between H and M ecotypes without an *a priori* assignment of individuals to specific ecotypes. In this exploratory analysis no strong clustering patterns were observed on the basis of site or position on the shore; however, this relationships were specifically tested later on in this work. This could mean that even if individuals grouped by ecotypes, there would be a continuum of intermediate variation between them. The only apparent difference in this PCA is that shape differences along PC3 (Fig. II.12, PC3) are more pronounced than along PC3 of all *L. saxatilis* individuals (Fig. II.9, PC3), i.e. shells are either more vertically compressed and have a bigger aperture, or are more vertically elongated and have a smaller aperture.

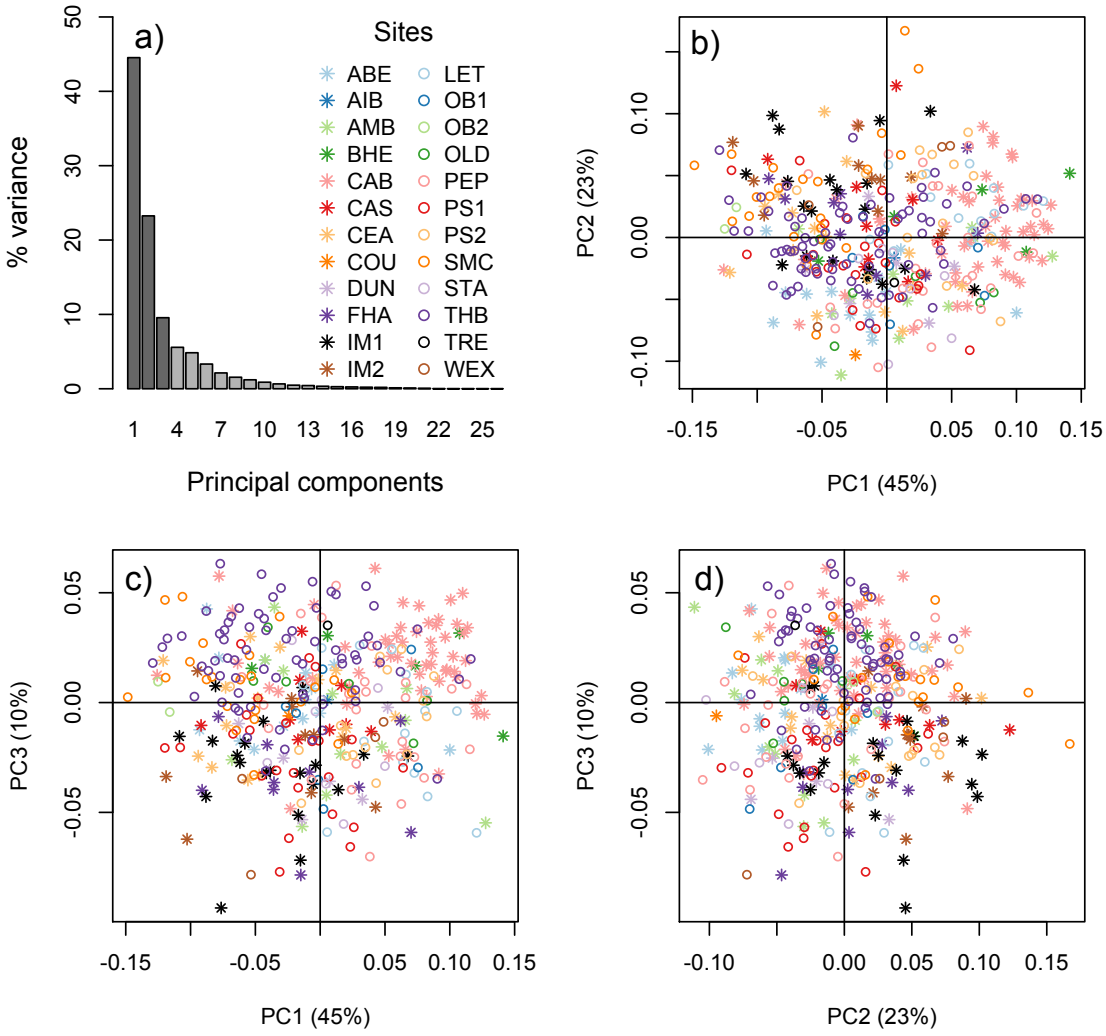


Fig. II.10 Principal component analysis of the 361 *L. saxatilis* ecotypes H and M individuals. (a) Scree plot of the variance described by each PC. Scatter plots showing scores on (b) PC1 and PC2, (c) PC1 and PC3, and (d) PC2 and PC3; dots represent single individuals and are coloured by collection site.

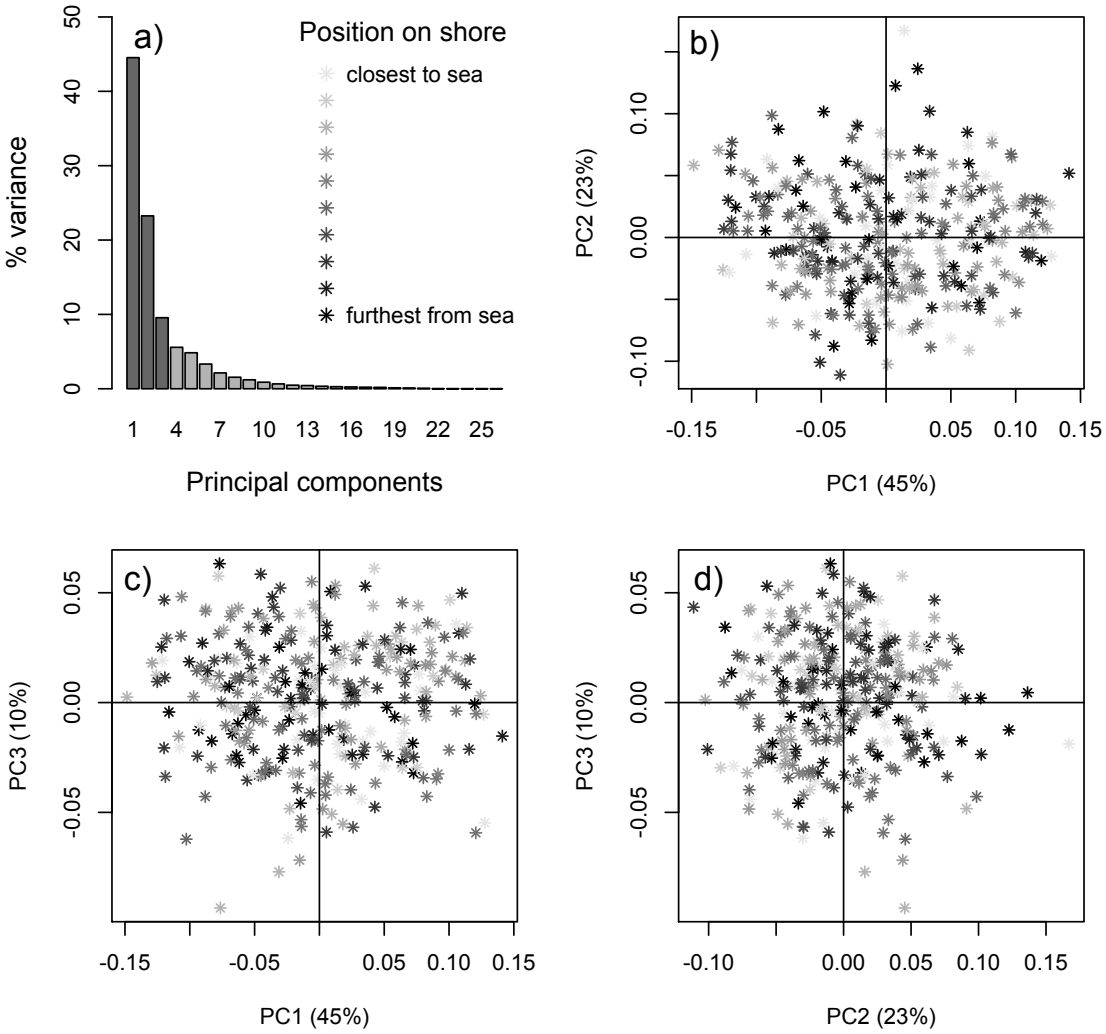


Fig. II.11 Principal component analysis of 324 *L. saxatilis* ecotypes H and M individuals from 20 sites in the British Isles. (a) Scree plot of the variance described by each PC. Scatter plots showing scores on (b) PC1 and PC2, (c) PC1 and PC3, and (d) PC2 and PC3; dots represent single individuals and are coloured by position on the shore.

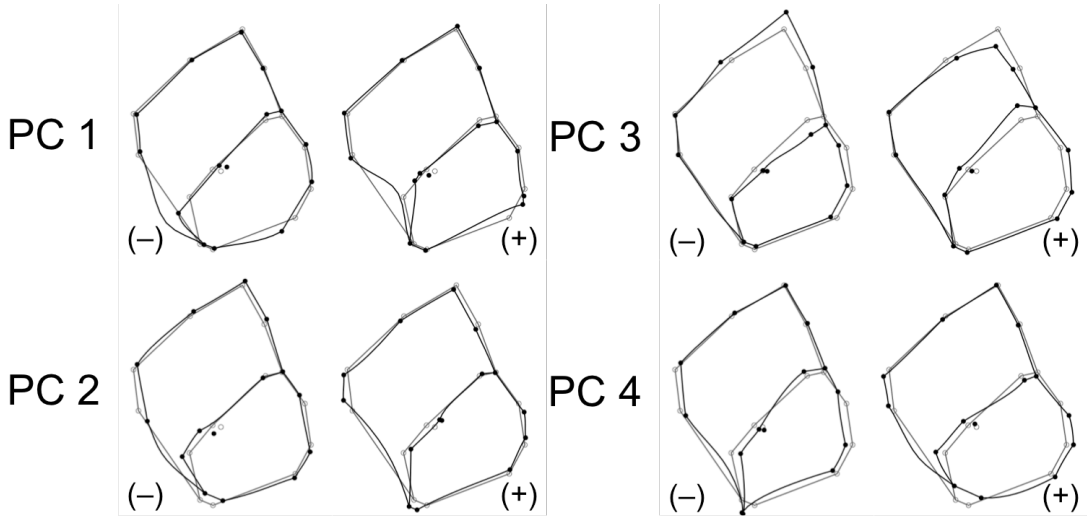


Fig. II.12 Patterns of shape variation along the first four PCs on the 361 *L. saxatilis* ecotypes H and M individuals. Soft wireframe graphs show the most extreme negative and positive deformations of landmarks (black) with respect to the average configuration (grey). Lines connect some landmarks to facilitate appreciation of shape differences.

The PCA of the individuals with a jelly gland, which included *L. arcana*, *L. compressa* and unidentified individuals, revealed that the first four principal components (PC) individually explained more than 5% of the variance, accounting for a total of 83% of the variance in the sample (Fig. II.13a). The variation along PC1 and PC2 together showed a distinction between *L. arcana* and *L. compressa*, with unidentified individuals clearly overlapping with *L. arcana* (Fig. II.13b). The variation along PC1 (40%) was associated with individuals in one extreme having a wider aperture and inner lip (Fig. II.14, PC1), characteristic of *L. arcana*. The variation along PC2 (27%) showed similar variation to PC1, with one extreme of variation having a much wider and more rounded aperture, as well as a thicker lip (Fig. II.14, PC2), again characteristic of *L. arcana*. The variation along PC3 (11%) showed that individuals with a smaller aperture relative to body size clearly had a much taller spire (Fig. II.14, PC3). The variation along PC4 (5%) was in general associated with the shell's globosity, with individuals in one extreme having a more compressed shape along the columella axis and a bigger aperture relative to body size (Fig. II.14, PC4).

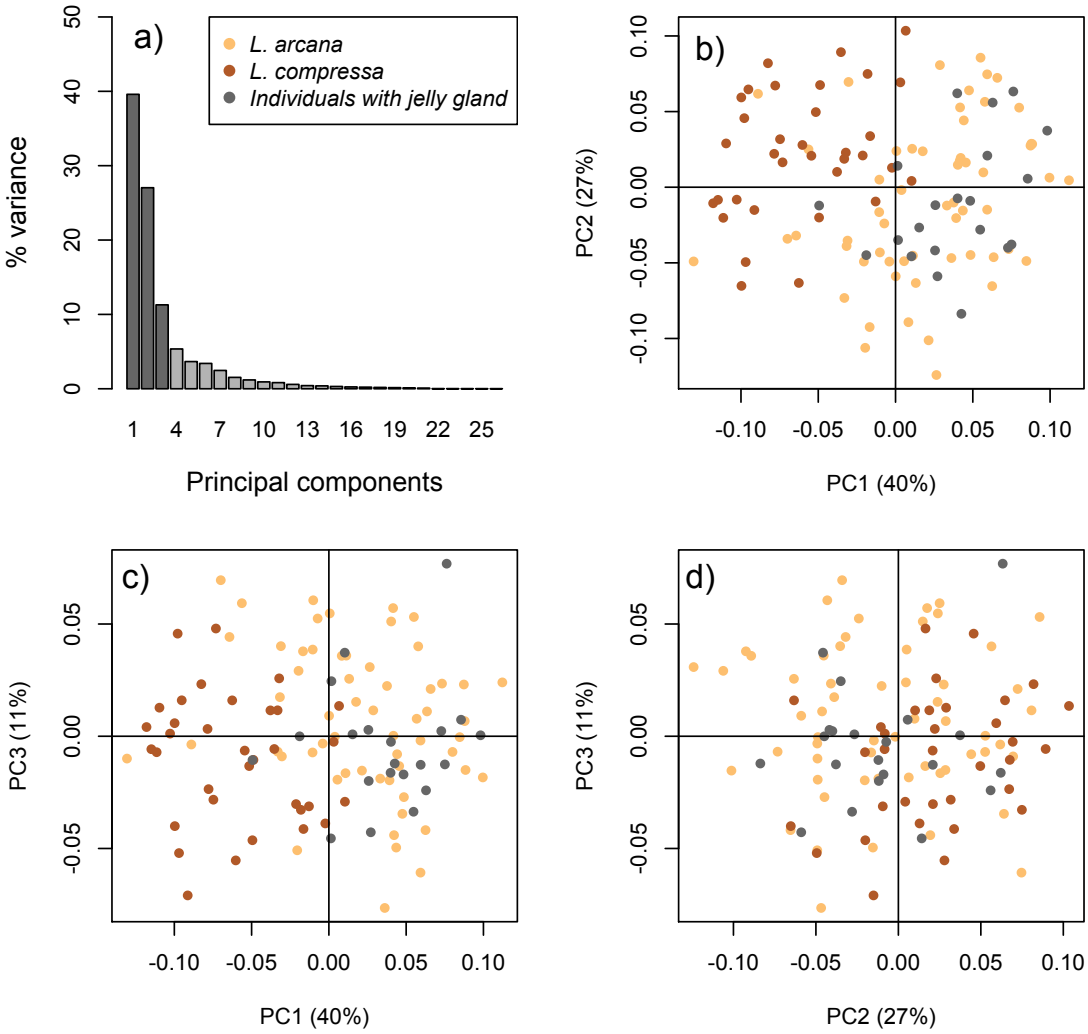


Fig. II.13 Principal component analysis of the 113 *L. arcana* and *L. compressa* individuals. (a) Scree plot of the variance described by each PC. Scatter plots showing scores on (b) PC1 and PC2, (c) PC1 and PC3, and (d) PC2 and PC3; dots represent single individuals.

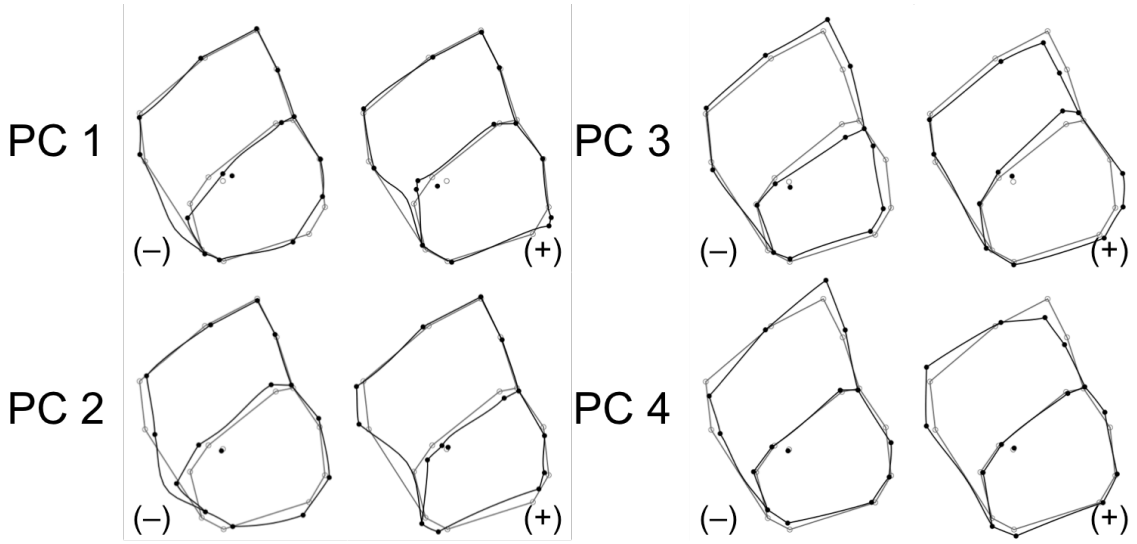


Fig. II.14 Patterns of shape variation along the first four PCs on the 113 *L. arcana* and *L. compressa* individuals. Soft wireframe graphs show the most extreme negative and positive deformations of landmarks (black) with respect to the average configuration (grey). Lines connect some landmarks to facilitate appreciation of shape differences.

The CVA (Fig. II.15) of *L. saxatilis* H and M, *L. arcana* and *L. compressa* individuals revealed the shell traits that vary the most between the three species relative to variation within species. This ordination method differs from PCA in that groups are defined *a priori* so that the differences within these predefined groups are minimised, while the differences between groups are maximised, highlighting thus the features that characterise each group. Variation along CV1 accounted for 74% of the among-group variation (Fig. II.15), and it distinguished *L. saxatilis* H and M from *L. arcana*, with the latter having much wider aperture and inner lip, and slightly more vertically compressed shell (Fig. II.16, CV1). Variation along CV2 accounted for 26% of the total variation (Fig. II.15), and it distinguished *L. saxatilis* H and M from *L. compressa*, with the latter having more rounded overall shape and aperture, and shorter spire (Fig. II.16, CV2). The greatest Mahalanobis distance was found between *L. arcana* and *L. compressa* (3.9; p-value < 0.0001), followed by that between *L. arcana* and *L. saxatilis* H and M (3.1; p-value < 0.0001), and finally by that between *L. compressa* and *L. saxatilis* H and M (2.4; p-value < 0.0001).

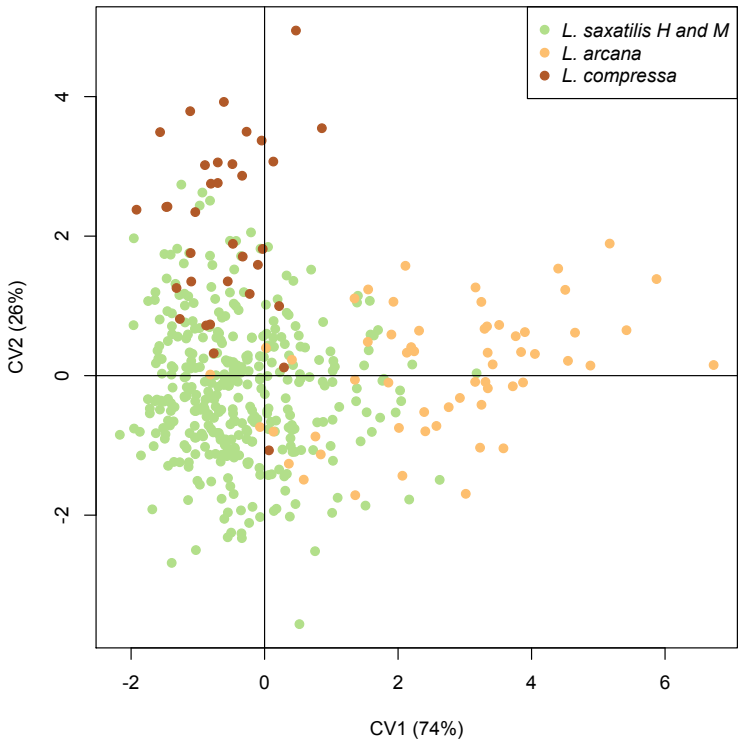


Fig. II.15 Canonical variate scores of the 361 *L. saxatilis* H and M, 58 *L. arcana*, and 34 *L. compressa* individuals. Dots represent single individuals.

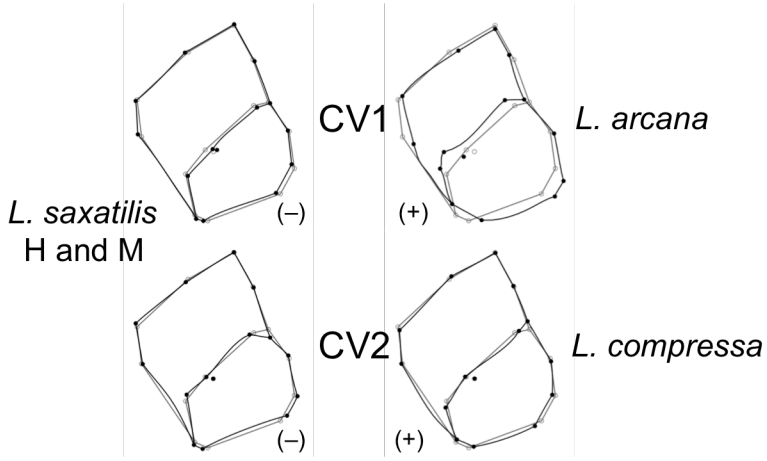


Fig. II.16 Patterns of shape differences along the CV1 and CV2 depicted as soft wireframe graphs that show the most extreme negative and positive deformations of landmarks (black) with respect to the average configuration (grey). Lines connect some landmarks to facilitate appreciation of shape deformation.

The three discriminant function analyses (Fig. II.17) – one for each possible pair of species (*L. saxatilis* H and M, *L arcana* and *L. compressa*) – revealed different degrees of shape differences between species. According to these results, the clearest separation was between *L. arcana* and *L. compressa* (Fig. II.17c), followed by a less clear separation between *L. saxatilis* H and M and *L. arcana* (Fig. II.17a), and finally by the least clear separation between *L. saxatilis* H and M and *L. compressa* (Fig. II.17b). The highest rates of correct classification in the cross-validation analyses were those between *L. arcana* and *L. compressa* (90% in both cases), followed by those between *L. saxatilis* H and M (93%) and *L. arcana* (83%), and finally by those between *L. saxatilis* H and M (86%) and *L. compressa* (74%). In comparison with *L. saxatilis* H and M and *L. compressa*, *L. arcana* had a much wider aperture and inner lip, and a shorter spire (Fig. II.18). The differences between *L. saxatilis* H and M and *L. compressa* were subtler, with the latter being more laterally compressed and having a more elongated last whorl (Fig. II.18).

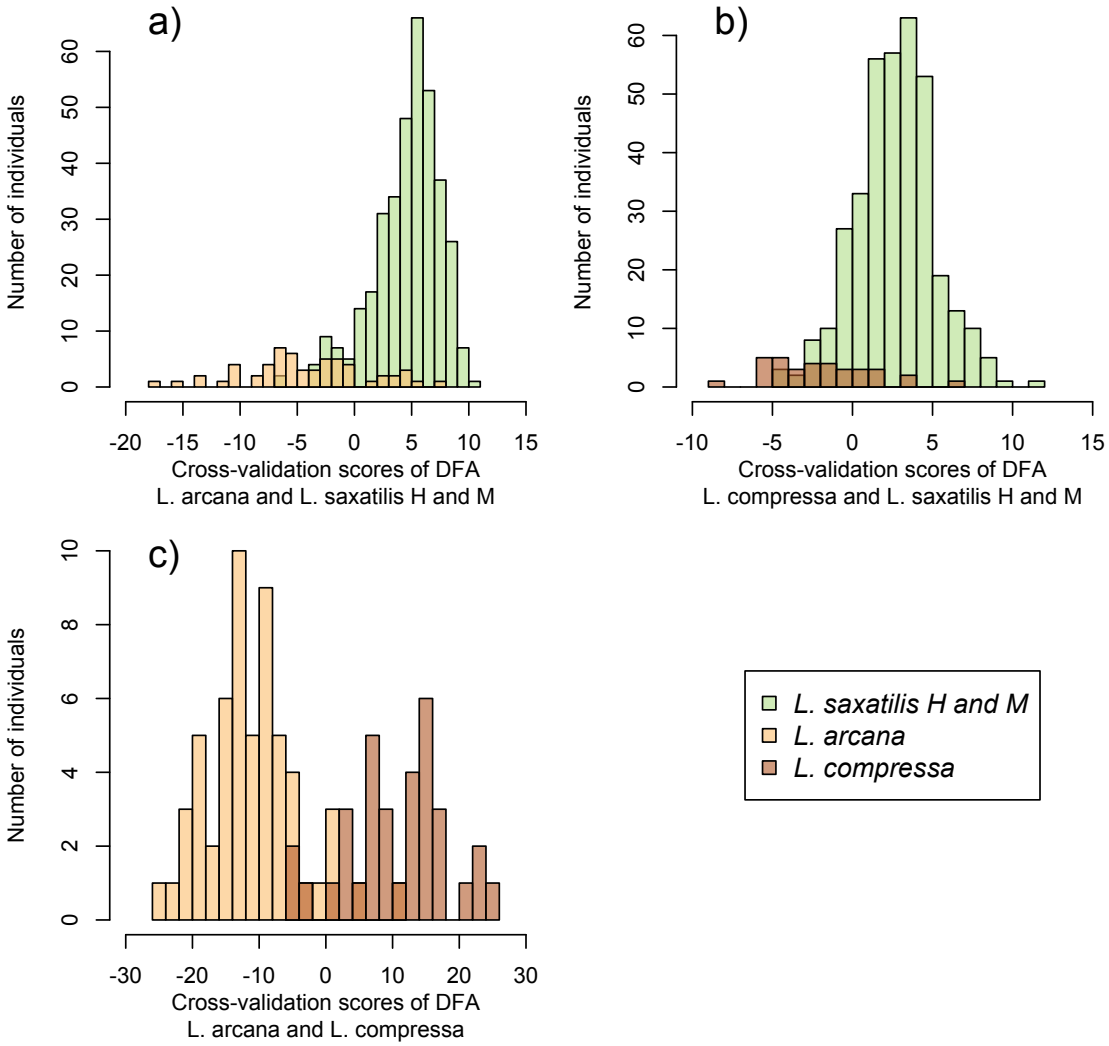


Fig. II.17 Frequency of the cross-validation scores of the discriminant function analysis of the 361 *L. saxatilis* H and M, 58 *L. arcana*, and 34 *L. compressa* individuals.

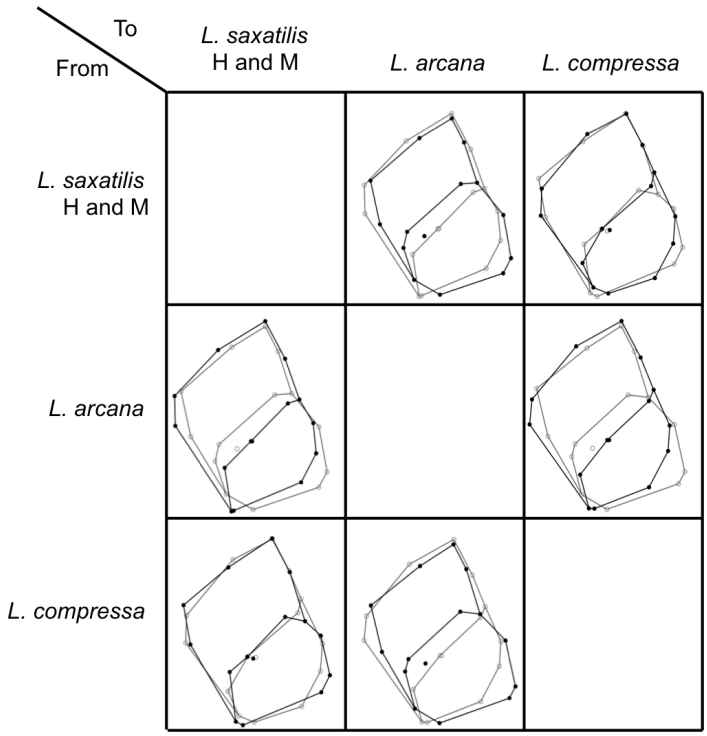


Fig. II.18 Pairwise shape differences between the three species means shown as the change from one species to another, as computed by the DFA. Lines connect some landmarks to facilitate appreciation of shape changes as wireframe graphs. Starting shapes are grey, and target shapes are black.

Covariation of shape with other variables

Simplification of maximal models led in all cases to minimal adequate models (Table II.3) that significantly differed from their respective null model (p-value < 0.05). The fraction of the variance in shape explained by each model was substantial (Table II.3, adjusted r^2). This means that the models described the data well, and shape was significantly influenced by environment. The variation in each PC was explained by a different combination of explanatory variables (Table II.4). However, model simplification procedures all had in common the deletion of specific explanatory variables – namely geographical region, latitude, bedrock geology, MDS2, and MDS3 – whose inclusion in the models did not significantly improve the goodness of the fit, so justifying their exclusion.

Table II.3 Minimal models for each of the principal components as computed in the morphometric analysis over all individuals.

Response variable	Explanatory variables	Adjusted r^2
PC1	Site + sitting structure + position + MDS1 + MDS4 + CS	0.40
PC2	Site + sitting structure + rep. structure + position + MDS1	0.40
PC3	Site + rep. structure + position + MDS5 + CS	0.38
PC4	Site + rep. structure + position + CS	0.27

Table II.4 Analysis of covariance for the four fitted models (i.e. one for each principal component analysed) showing the significance of each model term.

Response variable	Explanatory variable	Df[1]	Sum Sq[2]	Mean Sq[3]	F value	Pr(>F)[4]	
PC1	Site	3	0.5461	0.1820	76	< 2.2e–22	***
	Sitting structure	1	0.0479	0.0479	20	9.6e–06	***
	Centroid size	1	0.0461	0.0461	19	1.4e–05	***
	MDS1	1	0.0244	0.0244	10	1.5e–03	**
	Position	1	0.0165	0.0165	7	8.9e–03	**
	MDS4	1	0.0110	0.0110	5	3.2e–02	*
	Residuals	422	1.0060	0.0024			
PC2	Site	5	0.2974	0.0595	41	< 2.2e–22	***
	Rep. structure	1	0.0724	0.0724	50	5.9e–12	***
	Sitting structure	1	0.0434	0.0434	30	7.3e–08	***
	MDS1	1	0.0087	0.0087	6	1.5e–02	*
	Position	1	0.0028	0.0028	2	1.7e–01	
	Residuals	421	0.6078	0.0014			
PC3	Site	4	0.1347	0.0337	56	< 2.2e–22	***
	Rep. structure	1	0.0139	0.0139	23	2.2e–06	***
	Centroid size	1	0.0134	0.0134	22	3.3e–06	***
	Position	1	0.0011	0.0011	2	1.7e–01	
	MDS5	1	0.0007	0.0007	1	2.9e–01	
	Residuals	422	0.2549	0.0006			
PC4	Site	4	0.0387	0.0097	24	< 2.2e–22	***
	Centroid size	1	0.0163	0.0163	41	5.0e–10	***
	Position	1	0.0068	0.0068	17	4.9e–05	***
	Rep. structure	1	0.0039	0.0039	10	2.1e–03	**
	Residuals	423	0.1701	0.0004			

Significance codes: 0 *** 0.001 ** 0.01 * 0.05
1 = degrees of freedom; 2 = sums of squares; 3 = mean of squares; 4 = p-value

Scatterplots of the residuals against the fitted values revealed no obvious bias; however, small signs of heteroscedasticity were detected in all cases, i.e. the variance of the residuals was not constant across the fitted values of shape. This might be due to a greater variance being associated with particular values of the response variable. Alternatively, this pattern could be observed in highly structured data like, for example, a bimodal distribution of the response variable (e.g. discrete ecotypes without continuous shape variation connecting the two extremes). Also, using PC scores obtained in an analysis of all individuals, and then excluding some individuals in the statistical modelling may have had an impact on the results. Scatterplots of the standardised residuals against the leverage showed up to three highly influential data points in each model. Therefore, future studies should consider repeating these analyses without these points. Biological interpretation of the models presented here should be

made with caution considering that the models and data collection can be improved.

Variation in MDS1, MDS4 and MDS5 was significantly associated with variation in shape as measured by the PCs analysed in this study. Negative values of MDS1 were mostly associated with the presence of supralittoral lichens, *Melarhaphe neritoides*, and crabs, whereas positive values were associated with the presence of *Littorina littorea*, *Elminius modestus*, and *Patella*. Lichens and *Melarhaphe neritoides* are typical high shore species, whereas *Littorina littorea*, *Elminius modestus* and *Patella* are usually found closer to the sea. However, the association of crabs with high shore species makes it difficult to interpret the biological meaning of MDS1 axis. Negative values of MDS4 were mostly associated with the presence of supralittoral lichens, *Mytilus edulis*, and *Melarhaphe neritoides*, whereas positive values were associated with the presence of *Ulva*, *Elminius modestus*, and *Fucus vesiculosus*. Finally, negative values of MDS5 were mostly associated with the presence of supralittoral lichens and *Elminius modestus*, whereas positive values were associated with the presence of *Fucus spiralis*, crabs, and *Ulva*.

Variation in shape as described by PC1 was associated with site, sitting structure, position on the shore, MDS1, MDS4, and centroid size. Rounder shell shape with rounder aperture was associated with the presence of *Littorina littorea*, *Elminius modestus*, and *Patella*, whereas narrower aperture was associated with the presence of supralittoral lichens, *Melarhaphe neritoides*, and crabs (Fig. II.19). Model simplification led to the aggregation of sites into four groups with no evident clustering pattern related to geography, and also to aggregation of sitting structures into two categories: barnacles and all other structures (Fig. II.19). This was done following the methodology of model simplification previously described, which indicates that individual factor levels that do not significantly improve the model should be aggregated. Within the four groups of sites, snails sitting on barnacles had a less round shell with narrower aperture compared to those that were sitting on any other substrate.

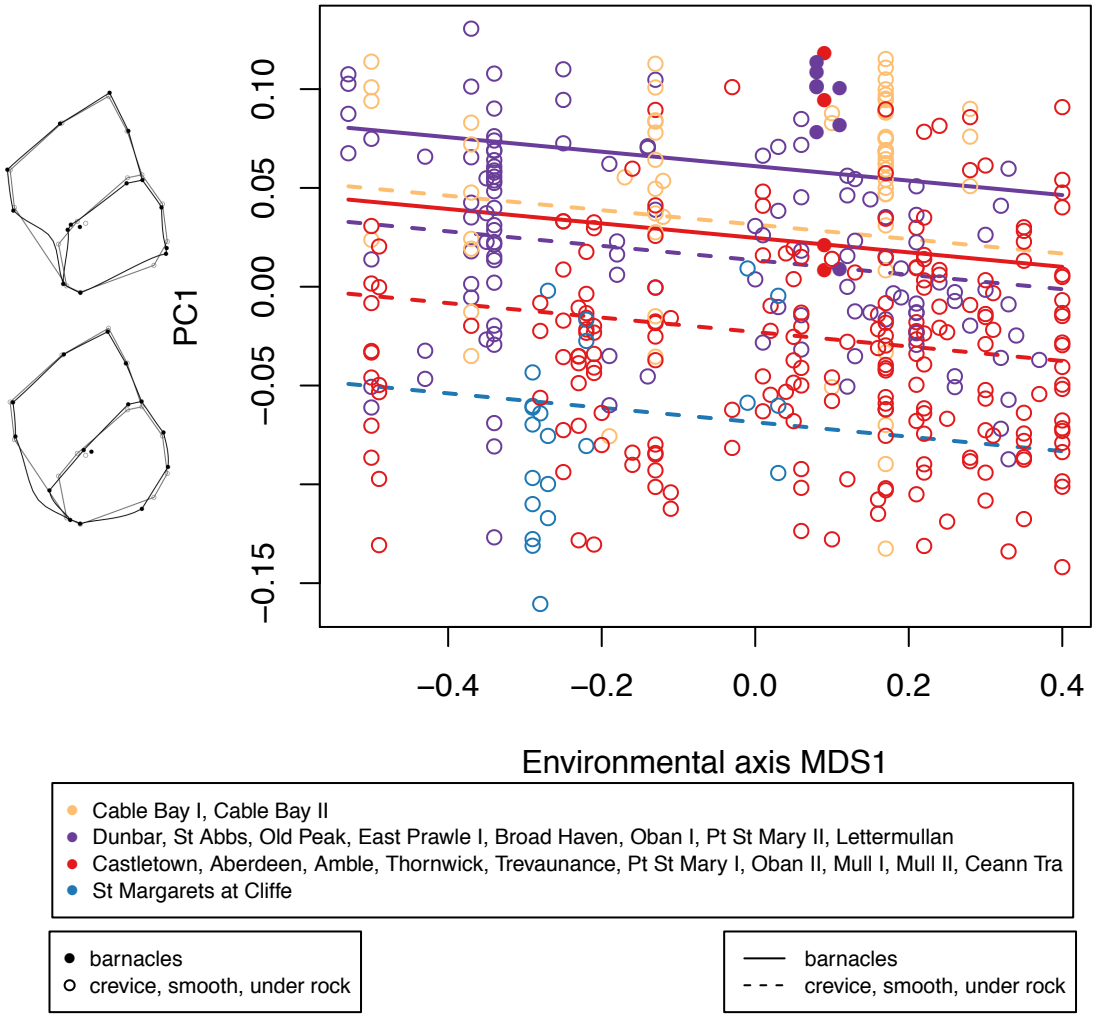


Fig. II.19 Relationship between shape (PC1) and environment (MDS1) across 431 individuals of the *L. saxatilis* complex. Colours indicate site group; symbols and lines indicate sitting structure. Lines were fitted according to model 1.

Variation in shape as described by PC2 was associated with site, sitting structure, reproductive structure, position on the shore, and MDS1. Shells with shorter spire and much wider aperture were associated with the presence of supralittoral lichens, *Melarhaphe neritoides*, and crabs, whereas shells with taller spire and narrower aperture were associated with the presence of *Littorina littorea*, *Elminius modestus*, and *Patella* (Fig. II.20). Model simplification led to the aggregation of sites into six groups with no clear clustering pattern related to geography, and also to aggregation of sitting structures into two categories: under rock and all other structures (Fig. II.20). Within the six groups of sites, snails found under rocks had taller spire and narrower aperture compared to those that were found on any other substrate.

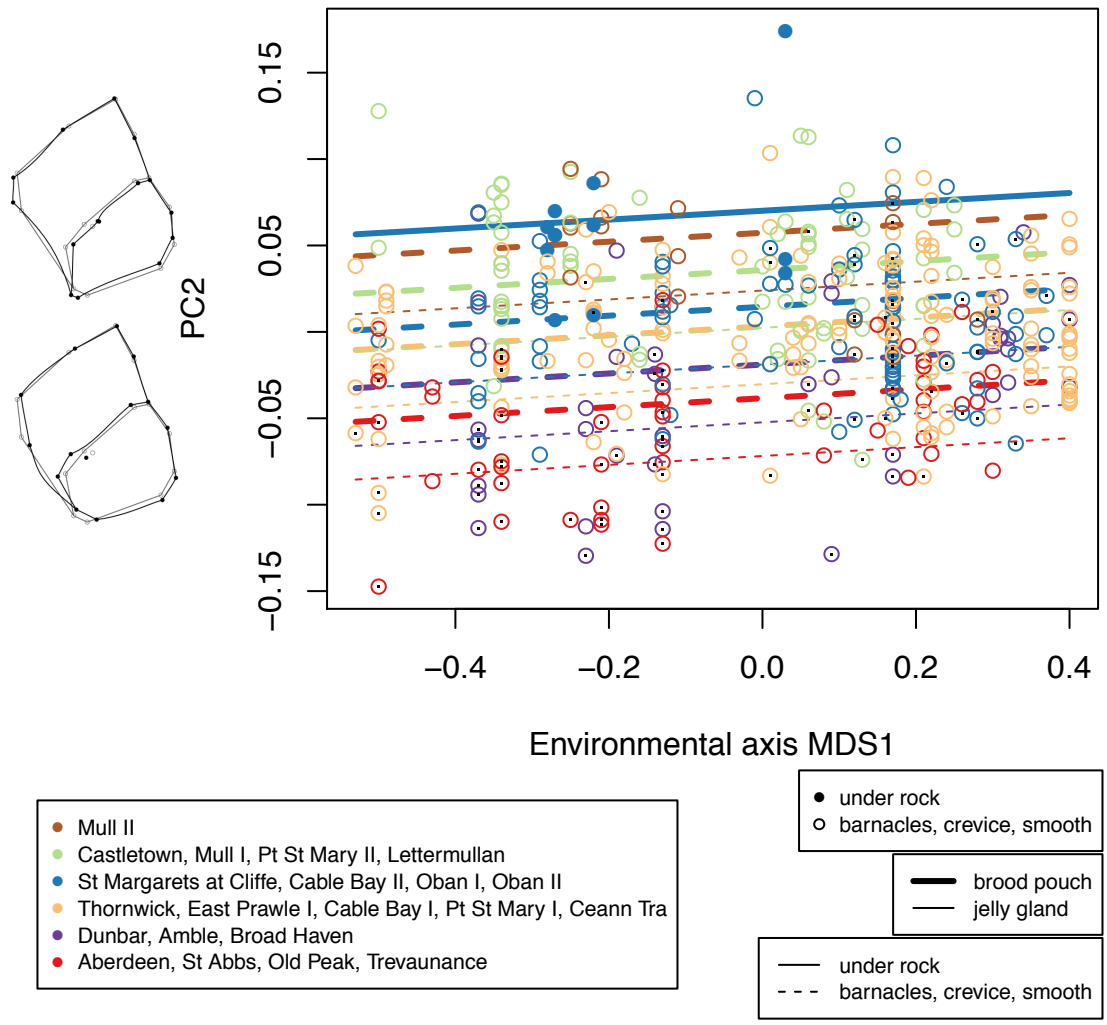


Fig. II.20 Relationship between shape (PC2) and environment (MDS1) across 431 individuals of the *L. saxatilis* complex. Colours indicate site group; symbols and lines indicate sitting structure; black dots "." identify individuals with a jelly gland. Lines were fitted according to model 2.

Variation in shape as described by PC3 was associated with site, reproductive structure, position on the shore, MDS5, and centroid size. Shells with shorter spire and much wider aperture were associated with the presence of *Fucus spiralis*, crabs, and *Ulva*, whereas shells with taller spire and narrower aperture were associated with the presence of supralittoral lichens and *Elminius modestus* (Fig. II.21). Model simplification led to the aggregation of sites into five groups with no clear clustering pattern related to geography (Fig. II.21). Within the five groups of sites, snails with a brood pouch had a taller spire and narrower aperture compared to those that had a jelly gland.

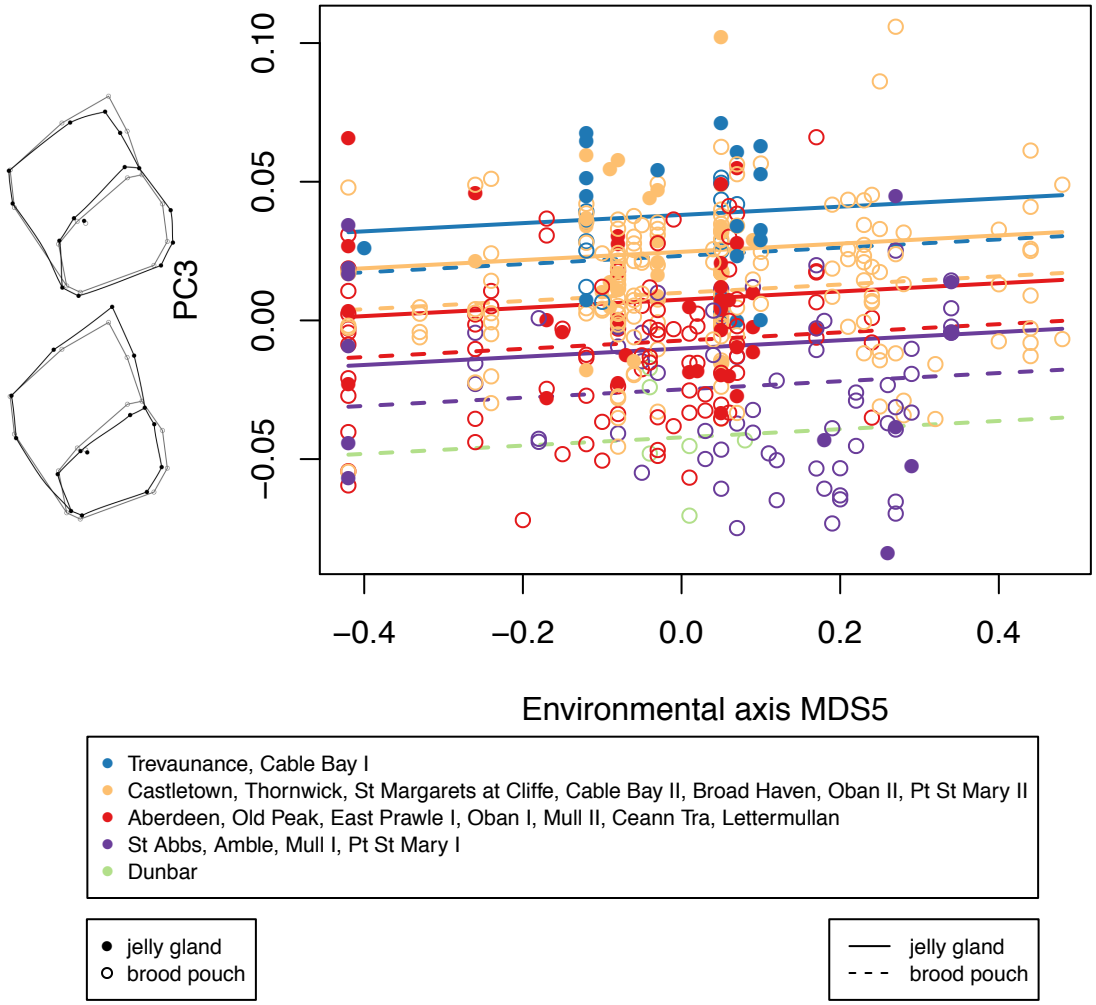


Fig. II.21 Relationship between shape (PC3) and environment (MDS5) across 431 individuals of the *L. saxatilis* complex. Colours indicate site group; symbols and lines indicate reproductive structure. Lines were fitted according to model 3.

Variation in shape as described by PC4 was associated with site, reproductive structure, position on the shore, and centroid size. Shells with shorter spire and narrower aperture were associated with positions closest to the sea, whereas shells with taller spire and wider aperture were associated with positions furthest from the sea (Fig. II.22). Model simplification led to the aggregation of sites into five groups with no clear clustering pattern related to geography (Fig. II.22). Within the five groups of sites, snails with brood pouch had shorter spire and narrower aperture compared to those that had jelly gland.

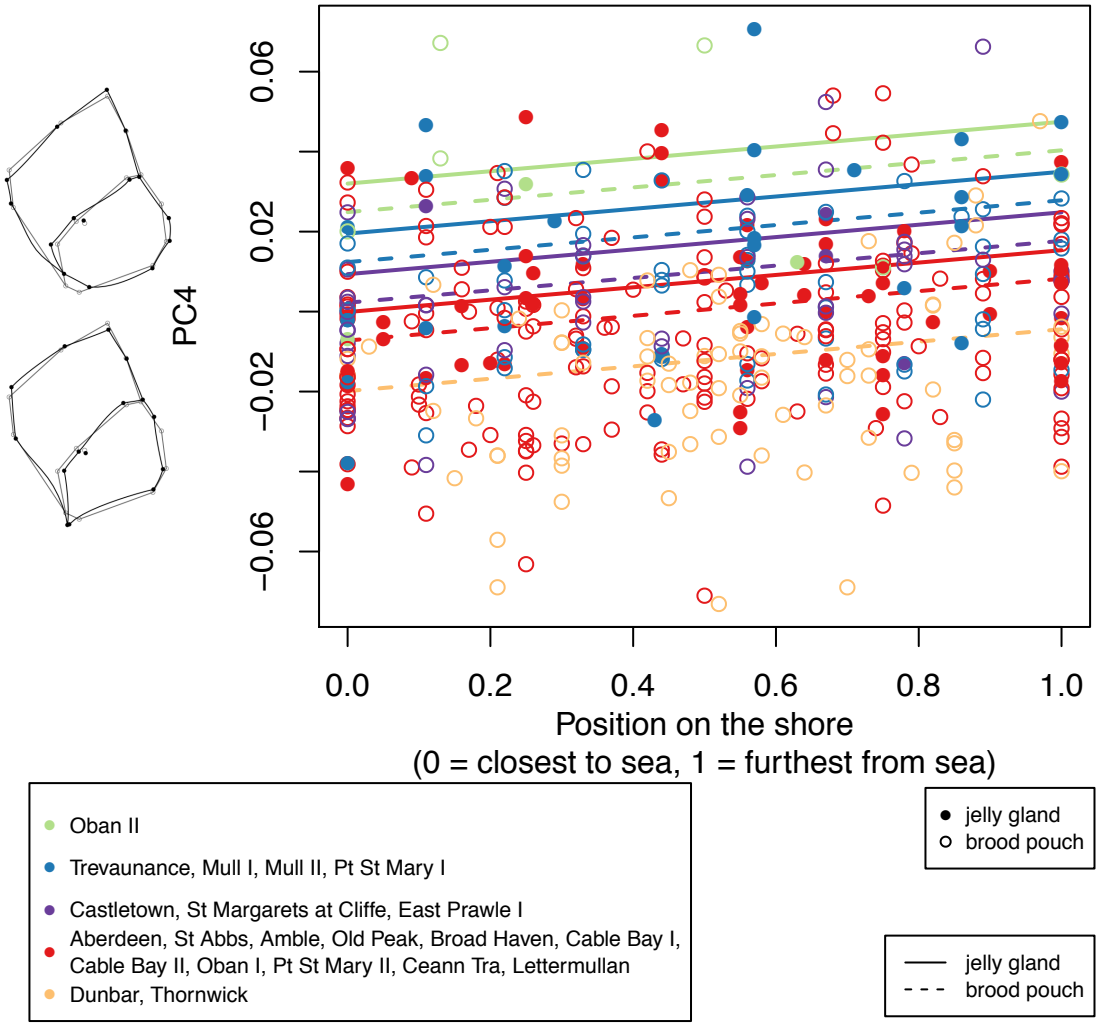


Fig. II.22 Relationship between shape (PC4) and position of the shore across 431 individuals of the *L. saxatilis* complex. Colours indicate site group; symbols and lines indicate reproductive structure. Lines were fitted according to model 4.

Discussion

A wide diversity of shell shape was found among species of the *Littorina saxatilis* complex, as well as within *L. saxatilis*. The most pronounced interspecific difference in shape and size was that between *L. arcana* and *L. compressa*. The observed shapes of these two species met the expectations, according to which snails from wave-exposed habitats (*L. arcana*) tend to have a stronger foot to prevent dislodgement, which in turn requires a wider aperture (Johannesson 1986). On the other hand, snails from sheltered habitats (*L. compressa*) tend to face higher levels of predation, and thus a narrower aperture is advantageous (Johannesson 1986).

Ecotypes H and M of *L. saxatilis* could not be unambiguously separated into discrete categories either *a priori* or *a posteriori*, suggesting the existence of

continuous variation between the two extremes when considering all sites together. However, the possibility that this variation could be discontinuous within sites should be further explored. It has previously been found that intermediate forms between different ecotypes exist where the typical microhabitats of those ecotypes meet (Johannesson & Johannesson 1990; Johannesson *et al.* 1993). However, considerable overlap both in shape and size was found between the group comprising ecotypes H and M of *L. saxatilis* with either *L. arcana* or *L. compressa*. This could be a reflection of habitat overlap between *L. saxatilis* H and *L. arcana*, and between *L. saxatilis* M and *L. compressa*. Given this habitat overlap, the observed separation of *L. saxatilis* from the other two species in the complex was reasonably good.

However, sampling along transects that were perpendicular to the shore may have introduced a bias in terms of sample sizes towards particular species and/or ecotypes. This could be due to the fact that not all habitats were equally sampled. This means that, for example, *L. saxatilis* M – most commonly found on the mid shore – may be overrepresented due to the fact that transects mostly fell within this habitat. Therefore, the observed separation between *L. saxatilis* and *L. arcana* may specifically represent a separation between *L. saxatilis* ecotype M and *L. arcana*. Moreover, it has been suggested that *L. arcana* migrates downwards in order to lay its egg masses in damp sites (Hannaford-Ellis 1985). So, future studies should consider this breeding migration when planning the collection of samples along vertical transects.

As previously mentioned, the aim of sampling along transects was to reduce the possibility of actively targeting snails that looked more like either H or M ecotype of *L. saxatilis*, and to capture the potential diversity that would have been missed should the snails had been collected targeting specific shapes or habitats. However, the aforementioned bias was not considered at the time. Future studies should first consider the distribution of habitats within specific sites (e.g. Sinfield 2008) in order to design a sampling scheme that is able to collect equal numbers of snails from the different habitats.

Ecotypes of *L. saxatilis* are similar phenotypes found in similar habitats, and they are partly genetically determined (Hollander *et al.* 2006b; Conde-Padín *et al.* 2007a, 2009). Convergence of phenotypes due to natural selection – and not

common ancestry – leads to the formation of ecotypes, which do not necessarily share the same genetic basis (Westram *et al.* 2014). According to this ecotype definition, the fact that *L. saxatilis* ecotype *neglecta* formed a coherent shape unit was expected. However, *L. saxatilis* ecotype *tenebrosa* formed two clearly separated groups (Ireland and Britain). At first, this finding may seem to differ from the ecotype concept because the two groups clearly differed in shape, despite both of them having been collected in lagoons.

There are a few possible explanations for the latter observation, one of them being that the separation seen here is a reflection of incomplete sampling, and so missing shapes. Another possibility is that the separation might seem larger than it actually is due to its examination along with the large variation observed among *L. saxatilis* ecotypes H and M. Yet another possible reason is that the two groups do differ that much in shape, but still might share other non-shape phenotypic traits, for instance, shell sculpture and tessellation, as well as metabolic and behavioural traits.

Given that shape is expected to be strongly associated with environment, it was expected that *L. saxatilis* ecotype *tenebrosa* both from Ireland and Britain would be rather more similar than observed in the current study. However, the time since colonisation and connectivity with adjacent environments should be taken into account when analysing shape. The pits in Wexford (SE Ireland), where *L. saxatilis* ecotype *tenebrosa* was collected, were isolated from the sea 115 years ago, and permanently filled with saltwater 40 years ago due to the construction of a site barrier (Small & Gosling 2000a), isolating snails in those pools from the outside shingle beach. In this system, *L. saxatilis* ecotype *tenebrosa* was found to have only one haplotype (mtDNA), which was also the most common haplotype on the outside beach, suggesting that *tenebrosa* ecotype had only about 25 years (at the time) to become established in the pit pools (Small & Gosling 2000a). Yet, the present study found that *tenebrosa* from this site had a very similar shape to *tenebrosa* from the western Irish lochs, despite some of these lochs having remained relatively stable for several thousands of years (Healy *et al.* 1997), but perhaps with constant connectivity with the shore animals (Small & Gosling 2000a).

On the other hand, the pits near Dersingham (SE Britain), where *L. saxatilis* ecotype *tenebrosa* were also collected, were dug out about 70 years ago for shingle extraction. Considering only time since habitat formation, the Wexford case would suggest that *tenebrosa* from Dersingham would have had enough time to reach a more typically *tenebrosa* shape. Furthermore, this pattern would be expected given the fact that the pits are situated in an extensive area of the coast with mudflats where *L. saxatilis* is not found, creating a gap in its distribution along the eastern coast of Britain. Given that the Dersingham pits are currently a natural reserve for birds, it is possible that snails may have arrived by hitchhiking on birds (Rees 1965).

Regarding the statistical modelling of shape, there are several ways in which the analysis can be improved. However, the models fitted the data better than the null models, explaining 27–40% of shape variation, suggesting that the explanatory variables kept in the minimal models do have effects on shape, although these effects are not easily interpretable. Only general trends can be observed given the current results.

For instance, geographical regions and latitude did not have an effect on shape; a result that was expected given that shell shape varies even within sites (i.e. ecotypic variation), perhaps masking any broad-scale geographical trend that may exist. However, further studies should consider clustering sites into geographical groups based on a population structure analysis, similar to that presented in Chapter III. Only two geographical varieties have been described for each of the three species investigated here (described by Reid 1996), and all the sampled sites fall within the range of one geographical variety. The aggregation of sites into groups with no particular geographical pattern resulting from the stepwise deletion procedure suggested that shape is not so dependent on geography. Instead, the similarities between distant geographical sites might perhaps be due to the presence of similar environments. Distribution of neither of the species examined here seems to be affected by composition of rocks, the only substrate requirements being hard surfaces where *L. compressa* and *L. arcana* can lay their egg masses (Reid 1996), and non-friable and erosion-resistant rocks that can sustain enough microalgae for the snails to feed on (Mill & Grahame 1990). Therefore, the deletion of bedrock geology from the models was reasonable.

One of the criticisms of using either presence/absence or abundance of species to describe environmental conditions is that it is not clear what, if anything, is being measured (Ballantine 1961). However, given that rocky shores of the British Isles are subject to a wide tidal range, strong patterns of zonation are observed, where most species are restricted to certain levels of the shore (Ballantine 1961). However, the MDS axes examined here did not appear to represent clear tidal level or exposure gradients and, therefore, it is difficult to interpret the way shapes were associated with these axes. Also, an effort should be made to identify indicator taxa to the species level. The reason for this is that similar species may have significantly different distributions, and recording those species as though they were only one would likely have a great impact on the results. Additionally, abundance rather than presence/absence should be used for a more accurate description of environment (Ballantine 1961).

In summary, shell shape was found to be very variable among species of the *Littorina saxatilis* complex, and also within *L. saxatilis*. Shell size and shape allowed a clear separation of *L. arcana* and *L. compressa*. Despite no distinction between *L. saxatilis* H and M being made, the overlap of this group with *L. arcana* and *L. compressa*, respectively, suggested the presence of these two ecotypes in the sample. Within *L. saxatilis*, ecotype *neglecta* had a distinct shape and size, whereas ecotype *tenebrosa* clustered into two rather distinct shape groups related to geography. Several suggestions have been presented as to how fieldwork can be improved, including prior assessment of sites, equal representation of habitats, and more accurate measurement of environment. Modelling using environmental variables suggested that fine-scale conditions (site and microhabitat) have an effect on shape, whereas broad-scale elements (geographical region and bedrock composition) do not. Future studies should consider modelling shell size as a function of the environment, as well as species and ecotypes as explanatory variables.

Acknowledgements

I am thankful to the following people and institutions for the help provided to carry out this research. John Grahame (University of Leeds) provided detailed information about some of the collection sites. The Society for the Study of

Evolution (SSE) provided travel expenses for fieldwork through the Rosemary Grant Graduate Student Research Award. Sophie Webster provided guidance on data collection and initial morphometric analyses. The following people helped during fieldwork: Steeves Buckland, Sara Cruz, Sara Hintz, Coralie Hopwood, Kerstin Johannesson, Anne-Lise Liabot, Cheryl Mills, Sophie Webster, and Anja Westram.

References

Ballantine WJ (1961) A biologically-defined exposure scale for the comparative description of rocky shores. *Field Studies Journal*, **1**, 1–19.

Bookstein FL (1991) *Morphometric tools for landmark data: geometry and biology*. Cambridge University Press, Cambridge, UK.

Caley KJ, Grahame J, Mill PJ (1995) A geographically-based study of shell shape in small rough periwinkles. *Hydrobiologia*, **309**, 181–193.

Carvajal-Rodríguez A, Conde-Padín P, Rolán-Alvarez E (2005) Decomposing shell form into size and shape by geometric morphometric methods in two sympatric ecotypes of *Littorina saxatilis*. *Journal of Molluscan Studies*, **71**, 313–318.

Conde-Padín P, Caballero A, Rolán-Alvarez E (2009) Relative role of genetic determination and plastic response during ontogeny for shell-shape traits subjected to diversifying selection. *Evolution*, **63**, 1356–1363.

Conde-Padín P, Carvajal-Rodríguez A, Carballo M, Caballero A, Rolán-Alvarez E (2007a) Genetic variation for shell traits in a direct-developing marine snail involved in a putative sympatric ecological speciation process. *Evolutionary Ecology*, **21**, 635–650.

Conde-Padín P, Grahame JW, Rolán-Alvarez E (2007b) Detecting shape differences in species of the *Littorina saxatilis* complex by morphometric analysis. *Journal of Molluscan Studies*, **73**, 147–154.

Cruz R, Rolán-Alvarez E, García C (2001) Sexual selection on phenotypic traits in a hybrid zone of *Littorina saxatilis* (Olivi). *Journal of Evolutionary Biology*, **14**, 773–785.

Doellman MM, Trussell GC, Grahame JW, Vollmer S V. (2011) Phylogeographic analysis reveals a deep lineage split within North Atlantic *Littorina saxatilis*. *Proceedings of the Royal Society B*, **278**, 3175–3183.

Efron B, Tibshirani R (1995) *Cross-validation and the bootstrap: estimating the error rate of a prediction rule*. Stanford University, Stanford, California.

Gómez Cano AR, Hernández Fernández M, Álvarez-Sierra MA (2013) Dietary ecology of Murinae (Muridae, Rodentia): a geometric morphometric approach. *PLoS ONE*, **8**, e79080.

Grahame JW, Wilding CS, Butlin RK (2006) Adaptation to a steep environmental gradient and an associated barrier to gene exchange in *Littorina saxatilis*. *Evolution*, **60**, 268–278.

Hannaford-Ellis CJ (1985) The breeding migration of *Littorina arcana* Hannaford Ellis, 1978 (Prosobranchia: Littorinidae). *Zoological Journal of the Linnean Society*, **84**, 91–96.

Healy B, Oliver G, Hatch P, Good J (1997) *Coastal lagoons in the Republic of Ireland* , Vol. 3, Parts 10 & 13. Report to the National Parks and Wildlife Services, Dublin.

Hollander J, Adams DC, Johannesson K (2006a) Evolution of adaptation through allometric shifts in a marine snail. *Evolution*, **60**, 2490–2497.

Hollander J, Butlin RK (2010) The adaptive value of phenotypic plasticity in two ecotypes of a marine gastropod. *BMC Evolutionary Biology*, **10**, 333.

Hollander J, Collyer ML, Adams DC, Johannesson K (2006b) Phenotypic plasticity in two marine snails: constraints superseding life history. *Journal of Evolutionary Biology*, **19**, 1861–1872.

Hollander J, Lindegarth M, Johannesson K (2005) Local adaptation but not geographical separation promotes assortative mating in a snail. *Animal Behaviour*, **70**, 1209–1219.

Janson K, Sundberg P (1983) Multivariate morphometric analysis of two varieties of *Littorina saxatilis* from the Swedish west coast. *Marine Biology*, **74**, 49–53.

Johannesson B (1986) Shell morphology of *Littorina saxatilis* Olivi: the relative importance of physical factors and predation. *Journal of Experimental Marine Biology and Ecology*, **102**, 183–195.

Johannesson B, Johannesson K (1990) *Littorina neglecta* Bean, a morphological form within the variable species *Littorina saxatilis* (Olivi)? *Hydrobiologia*, **193**, 71–87.
Johannesson B, Johannesson K (1996) Population differences in behaviour and morphology in the snail *Littorina saxatilis*: phenotypic plasticity or genetic differentiation? *Journal of Zoology*, **240**, 475–493.
Johannesson K, Johannesson B, Rolán-Alvarez E (1993) Morphological differentiation and genetic cohesiveness over a microenvironmental gradient in the marine snail *Littorina saxatilis*. *Evolution*, **47**, 1770–1787.
Johannesson K, Panova M, Kemppainen P *et al.* (2010) Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. *Philosophical Transactions of the Royal Society B*, **365**, 1735–1747.
Kendall DG (1977) The diffusion of shape. *Advances in Applied Probability*, **9**, 428–430.
Klingenberg CP (2011) MorphoJ: an integrated software package for geometric morphometrics. *Molecular Ecology Resources*, **11**, 353–357.
Mill PJ, Grahame J (1990) The distribution of *Littorina saxatilis* (Olivi) and *Littorina arcana* Hannaford Ellis in Lyme Bay, southern England. *Journal of Molluscan Studies*, **56**, 133–135.
Mill PJ, Grahame J (1995) Shape variation in the rough periwinkle *Littorina saxatilis* on the west and south coasts of Britain. *Hydrobiologia*, **309**, 61–71.
Mitteroecker P, Gunz P (2009) Advances in geometric morphometrics. *Evolutionary Biology*, **36**, 235–247.
R Core Team (2014) R: a language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*, URL http://www.R–project.org/.
Rees WJ (1965) The aerial dispersal of Mollusca. *Proceedings of the Malacological Society of London*, **36**, 269–282.
Reid DG (1996) *Systematics and evolution of* Littorina. The Ray Society, London.
Rohlf FJ, Marcus LF (1993) A revolution in morphometrics. *Trends in Ecology and Evolution*, **8**, 129–132.
Rohlf FJ, Slice D (1990) Extensions of the Procrustes method for the optimal superimposition of landmarks. *Systematic Zoology*, **39**, 40–59.
Sinfield CA (2008) Dynamics of the *Littorina saxatilis* hybrid zone. Ph.D. Thesis, The University of Leeds.
Small MP, Gosling EM (2000) Species relationships and population structure of *Littorina saxatilis* Olivi and *L. tenebrosa* Montagu in Ireland using single-strand conformational polymorphisms (SSCPs) of cytochrome b fragments. *Molecular Ecology*, **9**, 39–52.
Stone JR (1998) Landmark-based thin-plate spline relative warp analysis of gastropod shells. *Systematic Biology*, **47**, 254–263.
Venables WN, Ripley BD (2002) *Modern applied statistics with S*. Springer, New York.
Walker TN, Grahame JW (2011) Shell shape variation and fitness variables in the gastropod *Littorina saxatilis*. *Marine Ecology Progress Series*, **430**, 103–111.
Westram AM, Galindo J, Rosenblad MA *et al.* (2014) Do the same genes underlie parallel phenotypic divergence in different *Littorina saxatilis* populations? *Molecular Ecology*, **23**, 4603–4616.
Zelditch M, Swiderski D, Sheets H (2012) *Geometric morphometrics for biologists: a primer*. Academic Press, London.

Chapter III

Colonisation history of *Littorina saxatilis* from the British Isles inferred from genomic data

Abstract

Understanding the phylogeography and demographic history of species constitutes a platform to study speciation in the geographic context. Two recent phylogeographic studies have provided new insight into the evolutionary history of the intertidal snail *Littorina saxatilis*, which is increasingly becoming a model species in the study of ecological speciation. Those studies used mtDNA markers to reconstruct the phylogeographic structure of the species across the entire species range (North Atlantic), and found that one of the most complex phylogeographic patterns was found around the British Isles. Here, genome-wide SNP data from species in the *Littorina saxatilis* complex (*L. saxatilis*, *L. arcana*, and *L. compressa*) from various sites around the British Isles were used for characterising the genetic variation at a fine geographic scale. Principal component analysis and measures of F_{ST} across sites suggested certain structure within *L. saxatilis*. However, population structure analysis did not predict a clear number of genetic clusters, and genetic variation was rather better explained by isolation-by-distance. A neighbour-joining tree using Nei's estimator of F_{ST} did not solve the phylogenetic trichotomy among the species. Geographical distribution of the extant genetic lineages found on this study agreed with previous works that have suggested southwest Ireland, the Western English Channel and the Faroe Islands as glacial refugia. Overall, individuals did not cluster by ecotype in the principal components analysis, but rather by geography. Even though no statistical modelling could be performed here due to time limitations, one very important contribution of this work is that it provides future studies with an invaluable source of molecular information to conduct more detailed phylogeographic studies using, for example, an approximate Bayesian computation approach. This work also provides a wealth of sequence data to perform analyses aimed at detecting signatures of natural selection.

Introduction

Phylogeographic studies aim at understanding the relationship between patterns of genetic variation and historical geographical events (Avise 2000). Repeated glaciations during the Last Glacial Maximum (~ 18–25 ka) affected distributions and population sizes of marine taxa in the North Atlantic (Wares & Cunningham 2001; Maggs *et al.* 2008), with genetic evidence showing that some populations persisted in periglacial refugia (Maggs *et al.* 2008; Provan & Bennett 2008). These glacial cycles caused fluctuations in the sea level and changes in habitat availability, leading in many cases to population subdivision, local extinctions and repeated recolonisations from refugia through potentially different routes after the ice retreated (Coyer *et al.* 2003; Provan *et al.* 2005; Colson & Hughes 2007; Hoarau *et al.* 2007).

The intertidal snail *Littorina saxatilis* is broadly distributed across the rocky shores of the North Atlantic (Reid 1996) and yet no comprehensive phylogeographic studies of the species had been conducted until recently (Doellman *et al.* 2011; Panova *et al.* 2011). These two studies used mtDNA from samples that were collected all across the species range and found strong population structure. They also identified putative glacial refugia where *L. saxatilis* could have survived over glacial periods, including southwest Ireland and the Western English Channel, coinciding with suggested refugia for other North Atlantic taxa (Maggs *et al.* 2008). This contribution to the understanding of the glacial history of *L. saxatilis* is especially relevant because the species is increasingly becoming a model system to study ecological speciation (Johannesson 2001, 2003; Butlin *et al.* 2008), and understanding the phylogeography and demographic history of a species can be used as a platform to study speciation in the geographic context (Bolnick & Fitzpatrick 2007). However, no comprehensive examination of genetic variation within the *L. saxatilis* species complex from the British Isles exists.

There is no doubt that traditional Sanger sequencing has greatly progressed the study of phylogeography. However, high-throughput sequencing technologies and bioinformatic tools allow the discovery of enormous numbers of single nucleotide polymorphisms (SNPs) across multiple loci. Given that SNPs are found in high frequency throughout genomes, using them as molecular markers

has led to phylogeographic and population genetic studies of unprecedented resolution (Shaffer & Thomson 2007; Seeb *et al.* 2011; McCormack *et al.* 2013; Robinson *et al.* 2014). Nevertheless, acquiring more data does not necessarily solve certain phylogenetic problems, unless suitable markers that maximise phylogenetic signal are used (Philippe *et al.* 2011). To this end, high-throughput targeted capture offers an advantage in that a large, preselected set of loci can be captured with synthetic probes and sequenced (review by Jones & Good 2015).

This chapter reports a study of genetic variation in the *L. saxatilis* complex from the British Isles at a fine geographical scale using genome-wide SNP data generated by high-throughput targeted capture. The specific aims of this study were: (1) to characterise the genetic variation of *Littorina saxatilis*, (2) to describe the geographical distribution patterns of genetic variation, and (3) to discern the potential causes of these patterns.

Methods

Sampling

Sampling was described in Chapter II. Preservation of foot and/or head tissue in 100% ethanol immediately followed each individual dissection. A total of 198 females were used in this study, which included 182 *L. saxatilis*, 10 *L. arcana*, and 4 *L. compressa*, and 2 *L. fabalis* from Sweden (Table III.1). These last two individuals, kindly provided by Dr Anja Westram (University of Sheffield), were used as outgroup since they belong to the sister group of the *L. saxatilis* complex (Reid 1996; Reid *et al.* 2012).

Table III.1 Collection sites of the individuals of the *Littorina saxatilis* complex used in this study

Site	Site code	Latitude (°)	Longitude (°)	*L. saxatilis*			*L. arcana*	*L. compressa*
				H and *M*	*neglecta*	*tenebrosa*		
Thornwick Bay	THB	54.13267	–0.11503	9				
Cable Bay I	CAB	53.20724	–4.50024	3				
Cable Bay II	CAB	53.20703	–4.50458	5	1			
Broad Haven	BHE	51.60891	–4.91878	6			5	
Oban I	OB1	56.42207	–5.48392	7				
Oban II	OB2	56.41133	–5.48025	2	6			
Isle of Mull I	IM1	56.46981	–5.70344	8				
Isle of Mull II	IM2	56.47068	–5.69599	9				
East Prawle I	PEP	50.20544	–3.71664	6				
East Prawle II	PEP	50.20530	–3.71686	2				
St Margarets at Cliffe	SMC	51.14785	1.38415	9				
Castletown	CAS	58.59832	–3.38172	6				
Aberdeen	ABE	57.14018	–2.05124	12				
Dunbar	DUN	56.00360	–2.51174	3				
St Abbs	STA	55.89968	–2.13004	8				
Amble	AMB	55.33215	–1.56292	6	1		4	1
Old Peak	OLD	54.41036	–0.49196	4	5			
Port Saint Mary I	PS1	54.07602	–4.73618	9				
Port Saint Mary II	PS2	54.07695	–4.73746	6	3			
Lettermullan	LET	53.23096	–9.72898	9				
Loch Aibhnin I	AIB	53.32606	–9.57610			3		
Loch Aibhnin II	AIB	53.31895	–9.57471			3	1	
Loch Tanai	AIB	53.31477	–9.57206			5		
Loch Fhada	FHA	53.31041	–9.61229	6				
Ceann Trá	CEA	52.13205	–10.36071	8				3
Wexford I	WEX	52.35618	–6.42006	2				
Wexford II	WEX	52.35683	–6.41952			3		
Wexford III	WEX	52.35373	–6.37763			3		
Dersingham	DER	52.86750	0.44738			4		

Lab work

DNA isolation

DNA isolation was carried out in the Molecular Ecology Laboratory, Department of Animal and Plant Sciences, University of Sheffield. Genomic DNA was extracted using a modified version of the protocol from Wilding *et al*. (2001). Tissue was put in 500 µL 60 ºC CTAB buffer (2% CTAB, 1.4 M NaCl, 20 mM EDTA, 100 mM Tris-HCl pH 8, 0.2% β-mercaptoethanol) with 2 units of proteinase K, and incubated at 60 ºC for 15 h at an oscillation speed of 110

rpm. After this period, 0.3 mg ribonuclease was added, followed by incubation for 1 h at 60 ºC after vigorous mixing. Subsequently, 500 µL chloroform:isoamyl alcohol (24:1) was added, followed by gentle mixing for 10 min. DNA was isolated with 5 PRIME's Phase Lock Gel™ following the manufacturer's instructions. Then, 500 µL isopropyl alcohol was added to precipitate DNA, and incubated at room temperature for 5 min after gentle mixing. Samples were centrifuged at 4 ºC for 40 min at 13,000 rpm. After discarding the supernatant, DNA was washed with 1 mL 70% ethanol by gently mixing for 5 min, followed by centrifugation for 5 min at 13,000 rpm. Once the supernatant was discarded, the washing step was repeated with 500 µL 70% ethanol. Finally, the air-dried DNA pellet was dissolved in 50 µL 10% TE buffer (10 mM Tris-HCl pH 8, 1 mM EDTA). Concentration was measured by fluorometry.

Probe design and sequencing

As discussed in Chapter I, targeted capture techniques combined with NGS technologies constitute a suitable method for genotyping individuals with prior sequence information (Gnirke *et al.* 2009; Jones & Good 2016). Similarly to other reduced-representation methods, this approach concentrates sequencing resources and efforts on a subset of genomic regions using parallel enrichment of such regions. However, only with high-throughput targeted capture is it possible to preselect those genomic regions using probes that are designed based on prior sequence information. This approach was used for genotyping individuals in this study given the advantages of the method and the availability of prior sequence information for *Littorina saxatilis*.

In total, 3684 120-bp capture probes were used for targeting 3684 contigs of a draft *Littorina saxatilis* reference genome (a single 'crab ecotype' snail from Tjärnö, Strömstad, Sweden, sequenced as part of The IMAGO Marine Genome project, http://cemeb.science.gu.se/research/imago-marine-genome-projects/, coordinated by Profs. Kerstin Johannesson and Anders Blomberg, University of Gothenburg; second assembly version called "littorina_a1234_CLC_Soft_ss_100_400_novo.fasta" completed in October 2012). This reference genome consisted of 717,928 contigs of various lengths (minimum = 200, 1st quartile = 280, median = 422, mean = 659, 3rd quartile = 758; maximum = 23,810; all lengths in bp). GC content of the reference genome

was 41.5%, and the proportion of undetermined bases was 1.4%. The karyotype of this species (2n = 34) has previously been assembled using cytogenetic procedures, showing that males are the heterogametic sex (Rolán-Alvarez *et al.* 1996). However, a recent de novo RADseq study (Ravinet *et al.* 2015) has suggested that sex-linked loci may be rare among these markers.

Contigs of at least 500 bases long and coverage between 15x and 55x were considered for probe design. The large number of contigs in the reference genome might partially be a reflection of highly divergent alleles that are difficult to assemble together as single loci. Thus, only probes with complete (120 bp) or nearly complete (> 115 bp) alignment to at most two locations of the genome were retained in order to increase the probability of capturing divergent alleles at loci that might be represented in two contigs due to the high rate of heterogeneity, while still avoiding duplicated or repetitive regions. The ratio between potential probes with alignments to one and two contigs was 9:1. Considering this proportion, one probe per contig was randomly selected, resulting in 3316 and 368 probes aligning to a single and two locations in the genome, respectively. RapidGenomics (Florida, USA) designed and selected the probes, prepared individually barcoded libraries, captured the fragments ("Target-Seq"), and used an Illumina HiSeq 2000 machine to generate on average 2.7 million single-end 100-base reads per individual sample.

Data analysis

RapidGenomics (Florida, USA) delivered the entire set of raw sequences in FASTQ format, as well as single nucleotide polymorphism (SNP) information encoded in the variant call format (VCF). Ideally, one should use the raw sequences to extract any given information required, which allows analysis flexibility and ensures clarity in every processing step undertaken. However, filtered SNP data provided by RapidGenomics (Florida, USA) were used in this study due to time constraints. Only sites with mapping quality above 10 were kept. Alleles supported by fewer than three reads were discarded. Only alleles with a minor allele frequency of at least 0.01 were kept. Sites that were typed in less than 40% of individuals were removed. RapidGenomics (Florida, USA) performed all the processing steps up to this stage. Sites that were typed in fewer than two individuals in each collecting site were discarded. A total of

26,296 biallelic SNPs distributed over 3334 contigs were retained, that is an average of 8 SNPs per contig.

Exploratory analysis

The R package *adegenet* (version 1.4-2, Jombart & Ahmed 2011) was used for handling and analysing the SNP data. Principal component analysis (PCA) was used for summarising the genetic diversity among the sampled individuals without assuming an evolutionary model. PCA was also used for visual identification of potential outlier individuals and clusters of genetically similar individuals. Missing information was replaced by mean values. Collecting sites were grouped into five geographical regions (NE Britain, W Britain, S Britain, Isle of Man, and Ireland) according to the colonisation history suggested by Doellman *et al.* (2011). Nei's estimator of pairwise F_{ST} (Nei 1972, 1978) was computed across sites and species, also in the R package *adegenet*. F_{ST} values were used for estimation of a neighbour-joining tree (Saitou & Nei 1987), as implemented in the R package *ape* (version 3.3, Paradis *et al.* 2004).

Population structure

A set of 182 *L. saxatilis* individuals from all 23 sites was used in this analysis. Only one SNP per locus was kept since the population structure analysis required unlinked markers. For this, the SNP that was typed in the largest number of individuals for each locus was retained, and in cases where this number was shared by more than one SNP, then one SNP per locus was kept at random. Within each site, loci were tested for Hardy-Weinberg equilibrium (HWE) using the R package *pegas* (version 0.8-2, Paradis 2010). Loci departing from HWE (p-value < 0.01) at any collection site were removed from the entire dataset. The final dataset was made up of 2232 loci.

Population structure was inferred using the software *structure* (version 2.3.4, Pritchard *et al.* 2000), which implements a Bayesian, model-based clustering method. The analysis was run using the admixture model, assuming no prior information on collection site and considering correlated allele frequencies (Falush *et al.* 2003). A total of 10 independent replicates were run for values of genetic clusters (*K*) ranging from 1 to 23, i.e. maximum number of sites where samples were collected. A total of 100,000 Markov chain Monte Carlo (MCMC)

repetitions were simulated for each independent replicate and *K* value, discarding the first 20,000 repetitions as burnin. A new random seed was used for each run. The estimated natural logarithm of the posterior probability of the data (L(*K*)) for each replicate and *K* was plotted in order to determine the predicted number of *K*. This plot was then compared with the delta-K (Evanno *et al.* 2005), as computed in the software *structure harvester* (version 0.6.94, Earl & VonHoldt 2012), in order to determine the optimal *K* value.

Optimal alignment of all 10 replicate *structure* runs was found using the algorithm *LargeKGreedy* with 100 random input orders tested, as implemented in the software *clumpp* (version 1.1.2, Jakobsson & Rosenberg 2007). This algorithm was selected following the recommendations of the authors based on the *K* value, number of replicate runs, number of individuals, and number of input orders to be tested. The mean of aligned individual *Q*-matrices across replicates generated in *clumpp* was visualised using *distruct* (version 1.1, Rosenberg 2004).

Isolation by distance

The same genotype data that was used for the population structure analysis (i.e. 2232 loci and 182 *L. saxatilis* from 23 sites) was used for an isolation-by-distance (IBD) analysis. The R package *adegenet* was used for computing the genetic distance between sites using the Nei's estimator of pairwise F_{ST} (Nei 1972, 1978). Geographic distance between sites was measured around the coast in steps of 10 km. For pairs of sites on different islands, geographic distance was measured around the coast up to the points where the two islands were closest, adding the Euclidian distance between such two points to the two coastal distances. After having measured all geographic distances in clockwise and anticlockwise directions, the shortest distance between any two given sites was kept. Association test between geographic and genetic distances was performed using Mantel's test with 1e5 permutations in *adegenet*. Genetic distance, computed as $F_{ST}/(1–F_{ST})$, was regressed on geographic distance (Rousset 1997).

Genotype and shell shape

The relationship between genotype and shell shape was explored using genetic and phenotypic distance. The genotype data set used for population structure and IBD analyses were used here, excluding 18 *L. saxatilis* for which no morphometric data was available due to damaged shells or low-resolution images. Genetic distance between the remaining 164 *L. saxatilis* was computed using the Nei's estimator of pairwise F_{ST} (Nei 1972, 1978) in *adegenet*. Individual landmark configurations resulting from the Procrustes superimposition were used to compute the Procrustes distance (i.e. Euclidean distance) between individual shapes, which was considered as the phenotypic distance in this analysis. Association test between genetic and phenotypic distances was performed using Mantel's test with 1e5 permutations in *adegenet*. Phenotypic distance was regressed on genetic distance.

Results

Exploratory analysis

Principal component analysis revealed structure within the data with a few potential outlier individuals. The first six principal components explained 18.5% of the total variation in the data and were selected for analysis after visual inspection of the scree plot (Fig. III.1). As expected, PC1 (7.27%) clearly distinguished *L. fabalis* – sister species of the *L. saxatilis* complex – from species in the *L. saxatilis* complex (Fig. III.2a). PC2 (3.05%) revealed some structure among species in the *L. saxatilis* complex, where most *L. arcana* and two *L. compressa* individuals differed from the main *L. saxatilis* group (Fig. III.2a). Furthermore, PC2 made a clear distinction between *L. arcana* from NE Britain and from W Britain (Fig. III.2b). PC1 and PC2 also showed some geographical structure within the main *L. saxatilis* cluster, with individuals from NE Britain and W Britain forming different adjacent groups (Fig. III.2b).

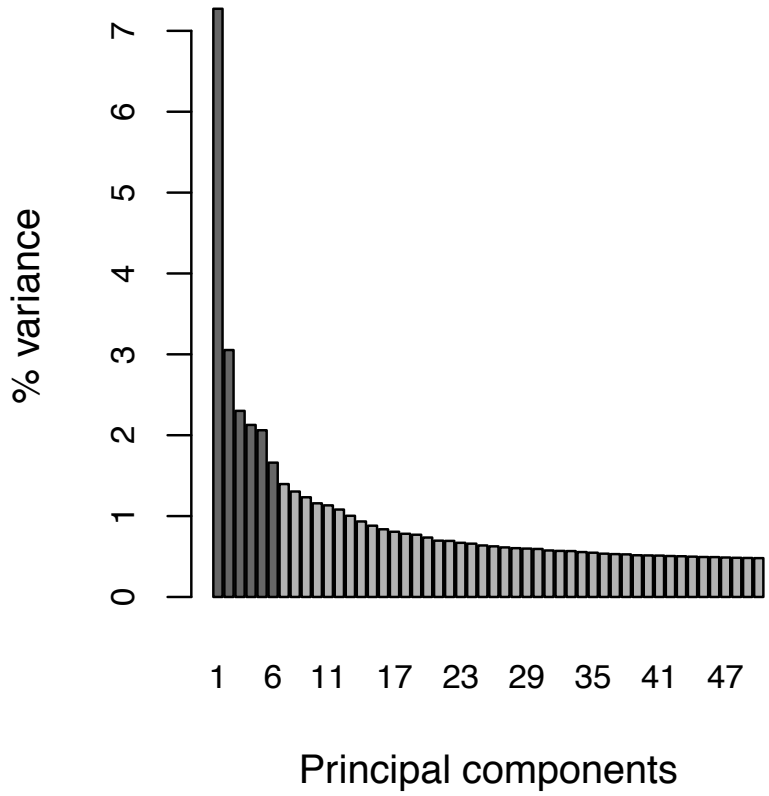


Fig. III.1 Scree plot of the first 50 principal components explaining a proportion of the variance among 26,296 SNPs distributed over 3334 contigs in 198 *Littorina* snails. Dark grey bars together explained 18.5% of the total variation.

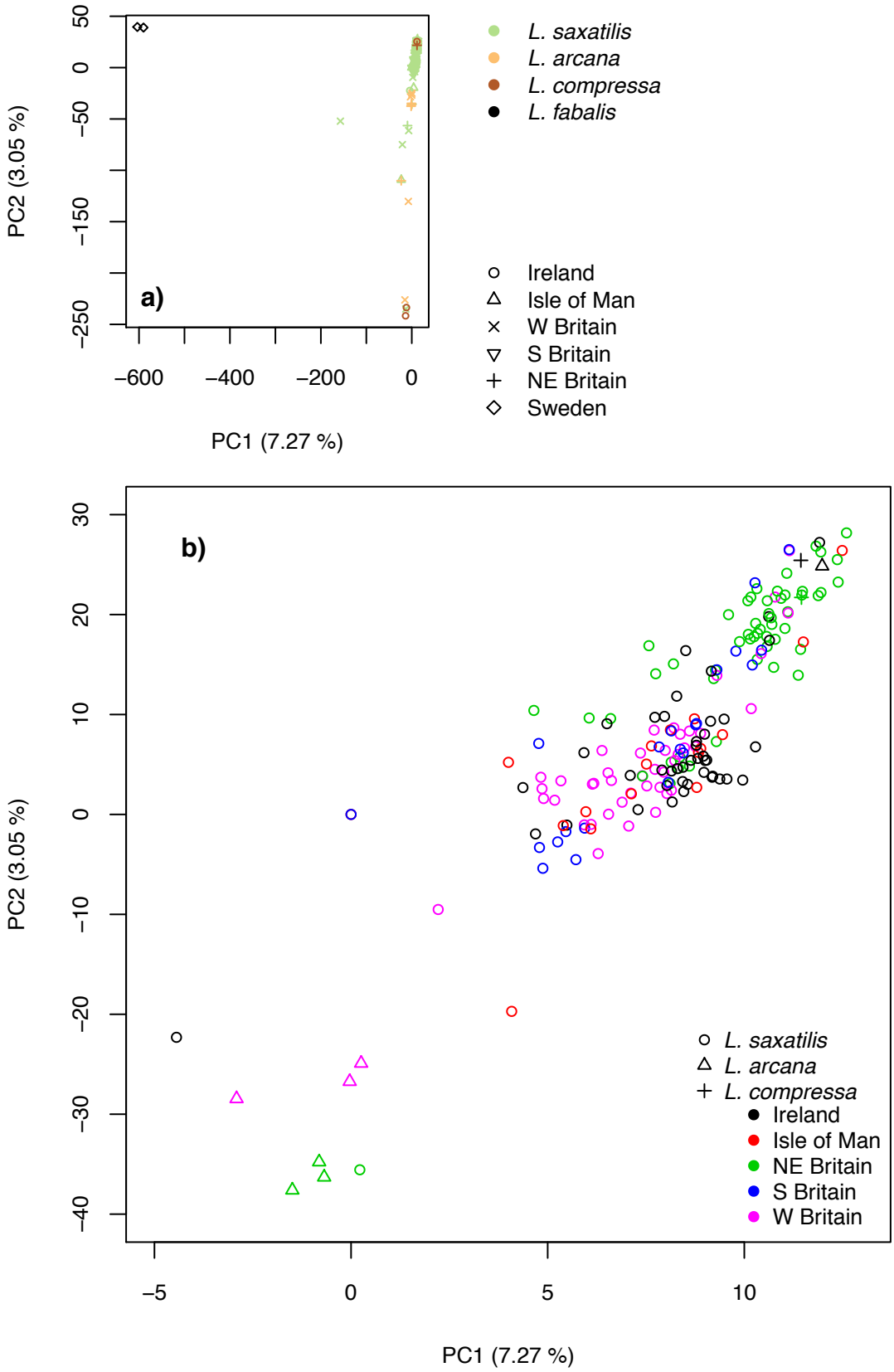


Fig. III.2 Principal component analysis of 26,296 SNPs from 198 *Littorina* snails. (a) Relationship between PC1 and PC2. (b) Close-up of (a). Note the different colour and symbol coding for figures (a) and (b).

PC3 (2.30%) and PC4 (2.13%) distinguished some *L. arcana* from NE and W Britain, as well as two *L. compressa* individuals from Ireland (Fig. III.3a). PC3 and PC4 also revealed a relatively clear geographical structure within *L. saxatilis* (Fig. III.3b). Within the latter, individuals from NE Britain and W Britain formed two distinct groups (Fig. III.3b), with all three S Britain sites being represented in both of them. Individuals from the S Britain sites St Margarets at Cliffe (H and M) and Dersingham (*tenebrosa*) were around the meeting point of the two main groups NE and W Britain. One individual from St Margarets at Cliffe and all individuals from the S Britain site East Prawle were all – but one – placed furthest from the NE group, being more similar to individuals from the W Britain sites Broad Haven, Cable Bay and Isle of Mull.

PC3 and PC4 (Fig. III.3b) showed that all *L. saxatilis* individuals from Castletown (far north of Scotland), originally assigned to the NE Britain group, appeared well within the W Britain group, as did one individual from the nearby site Aberdeen (Fig. III.3b; green within pink). However, most individuals from Aberdeen were within the NE Britain group and around the meeting point with the W Britain group. This was also the case for individuals from Dunbar and St Abbs, although they were further from the W Britain group than Aberdeen individuals. However, those individuals in the NE Britain group that were furthest from the W Britain group were from Amble, Thornwick Bay and Old Peak.

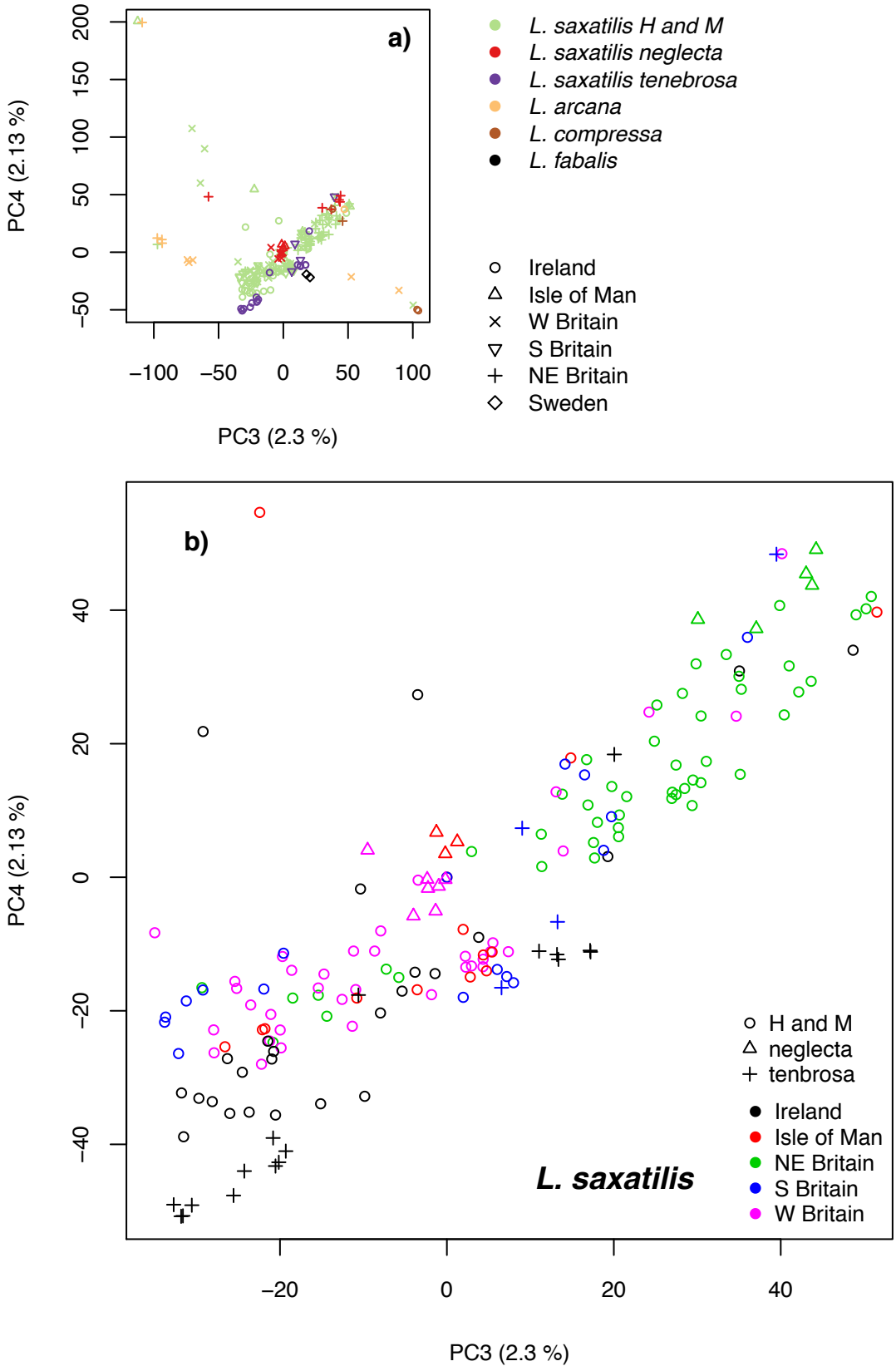


Fig. III.3 Principal component analysis of 26,296 SNPs from 198 *Littorina* snails. (a) Relationship between PC3 and PC4. (b) Close-up of (a), retaining only the 182 *L. saxatilis* individuals. Note the different colour and symbol coding for figures (a) and (b).

PC3 and PC4 (Fig. III.3b) showed that only four *L. saxatilis* individuals in total from the W Britain sites Broad Haven, Cable Bay, and Isle of Mull were well within the NE Britain group. All individuals – except two – from the Isle of Man overlapped with the W Britain group. *L. saxatilis* individuals from Ireland were also in both W Britain and NE Britain groups; however, they overlapped mostly with the W Britain group. Geographical structure was also found among *neglecta* individuals, with those from Amble and Old Peak clustering together far from the W Britain group, whereas those *neglecta* individuals from Cable Bay and the Isle of Man formed a tight group between the NE and W Britain groups.

PC4 clearly distinguished a subset of *L. saxatilis* from E Ireland (Fig. III.3b; black bottom-left), which in turn revealed clear substructure based on ecotype (i.e. H and M vs *tenebrosa*). Even further substructure was found among the *tenebrosa* individuals from the three nearby lochs in W Ireland, with individuals from Loch Aibhnin I and Loch Tanai (< 2 km apart from each other) not overlapping, and individuals of Loch Aibhnin II (between Aibhnin I and Tanai) overlapping with the two aforementioned groups. All but one of the *tenebrosa* individuals from Wexford clustered together. Of all four *tenebrosa* individuals from Dersingham, only three clustered together along PC3.

PC5 and PC6 (Fig. III.4a) revealed certain geographical structure within *L. arcana*, with one group including individuals from the NE Britain sites Thornwick Bay and Amble, and another group with individuals from the W Britain site Broad Haven. *L. arcana* individuals from W Britain were closer to the *L. saxatilis* group than to *L. arcana* individuals from NE Britain (Fig. III.4a). One *L. arcana* individual from Amble was standing alone on the opposite side with respect to all other individuals from the same species and site. Two *L. arcana* individuals from Broad Haven (S Britain) and one from Loch Aibhnin II (Ireland) were well within the main *L. saxatilis* group as defined by PC5 and PC6 (Fig. III.4a).

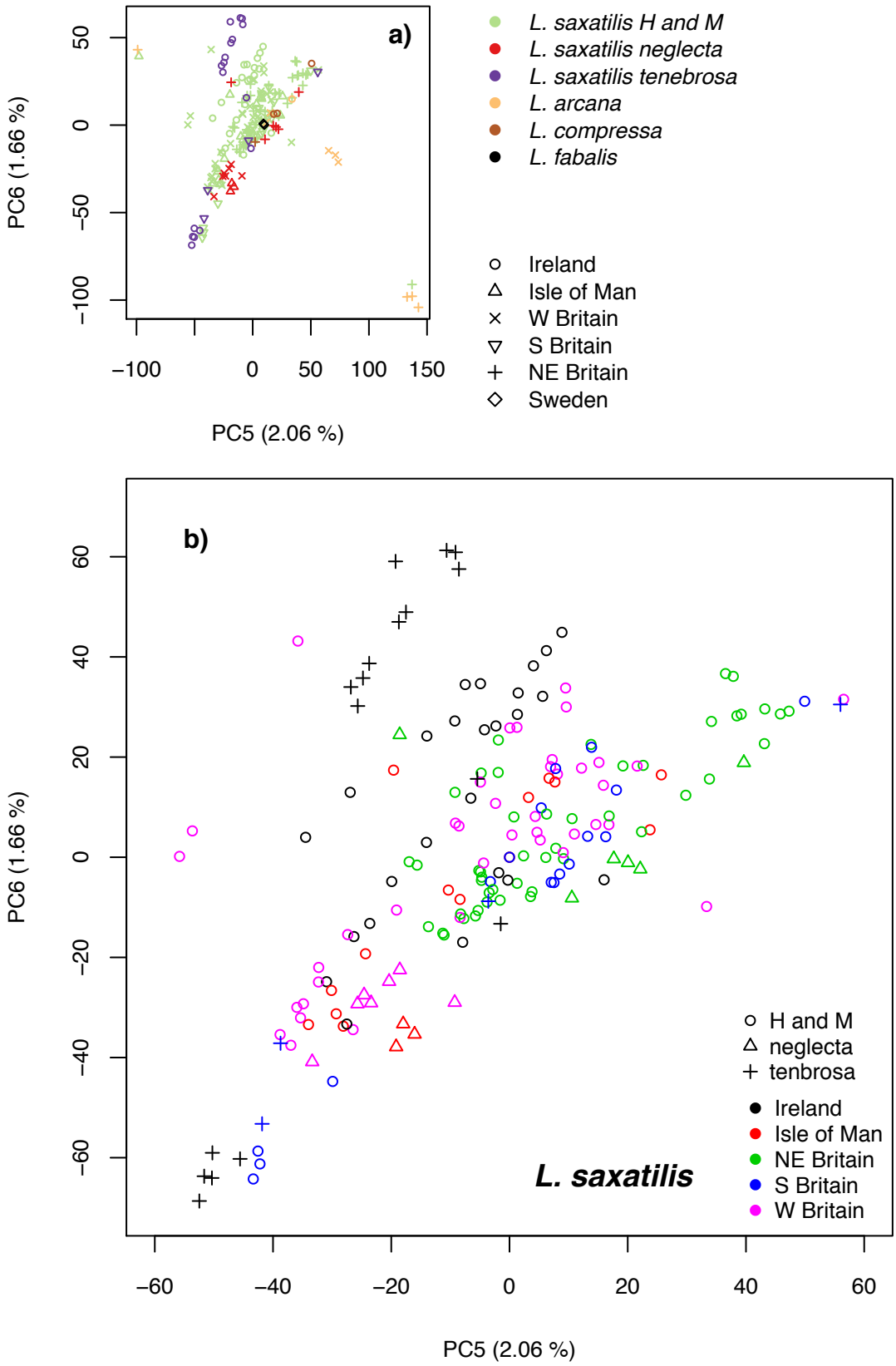


Fig. III.4 Principal component analysis of 26,296 SNPs from 198 *Littorina* snails. (a) Relationship between PC5 and PC6. (b) Close-up of (a), retaining only the 182 *L. saxatilis* individuals. Note the different colour and symbol coding for figures (a) and (b).

PC5 and PC6 showed spatial structure within *L. saxatilis* from Ireland (Fig. III.4b). Here, *tenebrosa* individuals from Wexford differed from H and M individuals that were collected at the same site, but on opposites sides of an artificial site barrier. On the other extreme of the plot were samples from the W Ireland lochs Aibhnin and Tanai, and Lettermullan. Individuals from the SW Ireland site Ceann Trá, which is geographically between the W Ireland sites and Wexford, appeared as intermediates between the two just mentioned groups.

In general, F_{ST} values were higher between species than within species (Fig. III.5). The highest F_{ST} values were those between *L. fabalis* and species in the *L. saxatilis* complex (0.15 ± 0.01). The mean F_{ST} value within the *L. saxatilis* complex was much lower (0.03 ± 0.01). Some geographical structure was observed in the neighbour-joining tree within the *L. saxatilis* complex (Fig. III.6). Here, *L. arcana* from W Britain and from NE Britain, along with *L. compressa* from Ireland, formed a distinct group. Whereas *L. arcana* from Ireland and *L. compressa* from NE Britain were in a group formed by all *L. saxatilis* from NE Britain, except *L. saxatilis* from Castletown, which was the north-eastern-most site. *L. saxatilis* from Castletown grouped with *L. saxatilis* from the W Ireland sites Lettermullan and lochs Tanai, Aibhnin and Fhada. *L. saxatilis* from the two S Britain sites St Margarets at Cliffe and Dersingham formed a distinct group along with *L. saxatilis* from the SW Ireland site Wexford. The other S Britain site East Prawle grouped with the two SW Britain sites Broad Haven and Cable Bay. *L. saxatilis* from the NW Britain sites Oban I, Oban II and Isle of Mull I formed a distinct group. *L. saxatilis* from Ceann Trá (SW Ireland) and from Port St May I (Isle of Man) grouped together. *L. saxatilis* from Isle of Mull II (W Britain) and from Port St May II (Isle of Man) did not show a clear aggregation pattern.

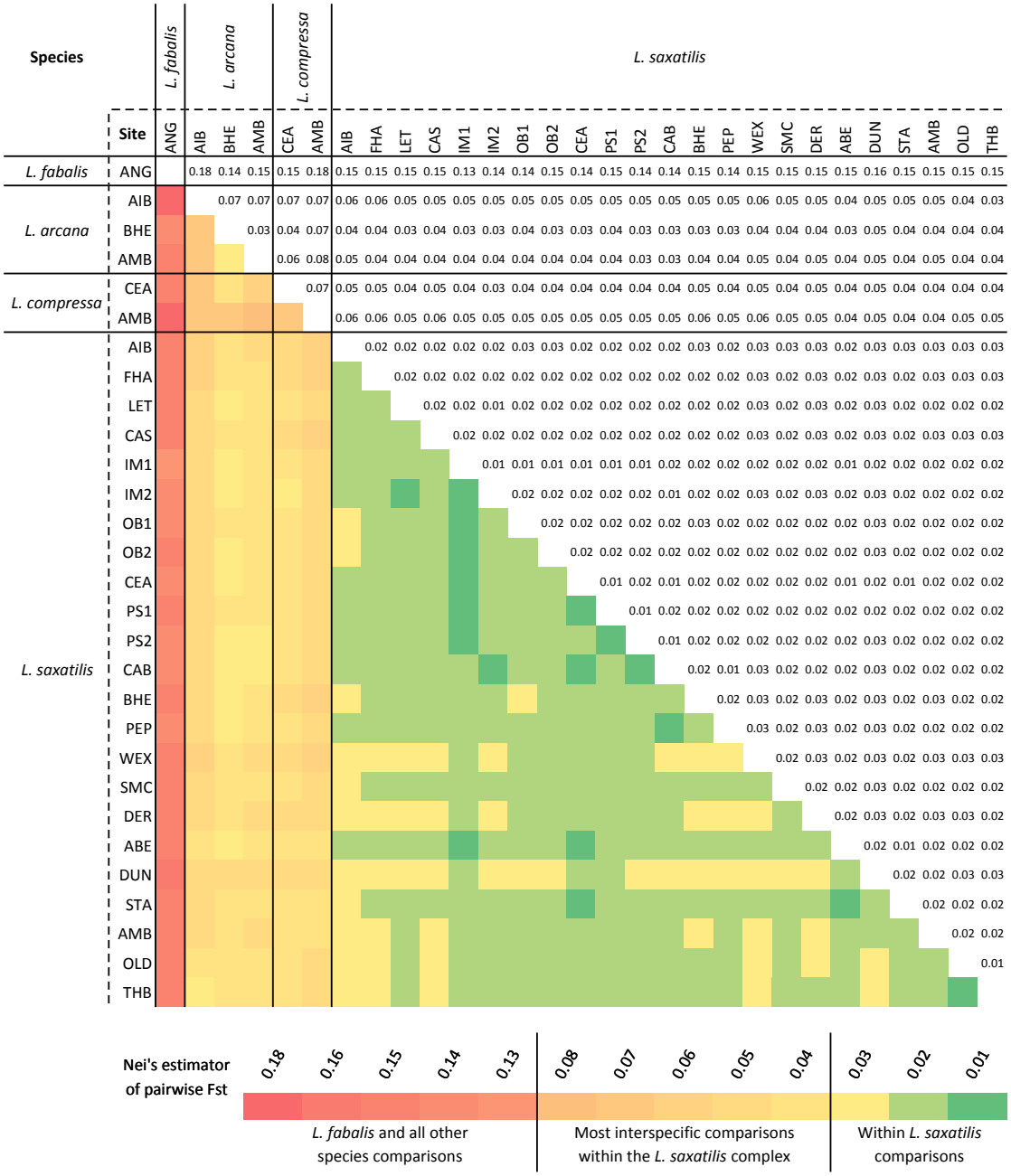


Fig. III.5 Nei's estimator of pairwise F_{ST} between pairs of sites where species in the *L. saxatilis* complex and *L. fabalis* were collected, above the diagonal; and conditional colouring of pairwise comparisons according to F_{ST} values, below the diagonal. Estimation based on 26,296 SNPs from 198 individuals. Site names as in Table III.1.

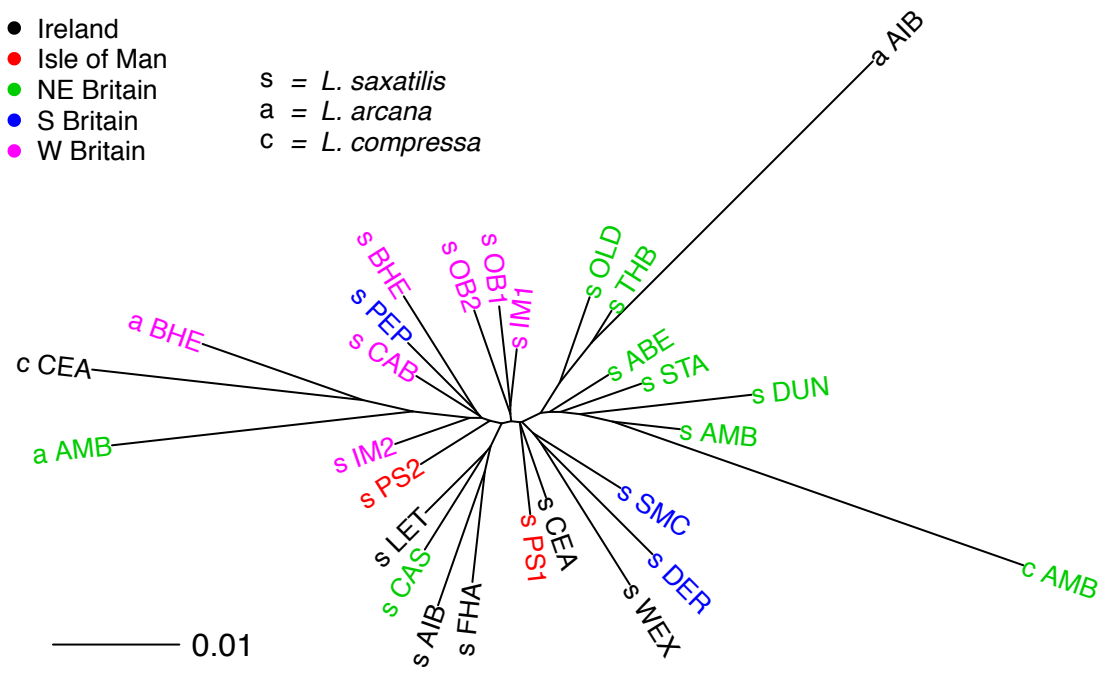


Fig. III.6 Unrooted neighbour-joining tree of species in the *Littorina saxatilis* complex based on pairwise F_{ST} across sites and species. Computation based on 26,296 SNPs from 196 individuals (*L. fabalis* from Sweden was excluded to facilitate visualisation). Sites names are indicated in Table III.1.

Population structure

The estimated natural logarithm of the posterior probability of the data L(*K*) showed a steady increase with *K* (Fig. III.7a; only L(*K*) > –2e–5 were plotted for visualisation purpose). The maximum mean of L(*K*) was at *K* = 12. The variance of L(*K*) markedly increased after *K* = 10, meaning that clustering became less repeatable after this *K* value; however, the delta-K graph indicated *K* = 12 (Fig. III.7b), but the signal was rather weak.

The graphs of the individual and population *Q*-matrices (Fig. III.8a and Fig. III.8b, respectively) revealed rather complex patterns, with no single genetic cluster found at only one site. In other words, at least two genetic clusters were found at every site. Within sites, diversity of genetic clusters was found within single individuals and between individuals. For example, at sites such as Fhada (Fig. III.8a FHA) most individuals had relatively similar proportions of membership to clusters 1 and 2. On the other hand, at sites such as St Margarets at Cliffe (Fig. III.8a SMC) most individuals were assigned to either cluster 9 or 10. For each site, average proportions of membership to clusters were better appreciated on the population *Q*-matrix graph (Fig. III.8b), with most clusters being restricted to specific areas (Fig. III.8c).

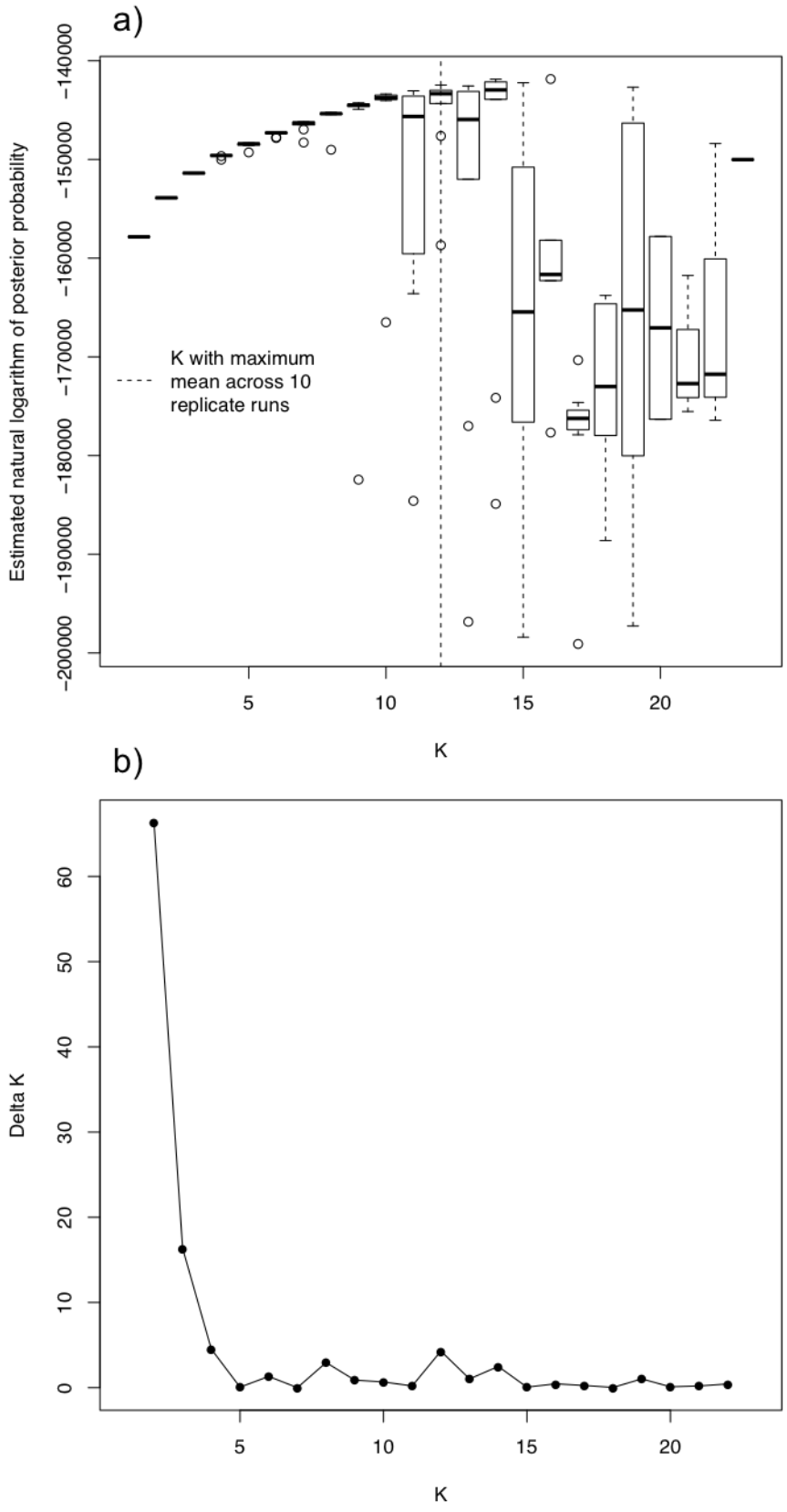
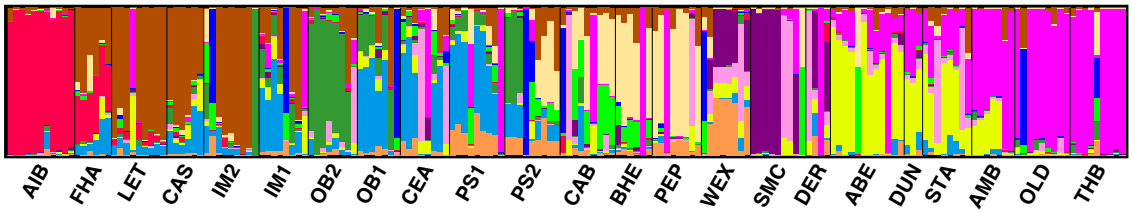
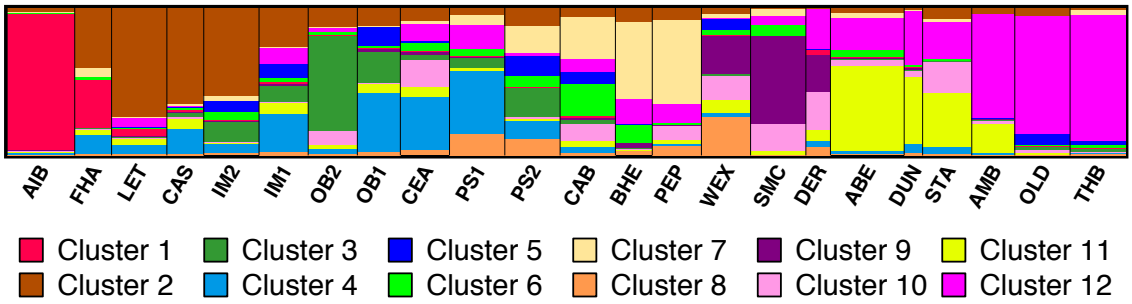


Fig. III.7 Predicted number of genetic clusters (*K*) according to a) the estimated natural logarithm of posterior probability and to b) delta-K. Results from *structure* and *structure harvester* for 182 *L. saxatilis* individuals from across 23 sites in the British Isles.

a)


AIB
FHA
LET
CAS
IM2
IM1
OB2
OB1
CEA
PS1
PS2
CAB
BHE
PEP
WEX
SMC
DER
ABE
DUN
STA
AMB
OLD
THB

b)


AIB
FHA
LET
CAS
IM2
IM1
OB2
OB1
CEA
PS1
PS2
CAB
BHE
PEP
WEX
SMC
DER
ABE
DUN
STA
AMB
OLD
THB
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
Cluster 11
Cluster 12

c)

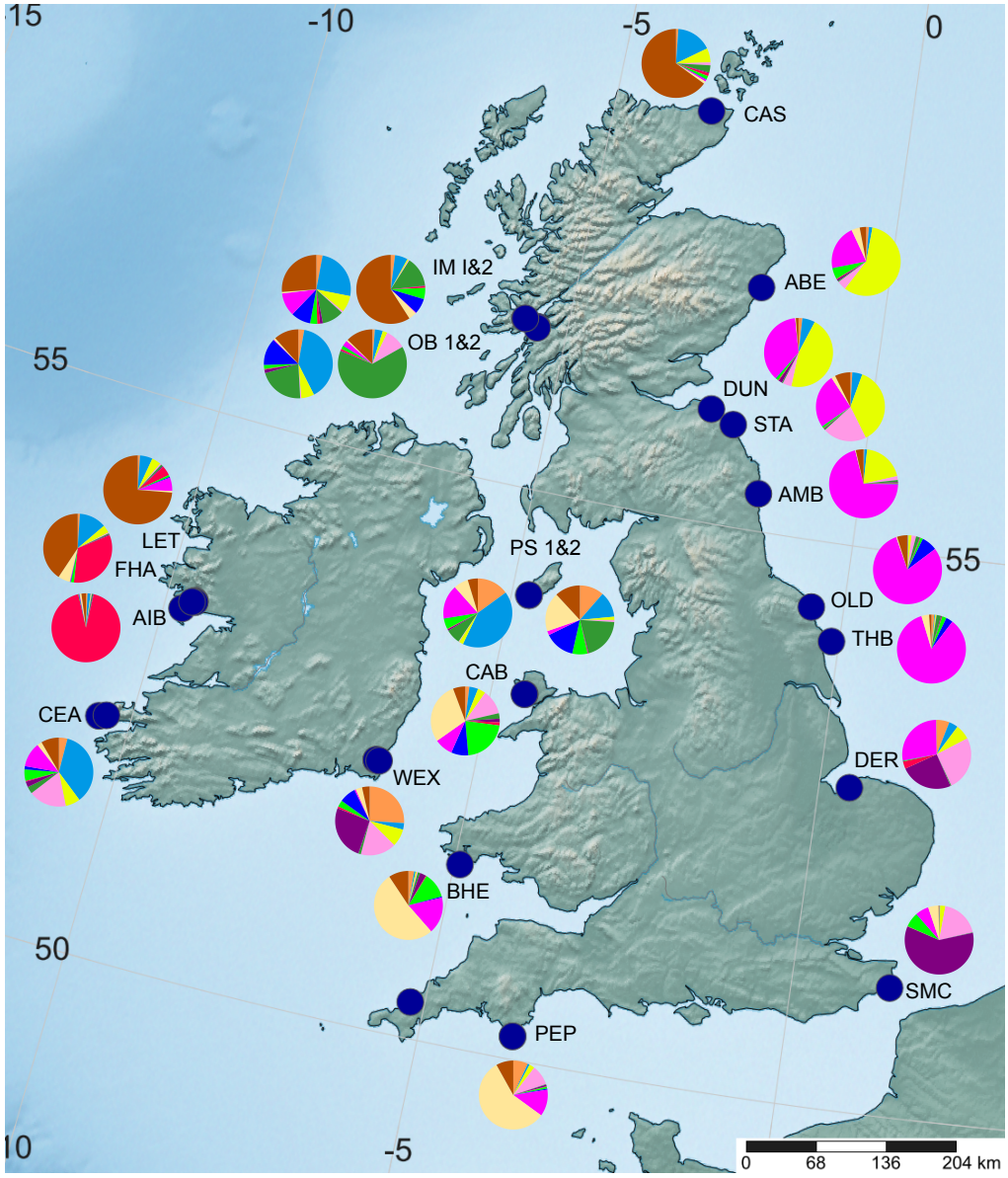

15
-10
-5
0
55
50
10
-5
CAS
IM I&2
OB 1&2
ABE
DUN
STA
AMB
PS 1&2
LET
FHA
AIB
CEA
CAB
WEX
BHE
PEP
OLD
THB
DER
SMC
0
68
136
204 km

Fig. III.8 Population structure analysis as performed in *structure* for 182 *L. saxatilis* individuals from 23 sites around the British Isles. a) Membership coefficients from the individual *Q*-matrix, where each individual is represented by a single vertical line partitioned into 12 coloured segments, with lengths proportional to each of the 12 inferred clusters. b) Membership coefficients from the population *Q*-matrix, where all individuals from each site are represented by a vertical line partitioned into 12 coloured segments, with lengths proportional to the mean of the 12 inferred clusters across all individuals from each site. In a) and b) black lines separate individuals from different sites, and labels correspond to the sites where samples were collected, mapped in c).

Spatial distribution of genetic clusters seemed to be associated with geography (Fig. III.8c). Cluster 1 was found in W Ireland, making up the largest proportion of membership at loch Aibhnin (AIB), where all individuals were ecotype *tenebrosa*. Cluster 1 also made up a large proportion of membership at Fhada (FHA), which is located at the connecting point between loch Aibhnin and the sea. Cluster 2 was mainly found in W Ireland and N Britain, making up a large proportion of membership at Fhada (FHA), Lettermullan (LET), Isle of Mull (IM2), and Castletown (CAS). Cluster 3 was mainly found in W Scotland and Isle of Man, making up the largest proportion of membership at Oban (OB2). Cluster 4 was mainly found in SW Ireland, W and N Scotland and Isle of Man, making up the largest proportion of membership at Ceann Trá (CEA), Oban (OB1) and Port St Mary (PS1). Cluster 5 was found in W Scotland, Isle of Man, N Wales and SE Ireland; however, it did not make up the largest proportion of membership of any site. Cluster 6 was found in SW Ireland, W and E Scotland, Isle of Man, Wales, and SE England, with its largest proportion of membership found at Cable Bay (CAB); however, this proportion was not the largest here.

Cluster 7 was mainly found in SW England, Wales, and Isle of Man, making up the largest proportion of membership at East Prawle (PEP) and Broad Haven (BHE), proportion that decreased northward. Cluster 8 was mainly found in SE Ireland and Isle of Man, with its largest proportion of membership found at Wexford (WEX); however, this proportion was not the largest of this site. Cluster 9 was found in SE Ireland and SE England, making up the largest proportion of membership at St Margarets at Cliffe (SMC). Cluster 10 had a broad distribution, having been found in SW and SE Ireland, W and E Scotland, SW and SE Britain, and N Wales; however, it did not make up the largest proportion of membership at any site. Cluster 11 was mainly found in NE Britain, making up the largest proportion of membership at Aberdeen (ABE), Dunbar (DUN), and St Abbs (STA). Cluster 12 was mainly found in E Britain, making up the largest proportion of membership at Thornwick Bay (THB), Old Peak (OLD), and Amble (AMB).

Isolation by distance

The observed correlation value of 0.24 between the geographic and genetic distance matrices was significantly greater (p-value < 0.001) than those between the permuted matrices (Fig. III.9). Randomly permuted matrices simulate scenarios with no spatial structure; therefore, the result of Mantel's test indicated a clear pattern of isolation by distance. Overall, genetic distance significantly (p-value < 0.001) increased with geographic distance (Fig. III.10). When fitting linear models to site pairs on the same island and on different islands separately, the increase of genetic distance with geographic distance remained significant only for between-islands comparisons.

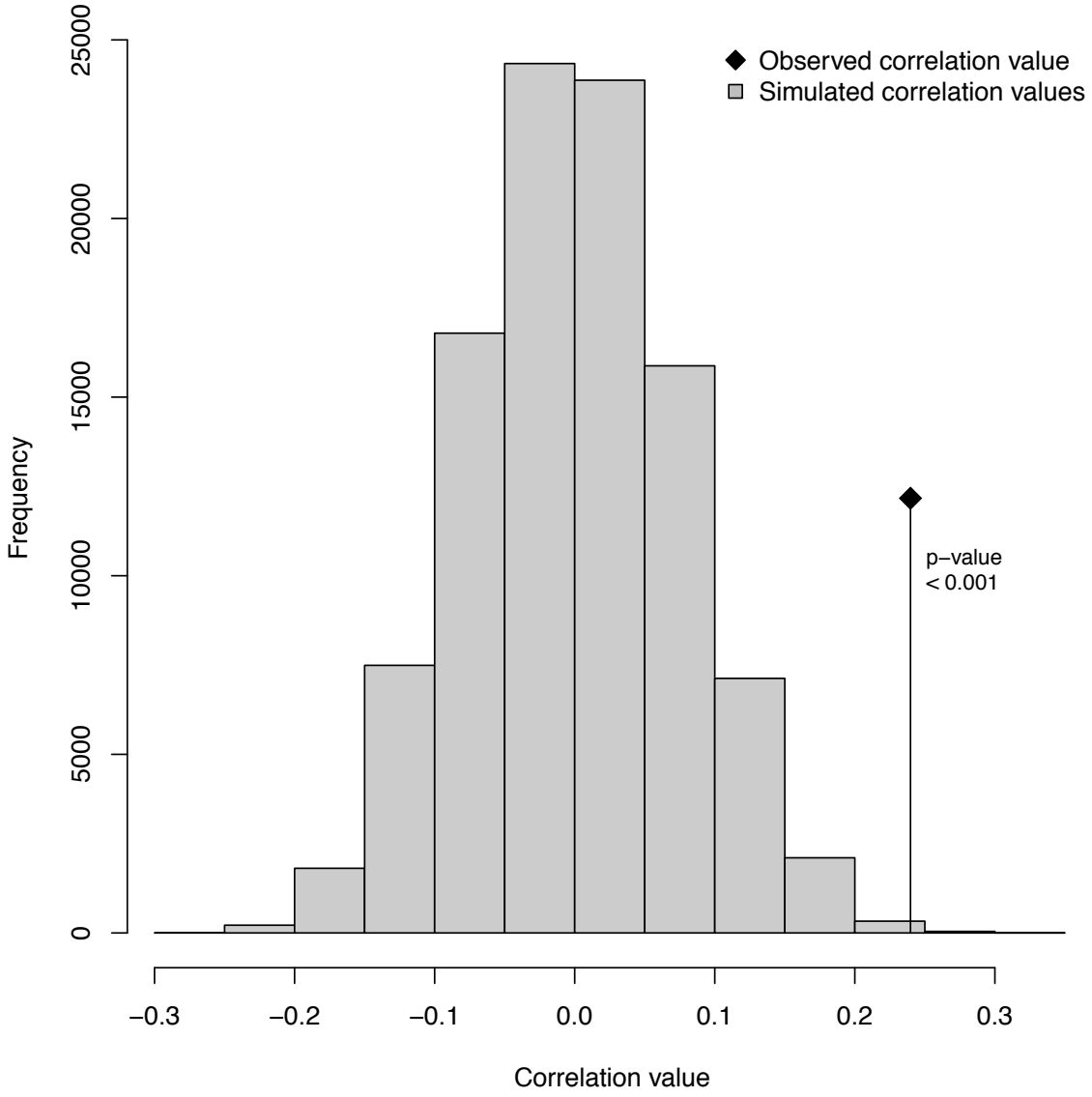


Fig. III.9 Mantel's test of isolation by distance on 182 *L. saxatilis* from 23 sites in the British Isles. Distribution of 1e5 correlation values of permuted geographic and genetic distance matrices (grey) compared to the observed correlation value (black).

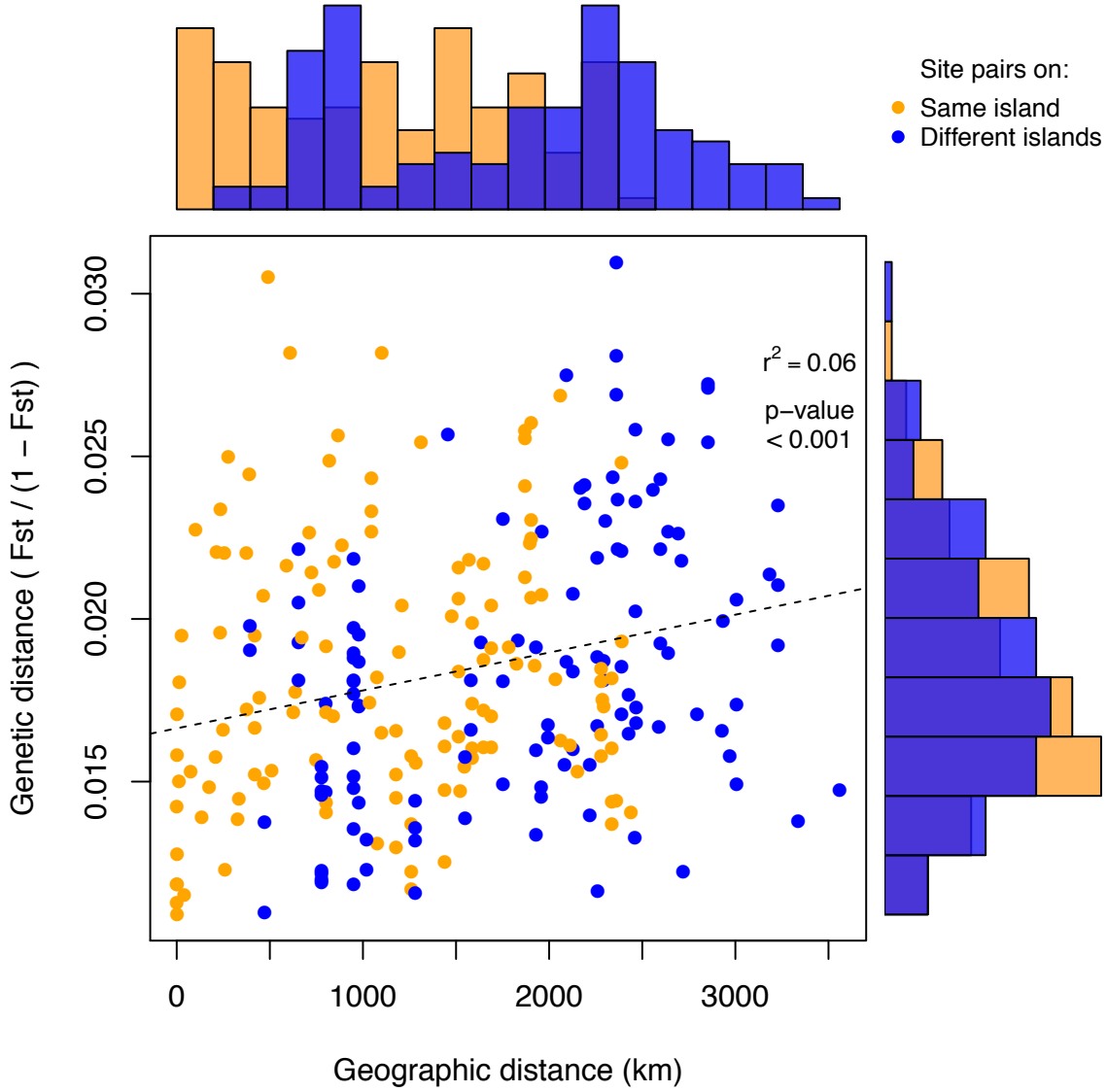


Fig. III.10 Relationship between geographic and genetic distance between all 23 sites from the British Isles, where 182 *L. saxatilis* were collected. Each dot represents a unique pair of sites coloured according to whether sites were on same or different islands. Linear regression of genetic distance on geographic distance is represented as a dashed line. Top histogram shows the distribution of geographic distance, whereas right-hand histogram shows the distribution of genetic distance.

Genotype and shell shape

The observed correlation value of 0.05 between the genetic and phenotypic distance matrices was not significantly greater (p-value = 0.1) than those between the permuted matrices (Fig. III.11), meaning that there was no significant relationship between genotype and shell shape within this dataset. However, phenotypic distance increased with genetic distance (Fig. III.12), although the fit was poor (r^2 = 0.003).

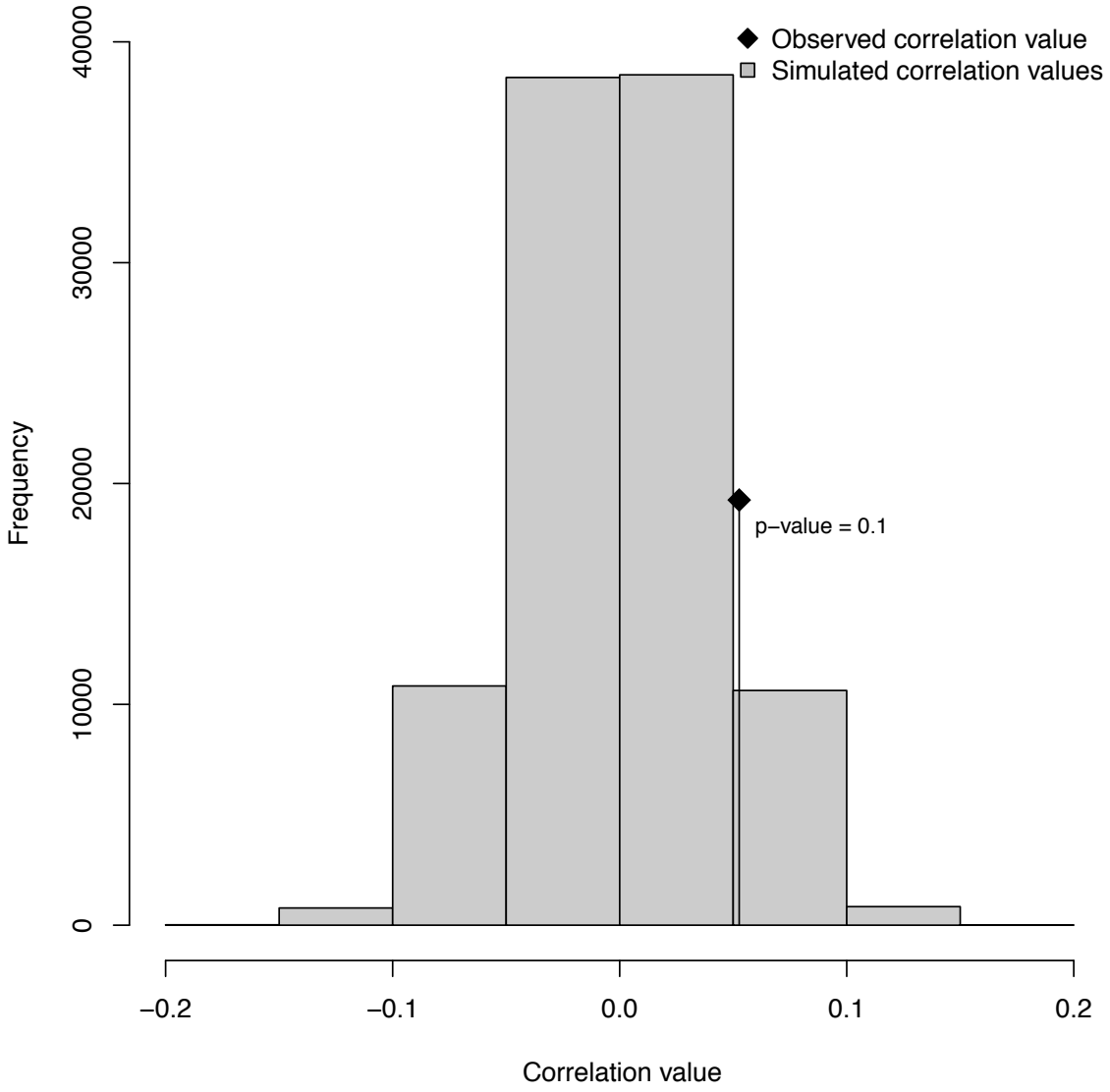
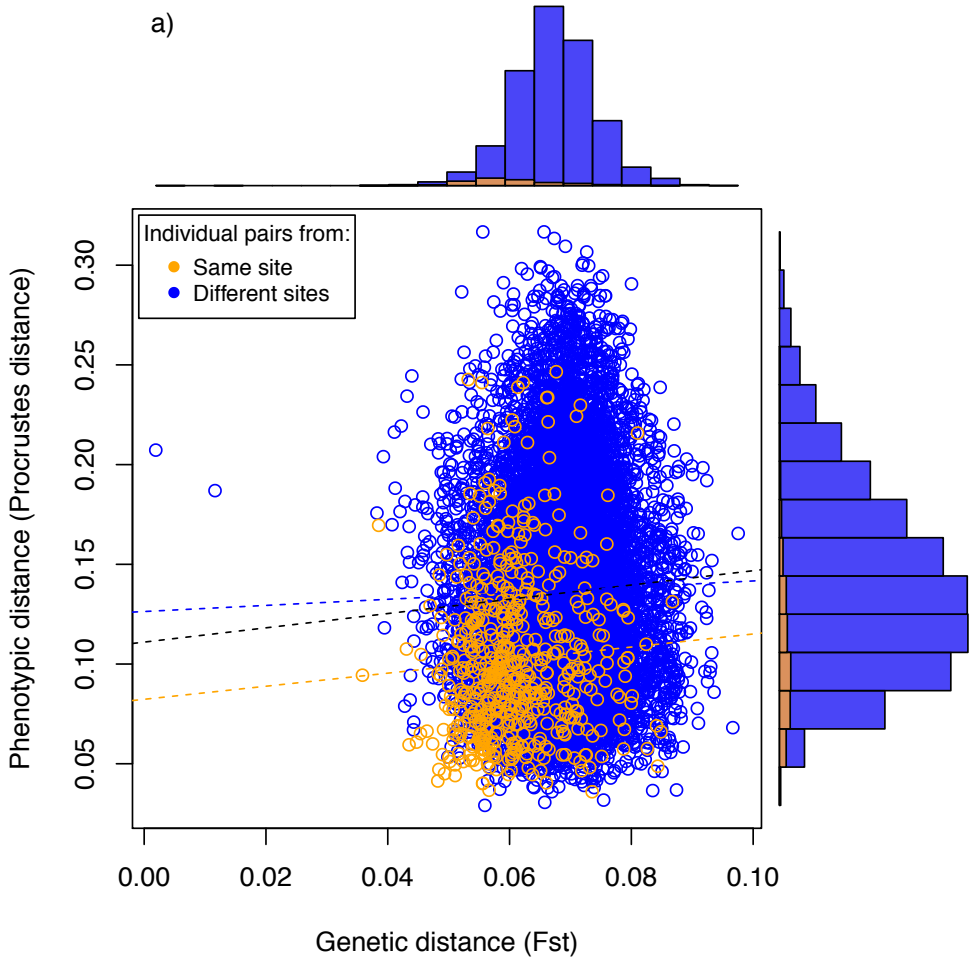


Fig. III.11 Mantel's test of association between genetic and phenotypic distance on 164 *L. saxatilis* from 23 sites in the British Isles. Distribution of 1e5 correlation values of permuted genetic and phenotypic distance matrices (grey) compared to the observed correlation value (black).


a)
Individual pairs from:
Same site
Different sites
Phenotypic distance (Procrustes distance)
0.05
0.10
0.15
0.20
0.25
0.30
0.00
0.02
0.04
0.06
0.08
0.10
Genetic distance (Fst)

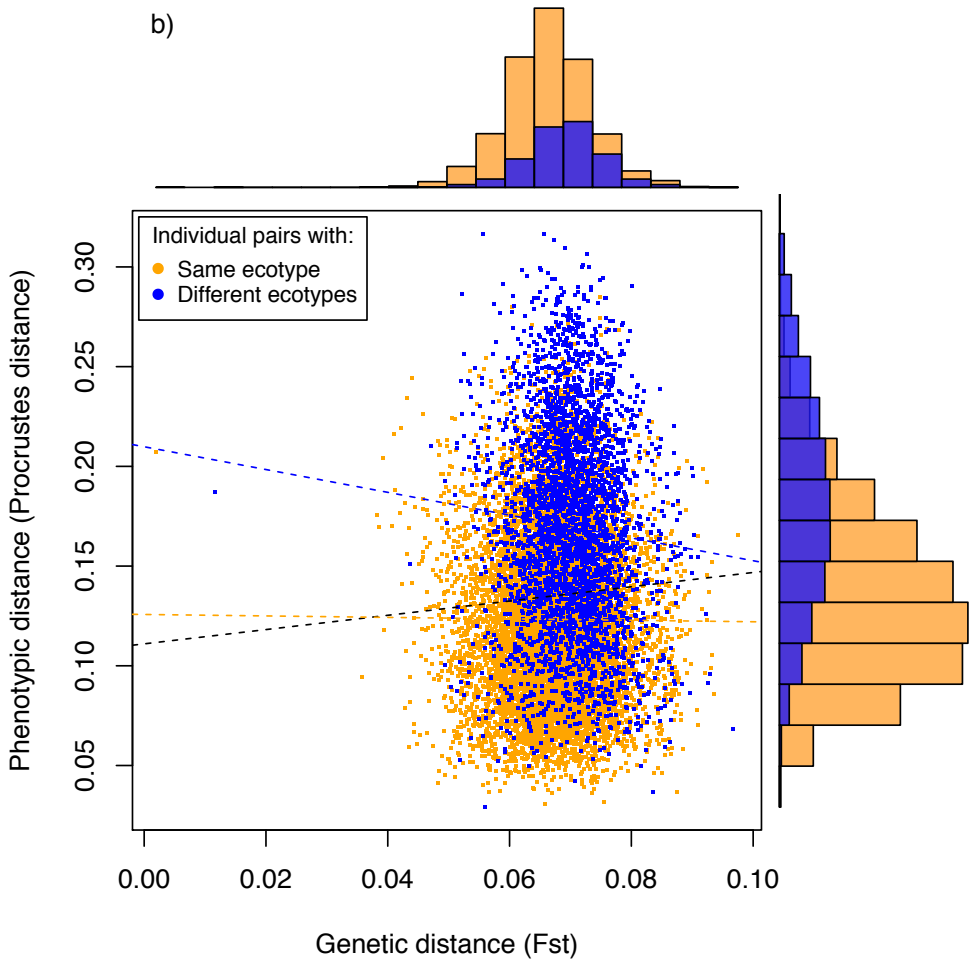

b)
Individual pairs with:
Same ecotype
Different ecotypes
Phenotypic distance (Procrustes distance)
0.05
0.10
0.15
0.20
0.25
0.30
0.00
0.02
0.04
0.06
0.08
0.10
Genetic distance (Fst)

Fig. III.12 Relationship between genetic and phenotypic distance between 164 *L. saxatilis* from 23 sites in the British Isles. Each dot represents a unique pair of individuals coloured according to whether: (a) individuals were from the same or different sites, and (b) individuals had same or different ecotypes. Linear regressions of phenotypic distance on genetic distance are represented as dashed lines. In a) and b), top histogram shows the distribution of genetic distance, whereas right-hand histogram shows the distribution of phenotypic distance.

Discussion

This study confirms the distinctiveness of *Littorina fabalis* as being genetically different from species in the *L. saxatilis* complex. The results also showed that *L. saxatilis*, *L. arcana* and *L. compressa* from the British Isles were not reciprocally monophyletic based on the markers analysed here. This agrees with previous works that reported identical mtDNA haplotypes and extensive lineage sharing between species (Small & Gosling 2000b; Wilding *et al.* 2000a; b; Doellman *et al.* 2011; Panova *et al.* 2011). However, the separate grouping of *L. arcana* and *L. compressa* with geographically adjacent *L. saxatilis* populations in this study suggests the possibility of ongoing gene flow between species since earlier reports suggest that *L. saxatilis* and *L. arcana* can hybridise (Warwick *et al.* 1990; Mikhailova *et al.* 2009). However, other evidence points towards lineage sharing being more likely due to common ancestry (Wilding *et al.* 2000a; b; Doellman *et al.* 2011). More specimens of *L. arcana* and *L. compressa* from more sites are needed to clarify the phylogenetic relationships among the species in the *L. saxatilis* complex. Alternatively, one could use molecular phylogenetic methods, such as maximum likelihood and Bayesian inference, which depend on appropriate models of DNA sequence evolution.

Genetic variation within *L. saxatilis* showed considerable geographical structure, which was especially clear at a more regional scale rather than at a broader one. The observed complexity of phylogeographic patterns agrees with previous reports (Doellman *et al.* 2011; Panova *et al.* 2011), which described that the most complex phylogeographic patterns in *L. saxatilis* were found in the region between the Faroe Islands and the Brittany Peninsula, a large part of which was examined in this study.

It has been suggested that the geographical complexity of the shore in this region, combined with repeated glaciation events that produced changes in the sea level and ice-cover, resulted in the persistence of marine taxa in glacial

refugia, including the Western English Channel, southwest Ireland, the Faroe Islands and Iceland (Maggs *et al.* 2008; Provan & Bennett 2008). Previous genetic evidence based on mtDNA markers has suggested that *L. saxatilis* could have persisted in all these refugia during glacial periods, and recolonised from there into different parts of the British Isles (Doellman *et al.* 2011; Panova *et al.* 2011). Compared to other species, Panova *et al.* (2011) have suggested that *L. saxatilis* could have better resisted the glacial disturbances due to the advantages that bearing live young confer – that is, a more effective dispersal over long distances and colonisation of remote areas given the potentially lower mortality rate of offspring occurring in egg masses or brooding animals (Johannesson 1988) – and to the diversity of habitats that the species currently occupies (Reid 1996).

In this context, it can be hypothesised that some of the genetic groups identified here may have each come from different refugia. However, identification of such refugia and routes of colonisation are beyond the scope of this study and require further investigation. The results of the neighbour-joining tree suggested a trend for sites to aggregate into five main such groups within *L. saxatilis*: a) NE Britain sites Thornwick Bay, Old Peak, Amble, St Abbs, Dunbar, and Aberdeen; b) S British Isles sites Dersingham, St Margarets at Cliffe, and Wexford; c) W Scotland sites Oban I, Oban II and Isle of Mull I; d) W Britain sites East Prawle, Broad Haven, Cable Bay and Isle of Mull II, and e) W Ireland sites Lettermullan and lochs Aibhnin, Tanai and Fhada, and the NE Scotland site Castletown. Of these, the NE Britain group was the most coherent and consistently identified cluster.

Despite the general phylogeographic patterns observed, current geographical proximity of sites did not in all instances correspond with genetic similarity, although the overall data showed a clear pattern of isolation by distance. One clear case of this was *L. saxatilis* from Castletown. However, it should be considered that the coastline during glacial periods was not the same as it currently is and that the ice cover might not have retreated at the same rate everywhere. During the Last Glacial Maximum (LGM) the Scandinavian Ice Sheet covered most of the British Isles, except SW Ireland and S Britain (Pflaumann *et al.* 2003). Different layout of the coastline and varying ice retreat rates could have led to recolonisation of different sites at different times, even

perhaps over relatively long distances if habitat availability was still limited, since there is evidence that *L. saxatilis* is capable of long-distance colonisation (Johannesson 1988).

Among the three classes of *L. saxatilis* ecotypes analysed here (H and M, *neglecta* and *tenebrosa*), genetic variation was more strongly structured by geography than by ecotype, which coincides with previous findings (Butlin *et al.* 2014). Despite this observation coming from an ordination method (i.e. PCA) that assumes no evolutionary model – although it has been shown that the method can be used to detect population structure in large datasets (Patterson *et al.* 2006) – there was a general trend of individuals to group by geographical region rather than by ecotype. Evidence of this was the fact that, for example, *neglecta* from NE Britain was genetically more similar to H and M from the same region, and that *neglecta* from W Britain and the Isle of Man were more similar to H and M from that region. These two groups of *neglecta* individuals greatly differed. Independent origin of *tenebrosa* could also be hypothesised given that individuals from Dersingham were genetically more similar to H and M from the nearby site St Margarets at Cliffe, whereas *tenebrosa* from W Ireland were clearly more similar to H and M from the nearby sites. These results constitute a powerful framework to test specific hypotheses about ecotype formation in *L. saxatilis*. Additionally, it has been suggested that certain scatter patterns of PCA plots can result from IBD gradients (Novembre & Stephens 2008; Reich *et al.* 2008), but interpretation needs to be very carful in these cases.

The Structure analysis showed that most individuals had mixed membership in multiple groups. In cases like this, Pritchard *et al.* (2000) have suggested that the cause might be the presence of rather arbitrary allele frequencies in different sites that gradually vary across geographical space, perhaps with some local dispersal. They have also suggested that the algorithms implemented in structure are not well suited to data from an IBD scenario, in which cases the interpretation of the results might be challenging. In cases like this, the inferred number of clusters (*K*) may also be rather arbitrary and without much biological sense. Furthermore, Falush *et al.* (2003) suggested that the correlated allele frequency model might improve the clustering, but it may also increase the risk of over-estimating *K*. Additionally, the steady increase of L(*K*)

with *K* might result from IBD rather than by a signature of strong subdivision in the data (Guillot *et al.* 2009).

The significance of the Mantel's test in the IBD analysis suggested a pattern of spatial structure. This pattern could be explained by the fact that *L. saxatilis* usually has small individual dispersal distances in the rage of 2–10 m per generation (Janson 1983; Erlandsson *et al.* 1998) compared to its wide distributional area (but see Johannesson 1988). However, a similar result of Mantel's test could be obtained in cases where barriers to gene flow exist between otherwise panmictic populations (Guillot *et al.* 2009). For example, Fontaine *et al.* (2007) studied the harbour porpoise *Phocoena phocoena* and found a significant association between geographic and genetic distance matrices. However, they showed that an environmental variable (availability of nutrients) was likely associated with the gene flow barrier and was therefore having a major impact on the spatial structure of the species. A similar possibility in *L. saxatilis* should be explored since not all proportions of inferred clusters seemed to gradually vary across the sampled region. For instance, it is perhaps unlikely that such high levels of isolation, as measured by genetic distance and population structure, between Castletown and Aberdeen, and between Ceann Trá and the rest of W Ireland sites, could be solely explained by geographic distance since they are relatively adjacent to one another.

Exploratory and descriptive analyses of the genetic data generated in this study revealed certain structure in the data, but interpretation is not so clear given that a pattern of IBD was supported. This evidences the potential of this dataset to test specific hypotheses about the demographic history and colonisation of *L. saxatilis* from the British Isles. For instance, different colonisation routes could be modelled and tested using coalescent simulations under an approximate Bayesian computation (ABC) approach (Beaumont 2010). Such models could include colonisation of sites in a stepping-stone migration way, as has been proposed for the colonisation of Greenland (Panova *et al.* 2011). Furthermore, the genetic dataset produced in this study holds the potential for genome scans in order to identify putative highly differentiated loci between ecotypes from the local scale to the whole region examined here.

In summary, this study provides new insight into the geographical distribution of genetic variation within the *L. saxatilis* complex in the British Isles, and illustrates the potential role of glacial cycles in the current distribution of this group. It also provides the data and sets the foundations for more in-depth examination of the demographic history and colonisation of the British Isles. Future studies should consider more stringent filters of genetic data in order to increase confidence in the patterns observed.

Acknowledgements

I am thankful to the following people and institutions for the help provided to carry out this research. John Grahame (University of Leeds) provided detailed information about some of the collection sites. The Society for the Study of Evolution (SSE) provided travel expenses for fieldwork through the Rosemary Grant Graduate Student Research Award. Sophie Webster provided guidance on data collection. The following people helped during fieldwork: Steeves Buckland, Sara Cruz, Sara Hintz, Coralie Hopwood, Kerstin Johannesson, Anne-Lise Liabot, Cheryl Mills, Sophie Webster, and Anja Westram. The following people from the Molecular Ecology Lab (University of Sheffield) provided technical support: Laura Brettell, Gavin Horsburgh, Andy Krupa, Maria-Elena Mannarelli, Rachel Tucker, and Anja Westram. I am thankful to Anja Westram and Roger Butlin for their help in the planning of probe design and data analysis.

References

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

Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 379–406.

Bolnick DI, Fitzpatrick BM (2007) Sympatric speciation: models and empirical evidence. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 459–487.

Butlin RK, Galindo J, Grahame JW (2008) Sympatric, parapatric or allopatric: the most important way to classify speciation? *Philosophical Transactions of the Royal Society B*, **363**, 2997–3007.

Butlin RK, Saura M, Charrier G *et al.* (2014) Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. *Evolution*, **68**, 935–949.

Colson I, Hughes RN (2007) Contrasted patterns of genetic variation in the dogwhelk *Nucella lapillus* along two putative post-glacial expansion routes. *Marine Ecology Progress Series*, **343**, 183–191.

Coyer JA, Peters AF, Stam WT, Olsen JL (2003) Post-ice age recolonization and differentiation of *Fucus serratus* L. (Phaeophyceae; Fucaceae) populations in Northern Europe. *Molecular Ecology*, **12**, 1817–1829.

Doellman MM, Trussell GC, Grahame JW, Vollmer S V. (2011) Phylogeographic analysis reveals a deep lineage split within North Atlantic *Littorina saxatilis*. *Proceedings of the Royal Society B*, **278**, 3175–3183.
Earl DA, VonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.
Erlandsson J, Rolán-Alvarez E, Johannesson K (1998) Migratory differences between ecotypes of the snail *Littorina saxatilis* on Galician rocky shores. *Evolutionary Ecology*, **12**, 913–924.
Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, **14**, 2611–2620.
Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.
Fontaine MC, Baird SJE, Piry S *et al.* (2007) Rise of oceanographic barriers in continuous populations of a cetacean: the genetic structure of harbour porpoises in Old World waters. *BMC Biology*, **5**, 30.
Gnirke A, Melnikov A, Maguire J *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, **27**, 182–189.
Guillot G, Leblois R, Coulon A, Frantz AC (2009) Statistical methods in spatial genetics. *Molecular Ecology*, **18**, 4734–4756.
Hoarau G, Coyer JA, Veldsink JH, Stam WT, Olsen JL (2007) Glacial refugia and recolonization pathways in the brown seaweed *Fucus serratus*. *Molecular Ecology*, **16**, 3606–3616.
Jakobsson M, Rosenberg NA (2007) CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.
Janson K (1983) Selection and migration in two distinct phenotypes of *Littorina saxatilis* in Sweden. *Oecologia*, **59**, 58–61.
Johannesson K (1988) The paradox of Rockall: why is a brooding gastropod (*Littorina saxatilis*) more widespread than one having a planktonic larval dispersal stage (*L. littorea*)? *Marine Biology*, **99**, 507–513.
Johannesson K (2001) Parallel speciation: a key to sympatric divergence. *Trends in Ecology and Evolution*, **16**, 148–153.
Johannesson K (2003) Evolution in *Littorina*: ecology matters. *Journal of Sea Research*, **49**, 107–117.
Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, **27**, 3070–3071.
Jones MR, Good JM (2015) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*.
Maggs CA, Castilho R, Foltz D *et al.* (2008) Evaluating signatures of glacial refugia for North Atlantic benthic marine taxa. *Ecology*, **89**, S108–S122.
McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.
Mikhailova NA, Gracheva YA, Backeljau T, Granovitch AI (2009) A potential species-specific molecular marker suggests interspecific hybridization between sibling species *Littorina arcana* and *L. saxatilis* (Mollusca, Caenogastropoda) in natural populations. *Genetica*, **137**, 333–340.
Nei M (1972) Genetic distance between populations. *The American Naturalist*, **106**, 283–292.
Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89**, 583–590.
Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, **40**, 646–649.
Panova M, Blakeslee AMH, Miller AW *et al.* (2011) Glacial history of the North Atlantic marine snail, *Littorina saxatilis*, inferred from distribution of mitochondrial DNA lineages. *PLoS ONE*, **6**, e17511.
Paradis E (2010) pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics*, **26**, 419–420.
Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.
Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics*, **2**, e190.
Pflaumann U, Sarnthein M, Chapman M *et al.* (2003) Glacial North Atlantic: sea-surface conditions reconstructed by GLAMAP 2000. *Paleoceanography*, **18**, 1065.

Philippe H, Brinkmann H, Lavrov D V. *et al.* (2011) Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biology*, **9**, e1000602.
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.
Provan J, Bennett K (2008) Phylogeographic insights into cryptic glacial refugia. *Trends in Ecology and Evolution*, **23**, 564–571.
Provan J, Wattier RA, Maggs CA (2005) Phylogeographic analysis of the red seaweed *Palmaria palmata* reveals a Pleistocene marine glacial refugium in the English Channel. *Molecular Ecology*, **14**, 793–803.
Ravinet M, Westram A, Johannesson K *et al.* (2015) Shared and nonshared genomic divergence in parallel ecotypes of Littorina saxatilis at a local scale. *Molecular Ecology*, **25**, 287–305.
Reich D, Price AL, Patterson N (2008) Principal component analysis of genetic data. *Nature Genetics*, **40**, 491–492.
Reid DG (1996) *Systematics and evolution of* Littorina. The Ray Society, London.
Reid DG, Dyal P, Williams ST (2012) A global molecular phylogeny of 147 periwinkle species (Gastropoda, Littorininae). *Zoologica Scripta*, **41**, 125–136.
Robinson JD, Bunnefeld L, Hearn J, Stone GN, Hickerson MJ (2014) ABC inference of multi-population divergence with admixture from unphased population genomic data. *Molecular Ecology*, **23**, 4458–4471.
Rolán-Alvarez E, Buño I, Gosalvez J (1996) Sex is determined by sex chromosomes in *Littorina saxatilis* (Olivi) (Gastropoda, Prosobranchia). *Hereditas*, **267**, 261–267.
Rosenberg NA (2004) DISTRUCT: A program for the graphical display of population structure. *Molecular Ecology Notes*, **4**, 137–138.
Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, **145**, 1219–1228.
Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.
Seeb JE, Carvalho G, Hauser L *et al.* (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources*, **11 S1**, 1–8.
Shaffer HB, Thomson RC (2007) Delimiting species in recent radiations. *Systematic Biology*, **56**, 896–906.
Small MP, Gosling EM (2000) Genetic structure and relationships in the snail species complex *Littorina arcana* Hannaford Ellis, *L. compressa* Jeffreys and *L. saxatilis* (Olivi) in the British Isles using SSCPs of cytochrome-b fragments. *Heredity*, **84**, 692–701.
Wares JP, Cunningham CW (2001) Phylogeography and historical ecology of the North Atlantic intertidal. *Evolution*, **55**, 2455–2469.
Warwick T, Knight AJ, Ward RD (1990) Hybridisation in the *Littorina saxatilis* species complex (Prosobranchia: Mollusca). *Hydrobiologia*, **193**, 109–116.
Wilding CS, Butlin RK, Grahame J (2001) Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *Journal of Evolutionary Biology*, **14**, 611–619.
Wilding CS, Grahame J, Mill PJ (2000a) Mitochondrial DNA Col haplotype variation in sibling species of rough periwinkles. *Heredity*, **85**, 62–74.
Wilding CS, Grahame J, Mill PJ (2000b) Nuclear DNA restriction site polymorphisms and the phylogeny and population structure of an intertidal snail species complex (*Littorina*). *Hereditas*, **133**, 9–18.

Chapter IV

Testing for reinforcement between two sister species of mangrove snails (*Littoraria cingulata* and *Littoraria filosa*) from Australia

Abstract

Natural selection can strengthen the reproductive isolation between two divergent lineages that come into secondary contact, a process known as reinforcement. However, how often this process contributes to speciation is still poorly understood. In many cases, reinforcement can produce stronger assortative mating in areas where the two lineages overlap, usually leading to reproductive character displacement, a pattern in which reproductive traits differ more in sympatry than in allopatry. These patterns have been observed in the pair of sister species of mangrove snails *Littoraria cingulata* and *Littoraria filosa* from Australia. Here, DNA sequences from a reduced representation of the genome were used to test specifically whether gene flow occurred or not upon secondary contact of these two species that are thought to have formed in allopatry. Among the interspecific comparisons of genetic differentiation, lower values were found between the sympatric populations, suggesting either higher levels of interspecific gene flow or lower levels of genetic drift in sympatry. Two demographic history models were tested under an ABC approach, differing only in the presence/absence of recent interspecific migration. The model with no interspecific migration received more support, suggesting that reinforcement could not have contributed to speciation in this system. However, it has been suggested that oversimplified models may obscure evolutionary processes in cases of complex demographic histories. Therefore, future studies on this system should test more complex and flexible demographic models, for instance, allowing for asymmetrical gene flow between populations and independent estimation of ancestral population sizes.

Introduction

The central idea of the biological species concept proposed by Mayr (1942) is that groups are reproductively isolated from each other, and thus not able to interbreed. This concept came after Dobzhansky (1937) discussed the role of reproductive isolation in the formation of new species. Despite the long-standing debate on species concepts in biology (reviewed by de Queiroz 2007), the understanding of how reproductive barriers between groups of organisms arise and evolve remains of great importance when studying biodiversity and the formation of new species. The exchange of genes between diverging groups through hybridisation plays major and varied roles in speciation depending on the stage of divergence (Abbott *et al.* 2013).

The level of gene flow between divergent groups upon secondary contact can have different effects depending on how divergent those groups are (Abbott *et al.* 2013). In this context, high levels of gene flow between divergent sister-lineages will tend to erode accumulated differences between groups and homogenise populations back into a single group. At the other extreme of the continuum of levels of gene flow, absence of gene exchange would indicate completion of the speciation process. However, gene flow at a certain intermediate stage of this continuum – where hybrids are selected against because they are unfit or mating is too costly – can tip the balance towards speciation by creating a selection pressure favouring the strengthening of the reproductive barriers, a process known as reinforcement (Dobzhansky 1937; Blair 1955; Butlin 1987; Servedio & Noor 2003).

The strict definition of reinforcement (Dobzhansky 1937; Butlin 1987) contemplates epistatic interactions between genes from two parental lineages with a negative effect on the offspring. Such intrinsic genetic interactions, known as Dobzhansky-Muller incompatibilities (Dobzhansky 1937; Muller 1942), may cause hybrid inviability, hybrid sterility or reduced fitness of hybrids (e.g. Wang *et al.* 2015). On the other hand, postzygotic extrinsic incompatibilities may be mainly related to ecological conditions (Coyne & Orr 1998; Kirkpatrick 2001), where fitness of hybrids is reduced due to their inability to fully exploit their environment (e.g. Nosil *et al.* 2003). Authors supporting broad-sense reinforcement (Servedio & Noor 2003) consider cases of increased reproductive

isolation between diverging lineages due to prezygotic incompatibilities (e.g. damage of genitalia due to heterospecific matings) as part of reinforcement's contribution to speciation (Servedio 2001).

Non-random mating is a key part in reinforcement and it can be achieved by a one-allele or two-allele mechanism at a single locus associated with loci affected by natural selection (Felsenstein 1981). It has been suggested that non-random mating would more likely lead to speciation by reinforcement under the one-allele model because no linkage disequilibrium is necessary to increase the frequency of the selected allele (Servedio & Noor 2003, and references therein). In contrast, in a two-allele model, recombination could break up linkage disequilibrium between the assortative mating locus and loci affected by selection. However, theory predicts that recombination suppressors could enhance processes leading to speciation involving reinforcement (Trickett & Butlin 1994; Butlin 2005), as could genetic modifiers that reduce recombination rates between diverging lineages in the face of gene flow (Ortiz-Barrientos *et al.* 2016). Furthermore, the reduced recombination rate of sex chromosomes suggests that these might have an important role in reinforcement (Ortiz-Barrientos *et al.* 2009), given that prezygotic isolation might spread more readily if either the mating preference loci or the genetic incompatibility loci are on the sex chromosome (Kelly & Noor 1996; Lemmon & Kirkpatrick 2006).

Most of the criticism that reinforcement has received is related to the homogenising effect that gene flow and recombination can have on diverging lineages, which in some instances would tend to break beneficial genetic associations between mating preference and underdominant loci (reviewed by Butlin 1995; Servedio & Noor 2003; Coyne & Orr 2004; Ortiz-Barrientos *et al.* 2009). However, since convincing evidence of reinforcement exists (e.g. Bewick & Dyer 2014; Beysard *et al.* 2015; Smadja *et al.* 2015), the focus has changed from questioning the occurrence of reinforcement to assessing its frequency and role in speciation (Servedio & Noor 2003).

In this context, rates and patterns of gene flow, which can be highly influenced by geography, have been analysed in some systems. For example, Nosil *et al.* (2003) analysed the effects of migration asymmetries in the stick insect *Timema cristinae*, and found that the effect of reinforcement was strongest when the

populations were of similar size. A study by Nadachowska-Brzyska *et al.* (2013) provide another example where different levels and patterns of gene flow were investigated, finding recent unidirectional gene flow between two species of *Ficedula* flycatchers, a system in which reinforcement has been suggested (Saetre *et al.* 1997). Rates and patterns of gene flow have even been shown to vary across loci within genomes of hybridising taxa (e.g. sea quirts; Roux *et al.* 2013), a situation that should be considered when looking for reinforcement.

One characteristic of reinforcement is a pattern of higher reproductive isolation in sympatry than in allopatry (Butlin 1987). Natural or sexual selection acting on mating cues at the time of secondary contact of diverging lineages can also generate this pattern by strengthening premating isolation (Servedio & Noor 2003). However, more theoretical and empirical studies should be conducted in order to discriminate between these two possible causes of increased reproductive isolation in sympatry.

A signature of reinforcement can be a pattern of reproductive character displacement (Butlin 1987), which means that a greater divergence of a mating trait in sympatric populations would be observed compared to allopatric populations, creating an increase in assortative mating. However, Butlin and Ritchie (2013) noted that the same pattern can arise in cases of reproductive interference, which is when genetically divergent populations come into secondary contact but do not exchange genes. This is a very important distinction to highlight because the lack of gene flow in the latter case indicates that the two parental forms have reached full species status, and therefore there is no longer a speciation process which reinforcement could contribute towards. However, some authors do include this case under a broad-sense reinforcement definition (Servedio & Noor 2003). Therefore, it is of vital importance to know whether genes were exchanged upon secondary contact of divergent populations that exhibit reproductive character displacement and assortative mating if it is to be known whether reinforcement contributed to speciation in such cases.

Examples of reproductive character displacement associated with mate recognition exist in the literature (e.g. Cooley *et al.* 2001; Jiggins *et al.* 2001; Höbel & Gerhardt 2003). Only a limited number of studies have investigated

pairs of sister species with diverse degrees of overlapping distributions within particular clades in order to analyse broader patterns of reproductive character displacement (Hollander *et al.* 2013) and assortative mating (Coyne & Orr 1989, 1997), specifically searching for cases of reinforcement. Hollander *et al*. (2013) showed that pairs of allopatric sister species of marine snails in the family Littorinidae tend to have more similar penis shape than those pairs with varying degrees of sympatry. This is in agreement with the idea that genital form has an important role in mate recognition and maintenance of reproductive isolation between closely related groups (Dufour 1844). This might be especially true in these marine snails, where penial morphology might be used for species recognition prior to insemination (Reid 1986, 1996). The results of Hollander *et al*. (2013), even though no specific comparisons between allopatric and sympatric populations were made, provide an important platform from which to study more in-depth patterns of reproductive character displacement and reinforcement.

The pair of sister species *Littoraria cingulata* and *Littoraria filosa* have recently been the focus of ongoing further investigation into this matter. The two species of snail inhabit mangrove forests on the north coast of Australia, with *L. filosa* having a much wider distribution extending from northern Western Australia to southeastern Queensland, while *L. cingulata* is restricted to the north and west of Western Australia (Reid 1986, 2001; Reid *et al.* 2010). This pair of species, both of which undergo planktotrophic development with free-swimming larvae that can potentially disperse over long distances, currently coexist in the north of Western Australia (Reid 1986), with only a very small number of putative hybrids yet found (D. G. Reid, personal communication).

This sympatric zone appears to be the result of secondary contact after allopatric divergence 2.6 (Reid *et al.* 2012) or 3.8 Mya (Reid *et al.* 2010), making this the youngest pair of sister species in the genus. Differential habitat choice is observed, with *L. cingulata* being mainly a trunk-dweller at higher tidal levels, and *L. filosa* a leaf-dweller higher up on trees at lower tidal levels (Reid 1986). Variation in allozymic loci and mitochondrial genes have revealed genetic structure within *L. filosa*, distinguishing between western and eastern samples (Inness-Campbell *et al.* 2003; Stuckey 2003; Reid *et al.* 2010). Genetic structure has also been reported for *L. cingulata* (Johnson & Black 1998; Reid

et al. 2010), with a subspecies being described in the most western part of its distribution.

Follow-up work on Hollander *et al.* (2013), who observed an overall penis similarity of 92% between the two species based on 18 traits of male genitalia, has found stronger assortative mating in sympatry than in allopatry for both species as measured by mating duration, suggesting greater isolation in sympatry (Hollander, Butlin and Reid, unpublished). The same team is also analysing penis shape in this system, testing for more direct evidence of reproductive character displacement, which is likely to be true given the observed assortative mating and penis shape differences already found by Reid (2001) between western and eastern *L. filosa*, coinciding with the geographical overlap with other *Littoraria* species and suggesting this to be a case of character displacement.

This chapter reports a study of population genomics and demographic history of *L. cingulata* and *L. filosa* from allopatric and sympatric sites, using molecular markers obtained from a reduced representation of the genome, in order to cast light on whether reinforcement contributes to speciation in this family of snails. The specific aims were to test whether gene flow between the two species has occurred since secondary contact and, if so, to explore different demographic history models to characterise gene flow. It was hypothesised that gene flow was likely to have occurred in the past – and therefore reinforcement contributed to speciation – given the low genetic divergence and short time since separation compared with other sister species pairs within the genus.

Methods

Sampling

Individuals of *L. cingulata* and *L. filosa* were collected from allopatric and sympatric sites by Johan Hollander (Lund University), Roger K. Butlin (University of Sheffield), and David G. Reid (Natural History Museum, London) (Table IV.1, Fig. IV.1). Allopatric sites were considered to be those where only one species occurred, whereas sympatric sites were considered to be those where both species occurred. Sites were selected based on previous sampling (Reid 1986, 2001; Reid *et al.* 2010). A total of eight samples were collected

according to the following scheme: *L. cingulata* from the two allopatric sites Denham and Monkey Mia, *L. cingulata* and *L. filosa* from the two sympatric sites Broome and Port Smith, and *L. filosa* from the two allopatric sites Darwin and Dundee Beach. For each of these eight groups, 30 female mantle tissue samples were preserved in 100% ethanol (absence of penis was scored as female), yielding a total of 240 individuals. For each species, pairs of sites that were geographically close together are further referred to as sympatric and allopatric regions.

Table IV.1 Collection sites of *L. cingulata* and *L. filosa*

Site	Site type	Species	Latitude (°)	Longitude (°)
Denham (De)	Allopatric	*L. cingulata*	–25.950482	113.559362
Monkey Mia (Mo)	Allopatric	*L. cingulata*	–25.798481	113.721269
Broome (Br)	Sympatric	*L. cingulata* *L. filosa*	–17.969273	122.237519
Port Smith (Po)	Sympatric	*L. cingulata* *L. filosa*	–18.512117	121.804787
Darwin (Da)	Allopatric	*L. filosa*	–12.408549	130.832430
Dundee Beach (Du)	Allopatric	*L. filosa*	–12.734647	130.356864

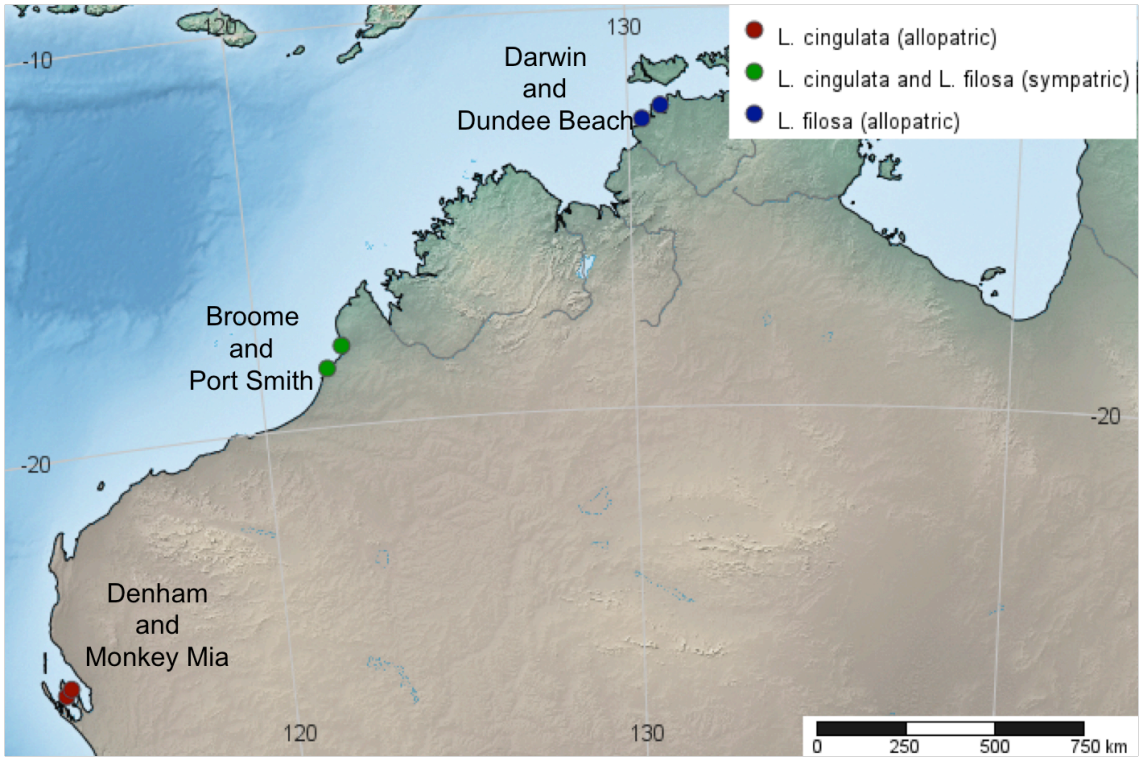


Fig. IV.1 Map showing the sites where *L. cingulata* and *L. filosa* were collected

Individuals were mostly collected from mangrove trees growing on sand, except at Monkey Mia, where snails were found under rocks. *L. cingulata* was mainly found living on the trunks at higher tidal levels, whereas *L. filosa* was found higher up on trees growing at lower tidal levels, living on the leaves. At all sites

snails were found on different trees, which sometimes were several tens of metres apart, with *L. filosa* at a lower tidal level than *L. cingulata*.

Lab work

DNA isolation

DNA isolation was carried out in the Molecular Ecology Laboratory, Department of Animal and Plant Sciences, University of Sheffield. Genomic DNA was extracted from 120 individuals (15 from each of the eight groups), using a modified version of the protocol from Wilding *et al*. (2001). Tissue was put in 500 µL 60 ºC CTAB buffer (2% CTAB, 1.4 M NaCl, 20 mM EDTA, 100 mM Tris-HCl pH 8, 0.2% β-mercaptoethanol) with 2 units of proteinase K, and incubated at 60 ºC for 15 h at an oscillation speed of 110 rpm. After this period, 0.3 mg ribonuclease was added, followed by incubation for 1 h at 60 ºC after vigorous mixing. Subsequently, 500 µL chloroform:isoamyl alcohol (24:1) was added, followed by gentle mixing for 10 min. DNA was isolated with 5 PRIME's Phase Lock Gel™ following the manufacturer's instructions. Then, 500 µL isopropyl alcohol was added to precipitate DNA, and incubated at room temperature for 5 min after gentle mixing. Samples were centrifuged at 4 ºC for 40 min at 13,000 rpm. After discarding the supernatant, DNA was washed with 1 mL 70% ethanol by gently mixing for 5 min, followed by centrifugation for 5 min at 13,000 rpm. Once the supernatant was discarded, the washing step was repeated with 500 µL 70% ethanol. Finally, the air-dried DNA pellet was dissolved in 50 µL 10% TE buffer (10 mM Tris-HCl pH 8, 1 mM EDTA). Purity was assessed by spectrophotometry, and concentration was measured by fluorometry.

DNA library preparation and sequencing

As previously discussed in Chapter I, several genotyping methods are available, none of which is suitable for all types of projects. However, technological and analytical advances have allowed NGS technologies to become widely employed for both discovery and genotyping of genetic markers (see Chapter I). In contrast with the study organism studied in Chapter III (*Littorina saxatilis*), there was no prior sequence information available for *Littoraria cingulata* or *Littoraria filosa*, and therefore a targeted capture would have been impractical for the time available to complete this study. Instead,

discovery and genotyping of SNPs was performed based on the method of multiplex shotgun genotyping (Andolfatto *et al.* 2011), which supersedes the traditional RAD approach (Baird *et al.* 2008) by avoiding DNA sharing and repair, and compensating with the use of a more frequent cutter. Furthermore, the multiplex shotgun genotyping method constituted a highly cost-effective option at the time of the study, this method being less labour-intensive that the original RADseq method.

Beijing Genomics Institute (Hong Kong) prepared and sequenced the DNA libraries using a protocol based on that of Andolfatto *et al*. (2011). DNA was digested with the *Ape*KI restriction enzyme to produce a high density of cut sites. This enzyme recognises the five-nucleotide sequence 5'-G/CWGC-3', where W can be A or T, and "/" represents the cutting site. Since no genome sequence is available for either species, the expected cutting frequency was calculated assuming that all four bases were present in equal proportions across the genome (i.e. 1/4 for each base). Thus, *Ape*KI was expected to cut every 512 bp (i.e. 1 / (1/4 * 1/4 * 1/2 * 1/4 * 1/4)). T4 DNA ligase was used for binding the DNA fragment ends to the sequencing adapters, which had a molecular identifier of four to eight bases to allow multiple individuals to be sequenced per lane. After inactivation of the ligase by heating, aliquots of each sample were pooled and non-ligated adapters removed using exclusion columns. Subsequently, DNA fragment pools were amplified by PCR using primers corresponding to the ligated adapters. Cleaned-up products were run on agarose gels, where amplified fragments were size-selected within a range of 300–600 base pairs, ensuring DNA insert length to be 200–500 base pairs, considering that the length of the adapters was 100 base pairs. This step was critical because it reduced the number of markers to positions where two cut sites fell within this range of separation. Finally, DNA libraries were validated with both Agilent 2100 Bioanalyzer (Agilent Technologies) and StepOnePlus™ Real-Time PCR (Life Technologies) systems. Pair-end 90 base sequencing was performed on the libraries using Illumina HiSeq systems.

Data analysis

Assembly of reads and genotyping

Reads were initially segregated based on their individual nucleotide identifier. After removing the adapter and individual identifier bases, reads with more than 50% low-quality bases (quality value ≤ 21 in a scale from 1 to 93) were removed. Reads with more than 10% non-determined bases were also removed. Remaining reads were assembled into tags allowing up to four base-pair mismatches, resulting in the identification of single nucleotide polymorphisms (SNPs). The decision to allow up to four mismatches was made considering divergence time between the two species and mutation rate, which ranged from 2.6e6 to 3.8e6 generations ago (Reid *et al.* 2010, 2012) – assuming one generation per year – and from 1.5e–9 to 1.5e–8 mutations per base per generation (Butlin *et al.* 2014), respectively. Thus, based on an expected read length of 82 bases after removing adapter and individual identifier bases, up to one mutation was expected every 18 bases, i.e. up to five mutations in 82 bases. Such a high number of mismatches could potentially lead to clusters of tags that do not actually represent the same locus, which would eventually lead to false SNP discovery. However, this might be rare as up to five SNPs per tag are expected. On the other hand, a lower number of mismatches could result in a given locus being split into more than one tag, and SNPs to be biased towards regions of low genetic diversity. From this stage on, data were differently processed for the population genetics and ABC analyses, as described within the respective sections.

Population genetics analyses

SNPs with fewer than five reads in total were removed, and heterozygous SNPs whose less frequent base was read fewer than two times were treated as homozygous. Beijing Genomics Institute did the bioinformatic analysis up to this point, i.e. they delivered genotype information per individual. Only the first SNP of each tag was kept in order to avoid treating SNPs in the same tag as independent from each other. Finally, tags genotyped in less than 80% of individuals were discarded.

The R package *adegenet* (version 1.4-2, Jombart & Ahmed 2011) was used for handling and analysing genotypes. After having removed loci with more than two alleles, a principal component analysis (PCA) was used for summarising the genetic diversity among the sampled individuals without assuming an evolutionary model. Another aim of the PCA was to visually identify potential outlier individuals as well as clusters of genetically similar individuals. Additional PCAs were used for further exploration of potential structure within species beyond the level of regions. Missing information was replaced by mean values in all PCAs. The lack of evident structure within regions allowed testing for Hardy-Weinberg equilibrium (HWE) at each locus within each of the four regions.

Loci significantly departing from HWE (p-value < 0.01) within any given region were removed from the whole dataset. In general, a locus can be considered to be in HWE when the difference between the expected and the observed heterozygosities (He–Ho) is minimal. Therefore, after having removed loci departing from HWE (p-value < 0.01), He–Ho was computed for each locus in the whole dataset and for each of the regions.

Population differentiation due to genetic structure was estimated using Nei's pairwise F_{ST}. Each value was computed as the expected heterozygosity minus the observed heterozygosity, all divided by the expected heterozygosity. Mean heterozygosities over all loci were used for each of 28 comparisons. Genetic distance, computed as $F_{ST}/(1–F_{ST})$, was plotted against geographic distance (Rousset 1997) to visualise genetic variation in the geographical context. Geographic distance between sites was measured around the coast in steps of 10 km. The R package *hierfstat* (version 0.04-14, Goudet 2005) was used for estimating hierarchical *F*-statistics over all loci across the nested levels of structure: species, regions, populations and individuals. In this analysis, the individual level within populations is equivalent to F_{IS} – a measure of departure from HWE after allowing for higher levels of structure (de Meeûs & Goudet 2007). This analysis was first applied to the whole dataset, and then to the two species separately because of their rather different regional structures.

Overview of the Approximate Bayesian computation analysis

An overview of the ABC approach is presented here, as described by Csilléry *et al.* (2010); specific details of it are given throughout the following sections. In brief, the observed genetic data are compared with simulated data that are generated under different models, which are unique hypothetical evolutionary scenarios to be tested. Each model consists of a set of parameters (e.g. population sizes, times since divergence, migration rates, etc.), whose ranges are specified based on prior knowledge. Then, under each model, a large number of datasets is simulated, each with the same known features (e.g. number of individuals, number of populations, number of loci, number of segregating sites per locus, etc.) of the observed data, but differing in the values of the parameters (known as priors) used for simulating different datasets. In other words, for each simulated dataset, priors are randomly drawn from their predefined range. Then, each dataset is reduced to a set of summary statistics, whose sampled priors are either accepted or rejected based on a pre-established distance (or tolerance) between the simulated and the observed summary statistics. The accepted priors are then transformed (or "adjusted") using local regression, and the resulting new parameter values constitute the so-called posteriors. These fitted parameter values allow evaluation of the uncertainty on the parameters given the observed summary statistics. Finally, models with higher posterior probability are preferred over those with lower posterior probability.

Preparation of the observed data for the ABC analysis

For each species, sites within regions were merged since population genetic analyses revealed no evidence of differentiation at this level. Therefore, four populations were considered for the ABC analyses, which are the allopatric and sympatric populations of each species. Only one allele for each individual and tag was retained because of the uncertainty associated with the low sequencing depth of the dataset, avoiding the risk of mistakenly calling a heterozygote as homozygote. For individuals scored as homozygotes, the allele was retained if it had more than one read. For individuals scored as heterozygotes, if only one allele had one read, then the other allele was retained, and if both alleles had more than one read, then one of them was randomly selected. At this stage, all

tags with any SNP with more than two states were discarded in order to avoid retaining tags potentially departing from the infinite alleles model (Kimura & Crow 1964), according to which the probability of a site to mutate back to its original state is very low. Thus, the removal of tags having SNPs with more than two states is justified because it would be difficult to distinguish these tags from those scored as monomorphic but where back mutations to the original state had occurred – all this assuming SNPs with more than two states are not just sequencing artefacts. Then, only tags with at least five individuals typed in each population were retained.

Then, only five individuals from each population were retained at random, which was expected to greatly simplify the simulation procedure and account for a bias induced by an originally truncated distribution of tags according to their number of SNPs. This approach simplified the procedure by having only one distribution of the bias for all tags and by avoiding the need to simulate locus-specific sample sizes. For some tags, these filters meant the loss of all alleles – either for low number of reads or at random – except one, making those tags monomorphic. For other tags, the filtering process caused the loss of individual SNPs within tags, making those positions monomorphic.

Tags that became monomorphic were excluded for simplicity and speed. Originally monomorphic tags were also excluded because they lacked sequencing depth information, and could therefore not be filtered in the same way (e.g. based on number of reads) as polymorphic tags were. Furthermore, the fact that their distribution according to the number of individuals typed was different from that of the polymorphic tags, and that the distribution of polymorphism among populations was also clearly different (Fig. IV.2), did not allow the retention of a number of monomorphic tags that was proportional to the retained number of polymorphic tags. At the end, 29,623 polymorphic tags were retained for passing the aforementioned filtering steps.

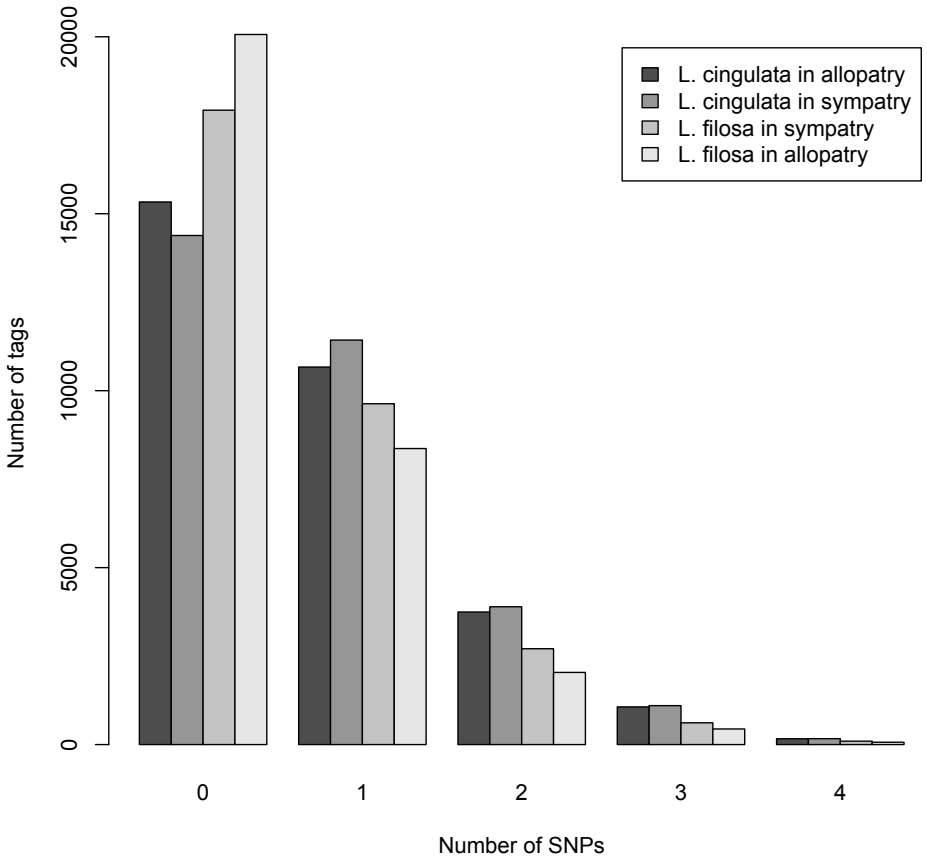


Fig. IV.2 Distribution of the 29,623 tags retained for the ABC analysis according to their number of SNPs in each population

After the filtering steps, the data were transformed into an ms alignment (Hudson 2002) in order to allow the further comparisons with the simulated data. Since no outgroup reference sequences were available for these species, ancestral and derived states of individual SNPs were assigned using a random haplotype as ancestral reference for each tag.

Simulations

Simulations were carried out with the programme msnseg, a modified version of the coalescent sampler msnsam (Hudson 2002; Ross-Ibarra *et al.* 2008) – kindly provided by Khalid Belkhir (Institut des Sciences de l'Évolution, Montpellier). Special attention was paid to simulating data with the same characteristics as the observed data. Therefore, each simulated dataset had 29,623 loci for 20 individuals – five from each population – and a constant locus length of 82 bases. Within datasets, for each simulated locus, up to 30 attempts were allowed before observing between 1 and 4 SNPs, which was the range of number of SNPs in the observed data. If the number of SNPs in the expected

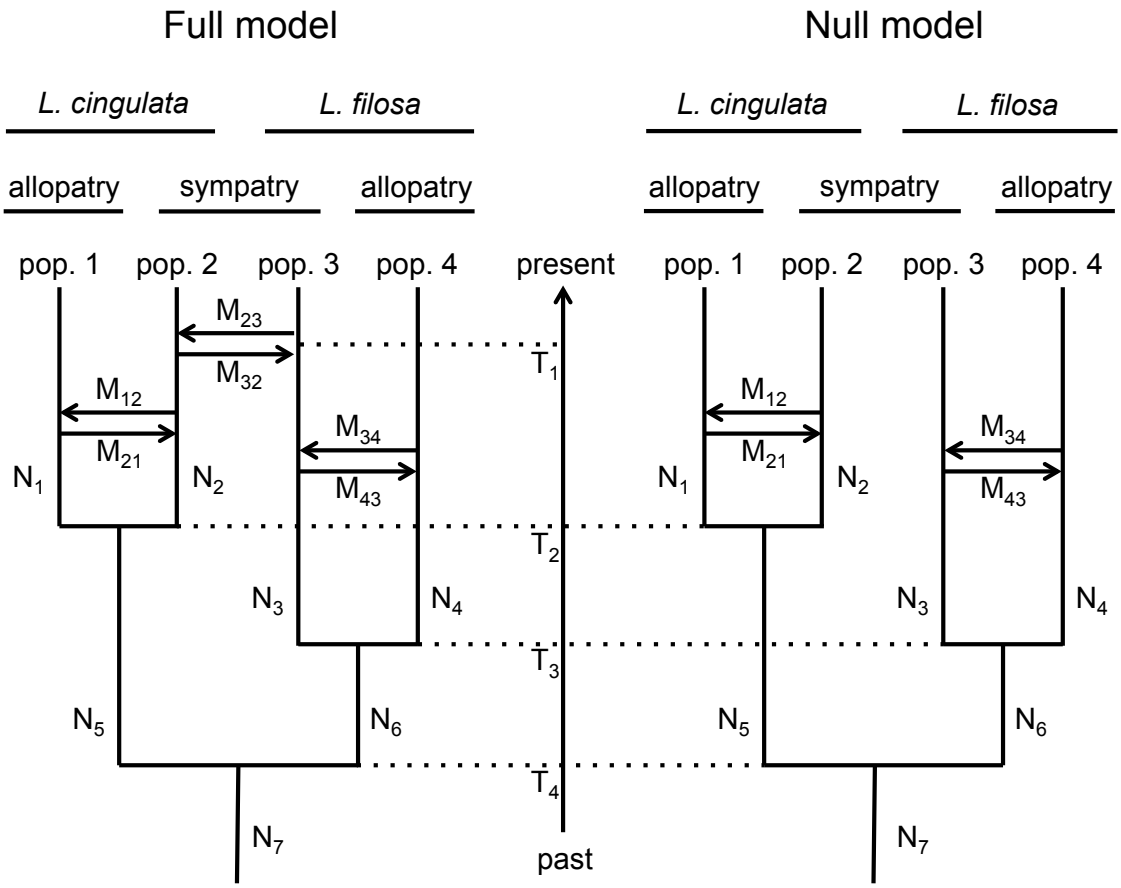
range was not observed after 30 attempts, then that specific locus was generated with however many SNPs were observed at the last attempt. Once a complete dataset of 29,623 loci had been simulated, if at least 1% of those loci did not have a number of SNPs in the expected range, then the dataset was discarded. Using this criterion, only about 1% of the datasets were discarded.

Demographic models and prior distributions

Two demographic models (Fig. IV.3) were tested in accordance with the aim of this chapter, which was to test whether interspecific gene flow occurred after secondary contact or not. Therefore, the two models exclusively differed in the presence/absence of interspecific migration. The model with interspecific migration is hereafter referred to as the full model, whereas that without interspecific migration is referred to as the null model. As previously mentioned, four populations were considered given the results of the population genetic analyses, i.e. allopatric *L. cingulata* (pop. 1), sympatric *L. cingulata* (pop. 2), sympatric *L. filosa* (pop. 3), and allopatric *L. filosa* (pop. 4). As required by the coalescent sampler, the parameters of the models were scaled by a factor of N_0, which was set to 1e4. Thus, effective population sizes (Ne) equalled N/N_0; times (T) since historical demographic events were simulated as $T/4N_0$, and migration parameters (M) were considered as $4N_0m_{ij}$, where m_{ij} is the fraction of population *i* that is made up of migrants from population *j*. One generation per year was assumed. Mutation rate (μ) was kept constant at 3e–9 mutations per base per generation.

For simplicity, the parameters of the full model are described first, followed by an account of which parameters were absent in the null model. From the past towards the present, the model starts with a common ancestral population size of N7. This population splits at time T4 into the ancestral populations of *L. cingulata* and *L. filosa* with population sizes of N5 and N6, respectively. Then, at time T3, the ancestral population of *L. filosa* splits into the sympatric and allopatric populations of sizes N3 and N4, respectively, allowing for symmetrical migration between these two populations until the present time. At time T2, the ancestral population of *L. cingulata* splits into the allopatric and sympatric populations of sizes N1 and N2, respectively, allowing for symmetrical migration between these two populations until the present time. Note that neither T2 nor

T3 was constrained to be greater than the other. After the two intraspecific splits, symmetrical interspecific migration begins at time T1 between sympatric populations (M23 and M32) until the present time. The null model differed from the full model in the exclusion of interspecific migration (M23 and M32) and time since this particular demographic event (T1).



Parameters

$N_{1,2,3,4}$ = current population sizes
N_5 = *L. cingulata* ancestral population size
N_6 = *L. filosa* ancestral population size
N_7 = common ancestral population size

T_1 = time of secondary contact
T_2 = time of *L. cingulata* split
T_3 = time of *L. filosa* split
T_4 = time of most ancient split

M_{ij} = migration rate into population *i* from population *j*

Fig. IV.3 Demographic models investigated in this study

The original ranges of priors were defined based on information available in the literature (Reid 1986; Turney *et al.* 2006; Reid *et al.* 2012) and on field observations – especially regarding density of snails – made by Johan Hollander (Lund University) and Roger K. Butlin (University of Sheffield). The assignment of an apparently large range to priors was necessary in order to properly estimate the posteriors, i.e. to ensure that the posteriors were included

within the priors. All parameter values were sampled from uniform distributions and set to a linear scale, except migration parameters, which were set to a log10 scale in order to concentrate the sampling on the rather relatively small values. After exploratory runs of 1e5 simulations for each model, the prior ranges were readjusted (Table IV.2) based on visual inspection of the posterior distributions until all the posteriors were included within the priors. Then, one million datasets were simulated under each model.

Table IV.2 Uniform prior distribution [low bound – high bound] after having ensured that they included the posteriors. Parameters marked with an asterisk (*) were unique to the full model. Population size parameters (N) are given in number of individuals; time parameters (T) are given in generations (assuming one generation per year), and migration parameters (M) are given in number of migrants per generation, i.e. $4N_0m_{ij}$, where N_0 = 1e4.

Parameter	Full model	Null model
N1	[5e4 – 1e6]	[5e4 – 1e6]
N2	[5e4 – 1e6]	[5e4 – 1e6]
N3	[5e4 – 1e6]	[5e4 – 1e6]
N4	[5e4 – 1e6]	[5e4 – 1e6]
N5, N6 & N7	[5e4 – 1e6]	[5e4 – 1e6]
T1*	[0 – min(T2, T3)]	
T2	[50 – T4]	[0 – T4]
T3	[50 – T4]	[0 – T4]
T4	[5e5 – 5e6]	[5e5 – 5e6]
M12 & M21	1e[–3 – 1]	1e[–3 – 1]
M23* & M32*	1e[–3 – 1]	
M34 & M43	1e[–3 – 1]	1e[–3 – 1]

Summary statistics

A set of summary statistics was computed for each population and for all possible population pairs, and used for the ABC inference. For each dataset – observed and simulated – mean and standard deviation were calculated across all 29,623 loci using msums, a modified version of mscalc (Roux *et al.* 2013) – kindly provided by Martin Hinsch and Ludovic Duvaux (University of Sheffield) – for the following summary statistics: sum of pairwise differences, number of segregating sites per locus, number of singleton sites per locus, Tajima's D (Tajima 1989), Tajima's theta (nucleotide diversity), Watterson's theta (Watterson 1975), Fu and Li's D* and F* (Fu & Li 1993), Ramos-Onsins and Rozas' R2 (Ramos-Onsins & Rozas 2002), raw nucleotide divergence (Nei's Dxy, equation 12.66, Nei & Kumar 2000), net nucleotide divergence (Nei's DA, equation 12.67, Nei & Kumar 2000), F_{ST} (Hudson *et al.* 1992), and number of biallelic sites. In total, 120 summary statistics were used considering nine

statistics for four populations, four statistics for six population pairs, and two values for each statistic, i.e. mean and standard deviation.

For each model, the full set of summary statistics was transformed via Partial Least Squares (Boulesteix & Strimmer 2006) in order to reduce its dimensions by extracting a set of orthogonal linear combinations of the summary statistics that best explained the variance in the model parameter space. Transformations were performed in the R package *pls* (version 2.5-0, Mevik & Wehrens 2007) and the 12 PLS components that best explained the variance were retained based on visual inspection of the root mean squared error plots (RMSEP plots). The transformed statistics were then used for computing the Euclidean distance between the observed and simulated datasets for the rejection step. Up to 1% (i.e. 1e4 datasets) of the transformed statistics nearest to the observed data were used for multivariate model parameter estimation via the non-linear regression correction algorithm “neural network” as implemented in the R package *abc* (version 2.1, Csilléry *et al.* 2012). Distributions of the posteriors were plotted using the R package *sm* (version 2.1, Bowman & Azzalini 2003). A symmetric credible interval of 0.9 was used for defining the lower and upper plausible limits for the estimated parameters.

Model selection

Posterior model probabilities were estimated based on the untransformed summary statistics using the R package *abc* (version 2.1, Csilléry *et al.* 2012). Euclidean distances were computed with the “neural network” algorithm and used for choosing the 1% of simulated datasets that were closest to the observed data. The model with higher posterior probability was chosen over the other model.

Results

Assembly of reads and genotyping

Genetic data were obtained for 113 individuals (Fig. IV.4), since seven of them failed the library preparation or sequencing step. The sequencing of DNA libraries generated 654,909,758 90-bp reads, 12% of which were assembled into 767,685 polymorphic tags (Fig. IV.5). After having filtered reads based on quality, SNPs on sequencing depth and number of SNPs per locus, and loci on proportion of genotyped individuals, as well as removing loci with more than two alleles, the dataset at this stage consisted of 1920 SNPs, one per locus.

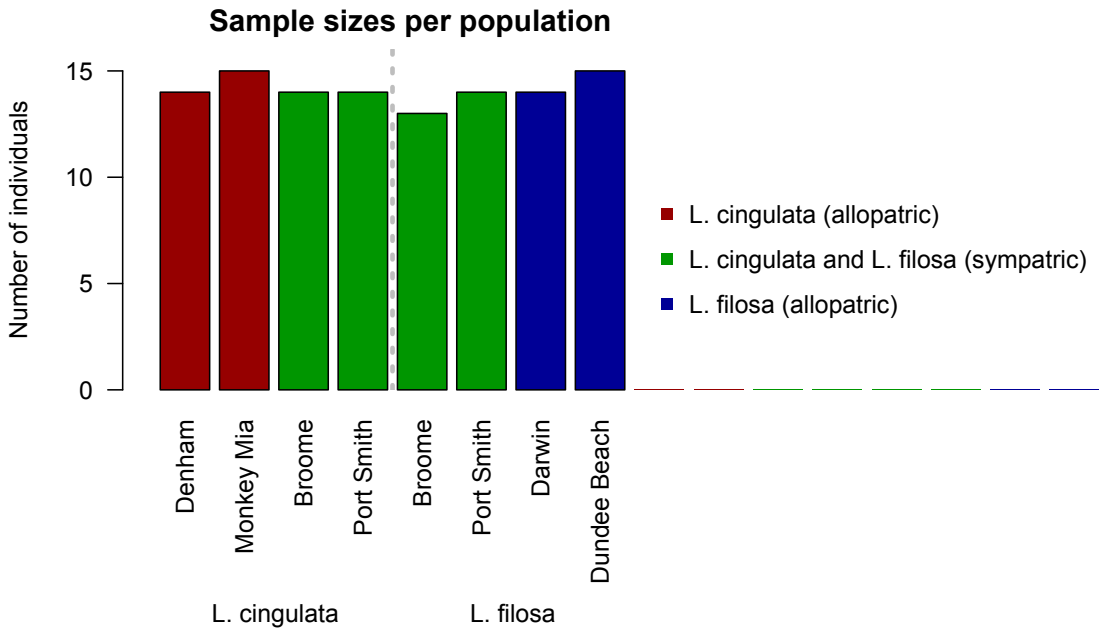


Fig. IV.4 Number of individuals per group used in this study

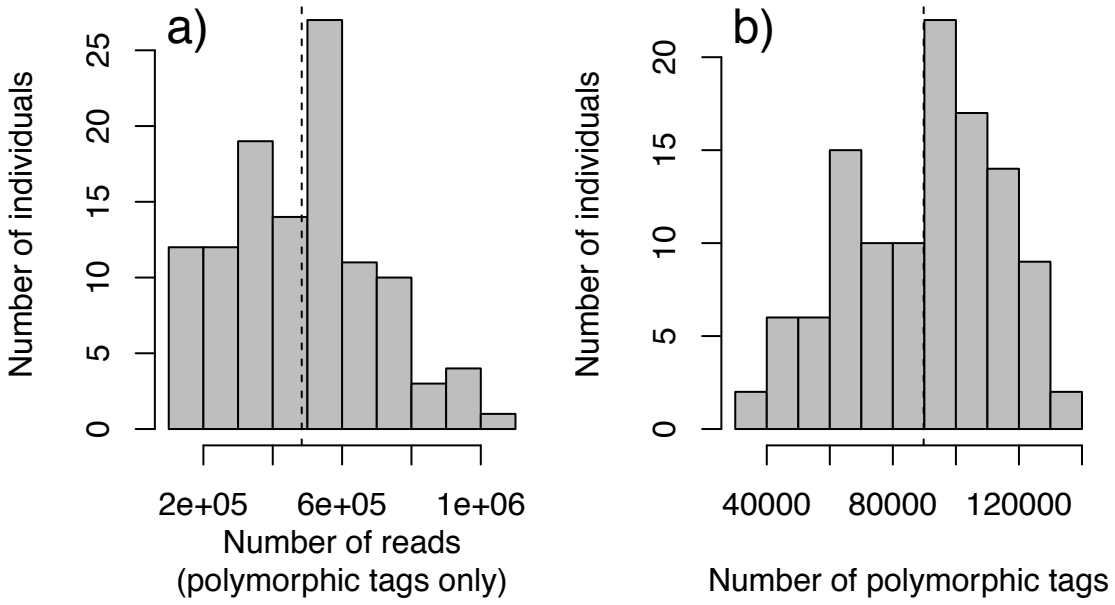


Fig. IV.5 Summary of reads that were assembled into polymorphic tags in each of the 113 *Littoraria* individuals. Reads and tags were counted immediately after assembling reads into tags. a) Number of reads per individual. b) Number of polymorphic tags per individual. Vertical dashed line in a) and b) represents the mean number of reads and tags per individual, respectively.

Population genetics analyses

The initial PCA revealed a clear structure within the data with no obvious outlier individuals. The same overall scatter pattern was observed after removal of loci departing from HWE (42 loci removed; details in the following section), thus the results presented here correspond to the latter. The first three principal components (PC) marked a clear separation at the expected different levels (Fig. IV.6): PC1 showed a distinction between *L. cingulata* and *L. filosa*, PC2 split *L. filosa* into the allopatric and sympatric regions, and PC3 did similarly for *L. cingulata*. The scree plots of the eigenvalues of the additional PCAs per species (Fig. IV.7) confirmed no further structure beyond the level of regions. The two regions within *L. cingulata* were found to be genetically more similar to each other than those within *L. filosa*. Notably, *L. cingulata* from the sympatric region, especially the one from Port Smith, showed a closer similarity to *L. filosa* along the species axis (PC1) and the *L. cingulata* axis (PC3) compared to the greater distance observed between *L. cingulata* from the allopatric region and *L. filosa*.

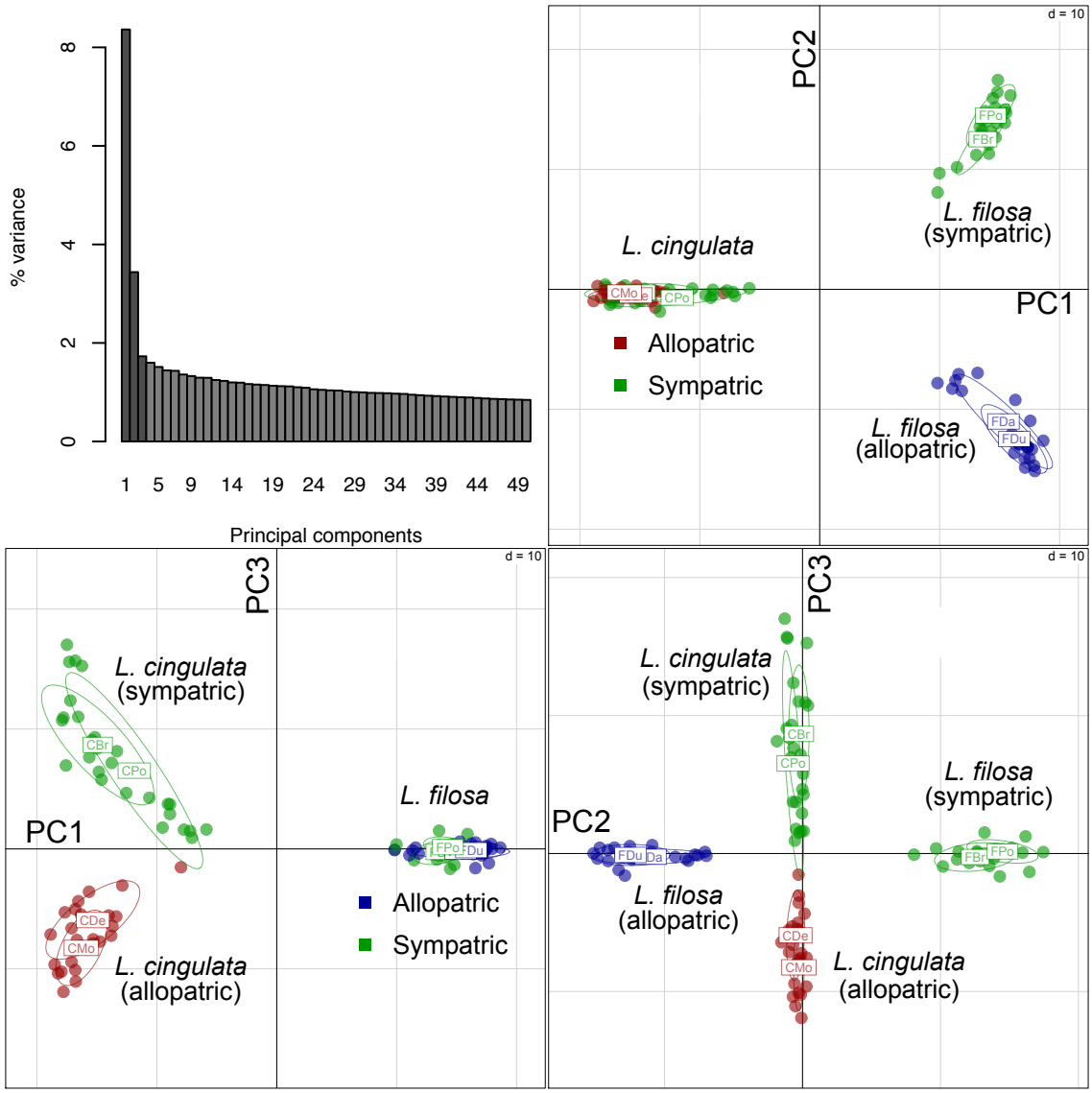


Fig. IV.6 Principal component analysis of all 113 individuals used in this study. Scree plot shows the proportion of the variance explained by the first 50 principal components. Scatter plots show the relationship between the first three principal components, where circles represent single individuals. Filtered loci had two alleles, were genotyped in at least 80% of individuals, and were in Hardy-Weinberg Equilibrium.

A total of 42 loci departing from HWE within regions (p-value < 0.01) were found and removed from the dataset. Of the remaining 1878 loci, 70% of them showed He–Ho values close to zero (–0.01 < x < 0.01) with a minimum of –0.1 and a maximum of 0.5 (Fig. IV.8). However, when computing He–Ho of loci within each of the four regions, the minimum values decrease beyond –0.1 (Fig. IV.9). Interestingly, considering each species separately, more biallelic loci are observed in the sympatric regions, meaning higher heterozygosity. This was accompanied by higher maximum He–Ho values in the sympatric region compared to the allopatric one, which means that there is a higher deficiency of heterozygotes within the sympatric region.

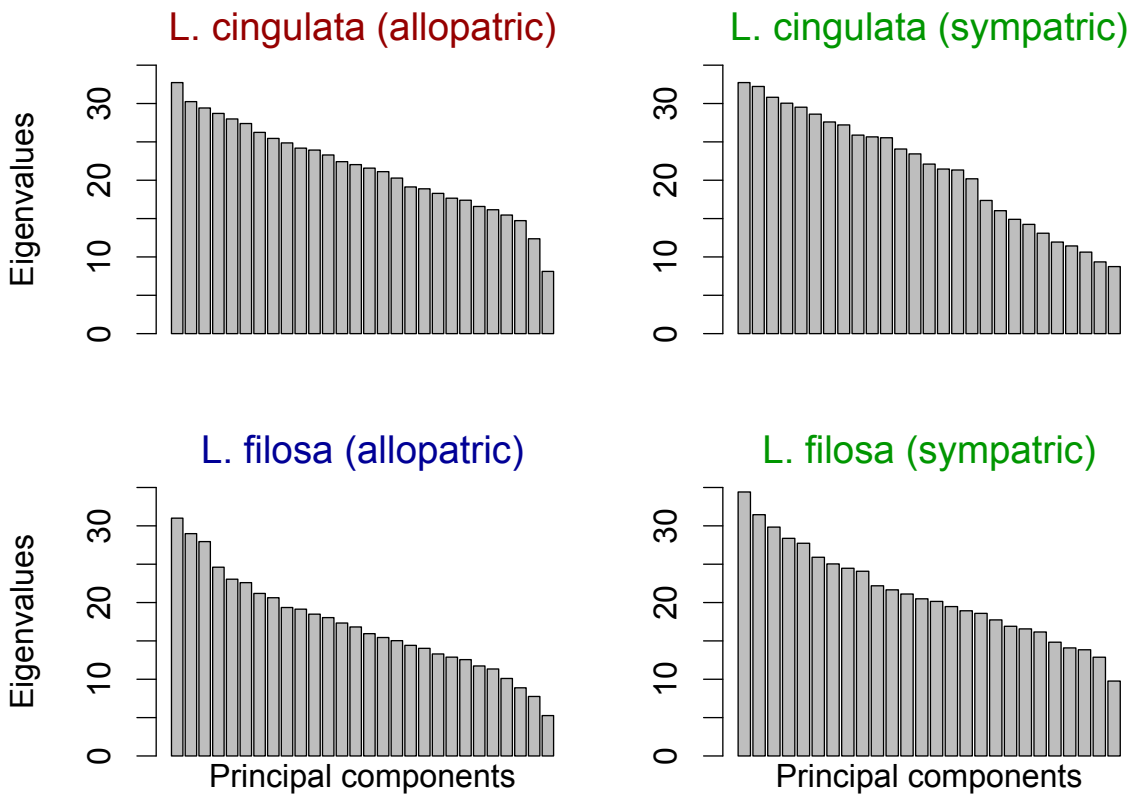


Fig. IV.7 Principal component analysis per species and region. Scree plots show the eigenvalues of the principal components.

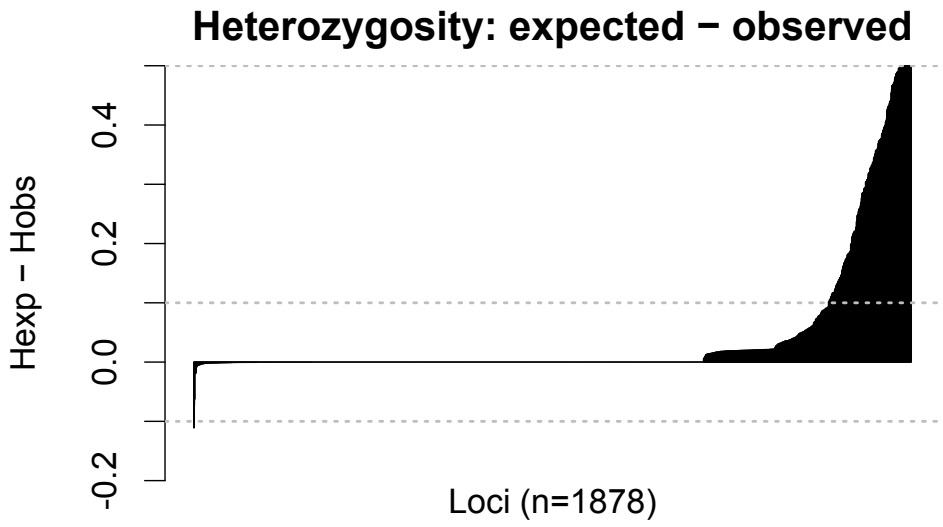


Fig. IV.8 Difference between expected and observed heterozygosities at each of the 1878 loci in Hardy Weinberg Equilibrium; loci sorted by He–Ho value

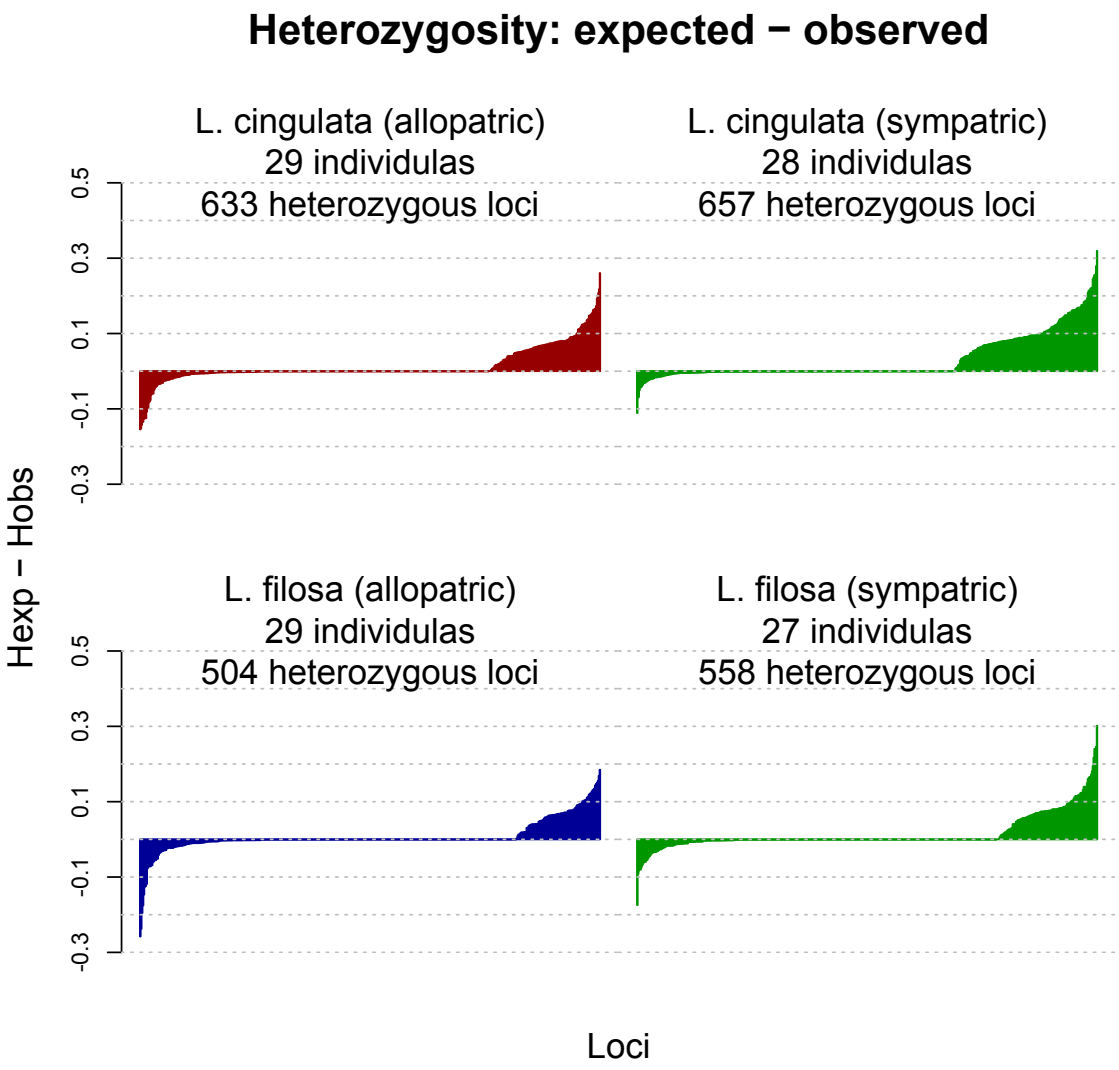


Fig. IV.9 Difference between expected and observed heterozygosities for each of the four regions; loci in Hardy Weinberg Equilibrium sorted by He–Ho value

Estimated F_{ST} values showed a pattern meeting the general expectation (Table IV.3), i.e. the greatest level of differentiation was found between species, whereas the lowest level of differentiation was observed between populations within regions (Fig. IV.10). Interspecific comparisons showed an interesting pattern as well, where F_{ST} was lowest among sympatric populations (0.42–0.44), followed by comparisons between allopatric *L. cingulata* and sympatric *L. filosa* (0.44–0.45), then by those between sympatric *L. cingulata* and allopatric *L. filosa* (0.46–0.48), and finally among allopatric populations (0.48–0.49). Intraspecific comparisons between sympatric and allopatric populations resulted in much lower F_{ST} in *L. cingulata* (0.07–0.08) than in *L. filosa* (0.25–0.30). Finally, intraspecific comparisons within regions led to slightly lower F_{ST} values between allopatric pairs (0.021–0.026) than between sympatric pairs (0.031–0.034).

Table IV.3 Nei's estimator of pairwise F_{ST}. Names in red and blue represent allopatric populations of *L. cingulata* and *L. filosa*, respectively; those in green represent sympatric populations of both species.

		L. cingulata				*L. filosa*		
		De	Mo	Br	Po	Br	Po	Da
L. cingulata	Mo	0.03						
	Br	0.07	0.07					
	Po	0.07	0.08	0.03				
L. filosa	Br	0.45	0.44	0.43	0.42			
	Po	0.45	0.45	0.44	0.43	0.03		
	Da	0.48	0.48	0.47	0.46	0.25	0.28	
	Du	0.49	0.49	0.48	0.46	0.27	0.30	0.02

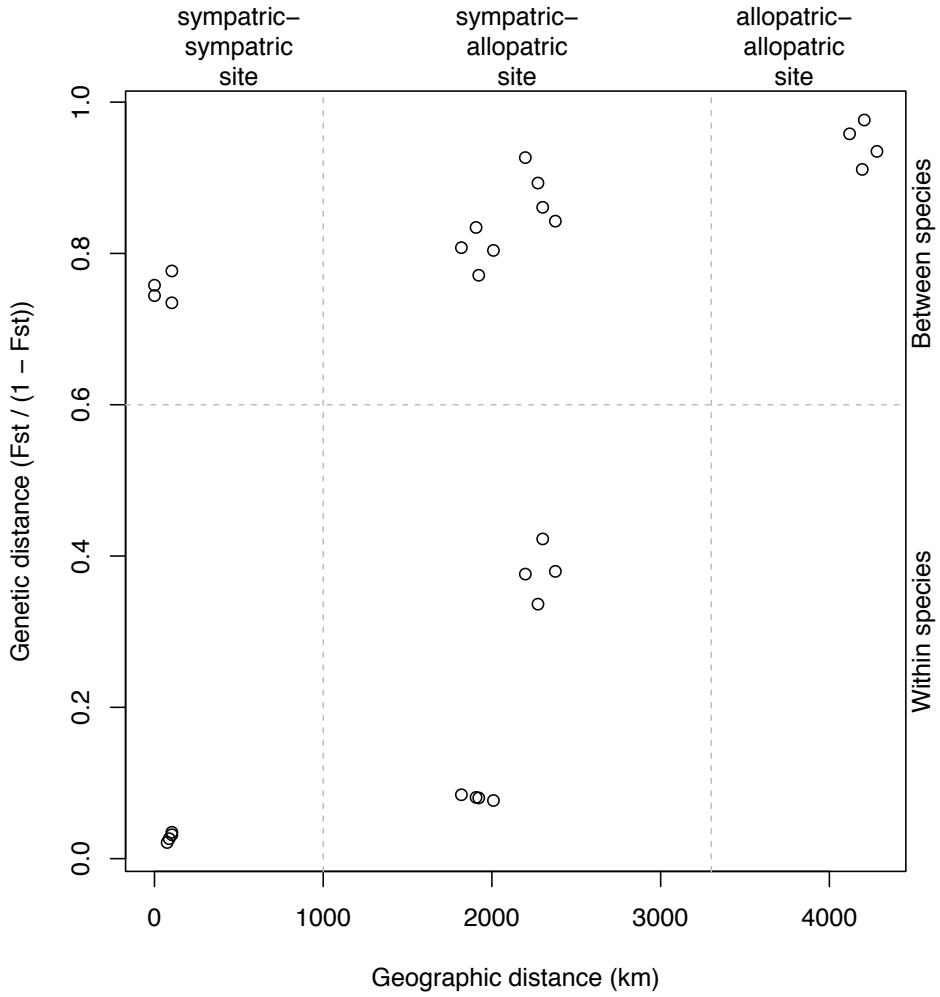


Fig. IV.10 Relationship between geographic and genetic distance between all eight sampled populations, representing *L. cingulata* and *L. filosa* both in allopatry and in sympatry. Each dot represents a unique pair of populations.

Hierarchical *F*-statistics over all loci across the nested levels of structure were estimated (Table IV.4). In general, the most relevant values in these tables are the *F*-statistic at any given level (*i*) within the immediately superior level (*i*–1) ($F_{\text{level}(i)/\text{level}(i-1)}$), which are located at the bottom of each column (de Meeûs & Goudet 2007). Considering the whole dataset, the maximum differentiation was

found between species ($F_{species/total}$ = 0.47), whereas the minimum one was found between populations within regions ($F_{populations/regions}$ = 0.01), rather than between individuals within populations ($F_{individuals/populations}$ = 0.14). When considering the two species separately, they both showed the lowest values when comparing populations within regions ($F_{populations/regions}$ = 0.01), results that are similar to their counterpart in the whole dataset. *L. cingulata* showed the greatest degree of differentiation between individuals within populations ($F_{individuals/populations}$ = 0.18), whereas *L. filosa* showed greater differentiation between the two regions ($F_{regions/total}$ = 0.40).

Table IV.4 Estimation of hierarchical *F*-statistics.

Both species		Species	Regions	Populations	Individuals
	Total	**0.47**	0.62	0.62	0.67
113 individuals	Species		**0.27**	0.28	0.38
	Regions			**0.01**	0.14
	Populations				**0.14**

L. cingulata		Regions	Populations	Individuals
	Total	**0.09**	0.10	0.26
57 individuals	Regions		**0.01**	0.18
	Populations			**0.18**

L. filosa		Regions	Populations	Individuals
	Total	**0.40**	0.41	0.47
56 individuals	Regions		**0.01**	0.10
	Populations			**0.09**

Approximate Bayesian computation analysis

The proportions of accepted datasets simulated under the null and full models were 59% and 41%, respectively. These were the proportions of the 1% simulated datasets closest to the observed data. The null model was the better one, having a much higher posterior probability (PP = 0.8) than the full model (PP = 0.2). This means that, given the observed data and the proportions of accepted datasets simulated under each model, the null model was four times more probable than the full model. Thus, the null model was chosen for parameter estimation (Table IV.5), although the posterior probabilities of all parameters of both the null and the full models (Fig. IV.11) are presented for comparison purposes.

Table IV.5 Parameter estimation under the null model, chosen for having received the highest posterior probability. Population size parameters (N) are given in number of individuals; time parameters (T) are given in generations (assuming one generation per year), and migration parameters (M) are given in number of migrants per generation, i.e. $4N_0m_{ij}$, where N_0 = 1e4.

				Credible interval 90%	
Parameter	Median	Mean	Mode	Lower	Upper
N1	6.4e5	6.5e5	5.9e5	4.5e5	8.8e5
N2	4.9e5	5.2e5	4.4e5	3.2e5	8.3e5
N3	5.5e5	5.5e5	5.2e5	4.4e5	6.8e5
N4	3.7e5	3.8e5	3.7e5	2.9e5	4.6e5
N5, N6 & N7	1.2e5	1.2e5	1.4e5	5.9e4	1.8e5
T2	9.6e5	9.9e5	8.8e5	7.2e5	1.4e6
T3	8.0e5	8.4e5	7.5e5	6.7e5	1.1e6
T4	1.5e6	1.5e6	1.5e6	1.2e6	1.9e6
M12 & M21	4.3e–2	9.2e–2	3.6e–2	9.7e–3	2.1e–1
M34 & M43	3.7e–3	4.8e–3	1.5e–3	9.0e–4	1.1e–2

All four current population sizes were estimated to be larger under the null model than under the full model (Fig. IV.11, N1–N4), and generally larger than the ancestral population sizes (Fig. IV.11, N5, N6 & N7). The unique parameter used to estimate ancestral population sizes did not vary much between the two models (Fig. IV.11, N5, N6 & N7). Under the null model, *L. cingulata* from allopatry had the largest population size (Fig. IV.11, N1, blue) and *L. filosa* from allopatry had the smallest population size (Fig. IV.11, N4, blue). Under the full model, *L. cingulata* in allopatry had also the largest current population size (Fig. IV.11, N1, orange); however, the smallest population sizes were found in the two sympatric populations (Fig. IV.11, N2 and N3, orange).

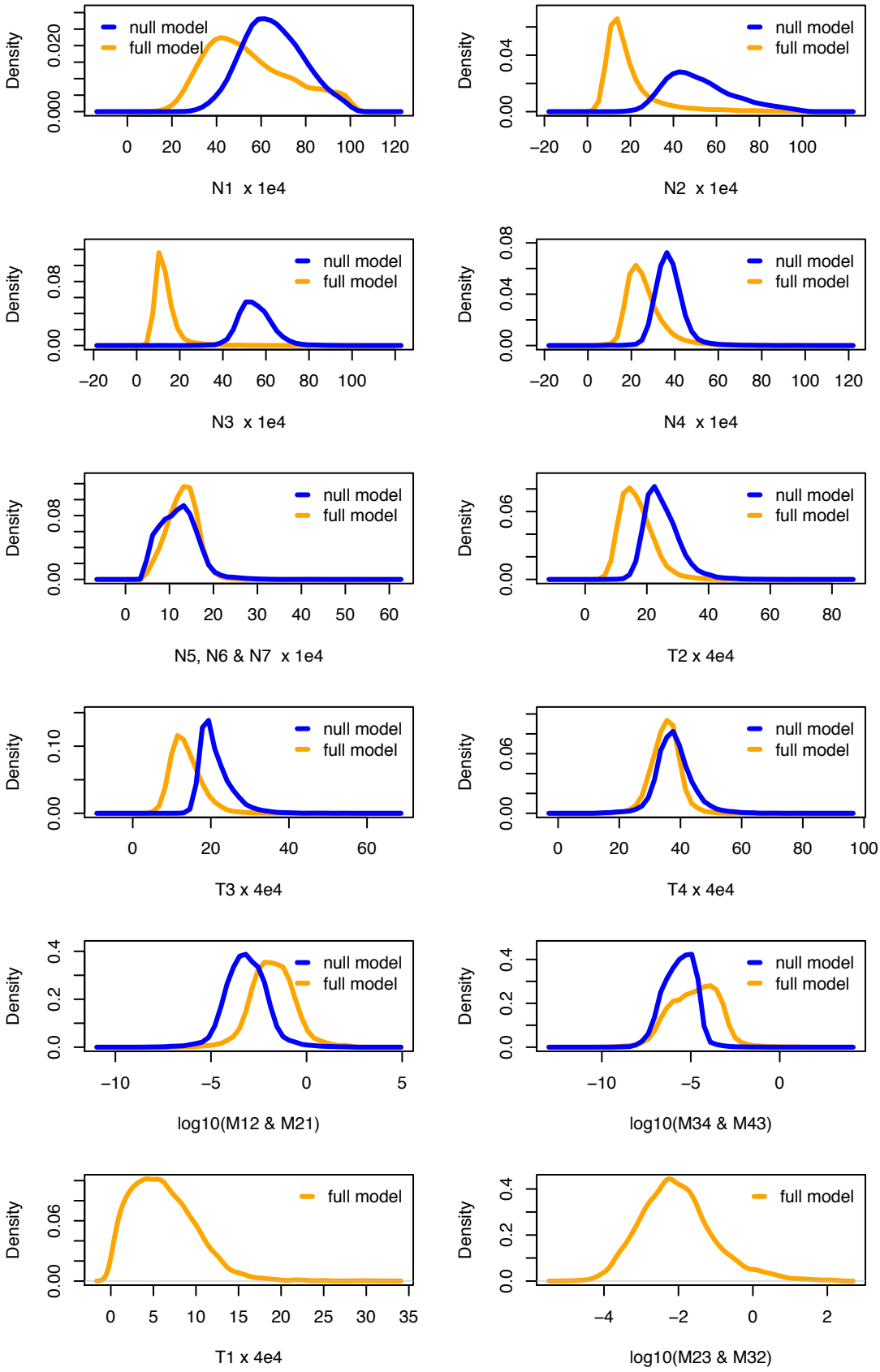


Fig. IV.11 Posterior probabilities of all parameters of both the null and full models. Population size parameters (N) are given in number of individuals; time parameters (T) are given in generations (assuming one generation per year), and migration parameters (M) are given in number of migrants per generation, i.e. $4N_0m_{ij}$, where N_0 = 1e4.

The two splits into the sympatric and allopatric populations for each species (T2 and T3) were estimated to be older under the null model (Fig. IV.11, T2 and T3). The estimation also revealed a slightly longer period since the split of *L. cingulata* (Fig. IV.11, T2) when compared with that of *L. filosa* (Fig. IV.11, T3). The null model showed that the time since the most ancient split (T4) was greater than that estimated by the full model (Fig. IV.11, T4) – a similar pattern as in T2 and T3, but much less pronounced. The time since migration between the sympatric populations started was estimated to be approximately 2.5e5 years ago.

Intraspecific migration for the two species was estimated to be lower under the null model (Fig. IV.11, M12 & M21, and M34 & M43), that is in the absence of interspecific migration. Under the two models, there was considerably less intraspecific migration in *L. filosa* (Fig. IV.11, M34 & M43) than in *L. cingulata* (Fig. IV.11, M12 & M21). Levels of interspecific migration fell between the levels of intraspecific migration found within each species under the full model.

Since the number of migrants per generation depends on the size of the receiving population, the comparison of such estimates is only logical within models due to different sizes of each population in the two models. In the null model, there was one migrant every 2 generations into allopatric *L. cingulata*, and one migrant every 3 generations into sympatric *L. cingulata*. Migration within *L. filosa* was lower, with the allopatric population receiving one migrant every 72 generations, and the sympatric population receiving one migrant every 51 generations. In the full model, there was one migrant per generation into allopatric *L. cingulata*, and one migrant every 5 generations into sympatric *L. cingulata*. Migration within *L. filosa* was also lower in this model, with the allopatric population receiving one migrant every 24 generations, and the sympatric population receiving one migrant every 46 generations. As for the interspecific migration between sympatric populations, there was one migrant every 6 generations into *L. cingulata* and one migrant every 7 generations into *L. filosa*.

Discussion

The results of the PCA were in agreement with the expectation considering the phylogeny (Reid *et al.* 2010, 2012) and geographic areas sampled. The clustering pattern of *L. filosa* into two groups fits with Reid's (1986, 2001) morphological descriptions of shell (sculpture and colour) and penis differences, distinguishing between western and eastern forms, with an intermediate shell morphology in the area of Cape Leveque, Western Australia, which lies between the two regions where *L. filosa* was sampled for this study. This finding is also consistent with the genetic structure found within *L. filosa* based on allozymic loci and mitochondrial genes (Inness-Campbell *et al.* 2003; Stuckey 2003; Reid *et al.* 2010) distinguishing between western and eastern samples.

The clustering pattern of *L. cingulata* into two groups corresponds to the subspecies designation *L. cingulata cingulata* and *L. cingulata pristissini* (Reid 1986), which are considered to be the same species based on genital anatomy (Reid 1986) and genetic data (Johnson & Black 1998; Reid *et al.* 2010), but described as subspecies due to differences in shell sculpture (Reid 1986) and a distribution gap of > 300 km between them (Johnson & Black 1998). The population structure observed in this study supports previous suggestions that the gap between the two subspecies of *L. cingulata* could favour the divergence between them (Johnson & Black 1998).

Another interesting aspect of the results of the PCA that is supported by previous work is the genetic distance observed between allopatric and sympatric regions in both species relative to each other. Despite previous findings in which the intraspecific Nei's (1972) genetic identity for both species was as high as > 0.98 (Johnson & Black 1998; Inness-Campbell *et al.* 2003), further investigation by Reid *et al*. (2010), computing Kimura two-parameter distances for COI sequences, found that allopatric and sympatric samples are separated by a greater distance in *L. filosa* (3.00) than in *L. cingulata* (0.61). This is also reflected in the most recent phylogeny of the genus that includes allopatric and sympatric samples of both species, where longer branches are observed between allopatric and sympatric *L. filosa* than between the corresponding comparison in *L. cingulata* (Reid *et al.* 2010).

After removing loci significantly departing from HWE in any one region from the whole dataset, a deficiency of heterozygotes (high He–Ho) at some loci was still expected in the combined dataset partially, but not entirely, due to the Wahlund effect (Wahlund 1928), i.e. differences in allele frequencies in different populations, even when Hardy-Weinberg proportions are kept within samples, or regions in this case. The Wahlund effect is usually observed when combining two or more different populations, and it is even more evident when including more than one species in the dataset (Hedrick 2011). However, the Wahlund effect is unlikely to explain the deficiency of heterozygotes in this study when the data were analysed separately by regions since the PCA showed no further structure at this level.

This apparent deficiency of heterozygotes within regions in both species is likely to be caused by some loci having large He–Ho differences that are not found to be significant in the HWE test. This would be partly overcome by using the weighted He–Ho, i.e. F_{IS} = (He–Ho)/He, instead of He–Ho. However, the positive values of $F_{individuals/populations}$, which are equivalent to F_{IS}, as computed in the hierarchical *F*-statistics analysis, within species (0.18 for *L. cingulata* and 0.09 for *L. filosa*) confirm the deficiency of heterozygotes. This remaining trend towards a lack of heterozygotes is most likely due to low sequencing coverage, which means that true heterozygotes could sometimes be miss-scored as homozygotes when the rarer allele at a particular locus has not been read enough times to be called as such; this is nevertheless unlikely to influence other analyses.

Both F_{ST} analysis and PCA suggested small signs of gene flow between the two species. Finding the lowest F_{ST} values (0.42–0.44) between the sympatric regions among all interspecific pairwise comparisons could suggest some gene flow between the overlapping populations of the two species, particularly from *L. filosa* into *L. cingulata*. Also, the fact that the PCA scatter plot revealed a shorter genetic distance between *L. cingulata* from the sympatric region and *L. filosa* compared to the one between *L. cingulata* from the allopatric region and *L. filosa* might also be a small sign of gene exchange. If this is true, due to the already mentioned asymmetry of the PCA scatter plot, it could be hypothesised that gene exchange happened from *L. filosa* to *L. cingulata* in sympatry.

Discerning among different biological explanations for this would need further investigation.

Even though the model without interspecific migration received higher posterior probability (contrary to the expectation), there are still several aspects that should be considered before ruling out either hypothesis. For instance, confidence in model choice should be calculated through a cross validation procedure, e.g. a false discovery rate control approach (Verhoeven *et al.* 2005). In this approach, a few simulated datasets are used as the observed data and the regression of the same simulated datasets is performed again. The frequency of the expected result is then used as an indication of the model choice robustness.

The parameter spaces of ancestral population sizes should be explored independently in order to represent more realistic scenarios. Regarding time parameters, the way in which the models were specified, the time since the most ancient split was firstly defined, and then all other times of the more recent demographic events were defined based on it. However, the model could also be built in reverse, i.e. defining a prior distribution for the time since the most recent demographic event (e.g. interspecific gene flow), and then constraining the older times around it. Getting similar outcomes from a pair of models differing only in this way would suggest that the order in which times are defined has no effect on the parameter estimation.

Time since the split into the two species was estimated to be younger (1.2–1.9 Mya) that previous estimates (2.6–3.8 Mya; Reid *et al.* 2010, 2012). The time estimate in this study coincides with the Pleistocene, a period which was characterised by repeated glacial cycles, which are known to have had an effect on the distribution of mangrove habitats due to changes in the sea level (Woodroffe & Grindrod 1991; Sun *et al.* 2000). Given than *L. cingulata* and *L. filosa* are obligate mangrove-dwellers, it is possible that changes in the distribution of mangroves during the Pleistocene associated with sea-level changes could have caused the divergence into these two species, as well as the intraspecific population structure (0.7–1.2 Mya), and even the interspecific secondary contact (0.25 Mya).

Regarding the ancestral population sizes, only general inferences can be made given that all three parameters used here were assumed to be equal. However, the estimates do suggest that ancestral populations were smaller than the current ones (null model) or nearly as big as the two sympatric populations (full model). These estimates of ancestral population sizes are reasonable when interpreted within the estimated timeframe of past demographic events. For instance, the reduction of suitable habitat in the past due to changes in sea levels may have forced the snails to retreat into refugia of mangrove forests, thus their population sizes would have been smaller than the current ones.

The fact that some parameters are assumed to be equal imposes constraints on the models built here. For instance, assuming symmetrical migration rates may represent a strong constraint on the models, as some marine currents are known to have direct impact on the dispersal of marine organisms with planktotrophic development (Wang *et al.* 2015). However, in the specific case of these mangrove snails, oceanic currents could also have an indirect impact on the patterns of gene flow by shaping the distribution of suitable habitat. In southeastern Asia and South America, it has been suggested that oceanic currents may act as barriers to gene flow in mangrove trees (Mori *et al.* 2015; Wee *et al.* 2015). Considering that the offspring of both *Littoraria* snails and mangrove trees disperse by water, it is possible that the patterns of gene flow in *L. cingulata* and *L. filosa* could be affected by oceanic currents too. Thus, relaxing the assumption of symmetrical gene flow in the models and allowing for asymmetrical migration rates should help clarify patterns of gene flow within and between *L. cingulata* and *L. filosa*.

It should also be considered that the interspecific migration clearly did not tend towards zero, nor did it approach the lower bound of the prior (1e–3), suggesting that gene flow is likely to have happened should this scenario be true. This result supports the hypothesis of interspecific gene flow – and so reinforcement – despite the model comparison result. This means that even though the full model did not receive the highest posterior probability, if there had actually been no interspecific gene flow at all, then much lower the levels of interspecific migration would have been expected here. The migration observed between the sympatric populations of *L. cingulata* and *L. filosa* was rather high, especially given that it was assumed to be constant from secondary contact

until the present time. However, this assumption might not be in agreement with the initially proposed hypothesis. In theory, if reinforcement contributed to speciation, then gene flow would have been at its maximum level upon secondary contact and gradually decreased towards the present. A model to test this hypothesis should incorporate such a decline. Alternatively, a less complex scenario could assume that interspecific gene flow both started and ceased in the past.

In summary, the population genetic analyses revealed lower levels of genetic differentiation between sympatric populations among all the interspecific comparisons. However, further studies should investigate whether this has been caused by recent gene flow between them or higher levels of genetic drift, and so more complete lineage sorting, experienced by the allopatric populations. Even though the model with no interspecific migration received more support in the ABC analysis, the model comparison should be validated and a wider variety of models, particularly allowing different patterns of gene flow, should be tested.

Acknowledgements

I am thankful to the following people and institutions for the help provided to carry out this research. Johan Hollander, David Reid and Roger Butlin designed the study and collected the samples. Johan Hollander obtained the funding from The Crafoord Foundation. The Molecular Ecology research group (University of Sheffield) provided lab support. Khalid Belkhir modified the coalescent sampler msnsam for the specific needs of this project. Ludovic Duvaux and Martin Hirsch set up the ABC pipeline and modified the summary statistic calculator mscalc. Ludovic Duvaux, Anja Westram and Roger Butlin provided insightful discussions throughout the development of this project. The Society for the Study of Evolution provided travel support to present this work at the Evolution conference, Brazil 2015. Finally, I would like to thank my supervisor Roger Butlin for his invitation to participate in this project in addition to my original Ph. D. project.

References

Abbott R, Albach D, Ansell S *et al.* (2013) Hybridization and speciation. *Journal of Evolutionary Biology*, **26**, 229–246.

Andolfatto P, Davison D, Erezyilmaz D *et al.* (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, **21**, 610–617.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Bewick ER, Dyer KA (2014) Reinforcement shapes clines in female mate discrimination in *Drosophila subquinaria*. *Evolution*, **68**, 3082–3094.

Beysard M, Krebs-Wheaton R, Heckel G (2015) Tracing reinforcement through asymmetrical partner preference in the European common vole *Microtus arvalis*. *BMC Evolutionary Biology*, **15**, 170.

Blair WF (1955) Mating call and stage of speciation in the *Microhyla olivacea–M. carolinensis* complex. *Evolution*, **9**, 469–480.

Boulesteix A-L, Strimmer K (2006) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, **8**, 32–44.

Bowman AW, Azzalini A (2003) Computational aspects of nonparametric smoothing with illustrations from the sm library. *Computational Statistics and Data Analysis*, **42**, 545–560.

Butlin RK (1987) Speciation by reinforcement. *Trends in Ecology and Evolution*, **2**, 8–13.

Butlin RK (1995) Reinforcement: an idea evolving. *Trends in Ecology and Evolution*, **10**, 432–434.

Butlin RK (2005) Recombination and speciation. *Molecular Ecology*, **14**, 2621–2635.

Butlin RK, Ritchie MG (2013) Pulling together or pulling apart: hybridization in theory and practice. *Journal of Evolutionary Biology*, **26**, 294–298.

Butlin RK, Saura M, Charrier G *et al.* (2014) Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. *Evolution*, **68**, 935–949.

Cooley JR, Simon C, Marshall DC, Slon K, Ehrhardt C (2001) Allochronic speciation, secondary contact, and reproductive character displacement in periodical cicadas (Hemiptera: *Magicicada* spp.): genetic, morphological, and behavioural evidence. *Molecular Ecology*, **10**, 661–671.

Coyne JA, Orr HA (1989) Patterns of speciation in *Drosophila*. *Evolution*, **43**, 362–381.

Coyne JA, Orr HA (1997) "Patterns of speciation in *Drosophila*" revisited. *Evolution*, **51**, 295–303.

Coyne JA, Orr HA (1998) The evolutionary genetics of speciation. *Philosophical Transactions of the Royal Society B*, **353**, 287–305.

Coyne JA, Orr HA (2004) *Speciation*. Sinauer Associates, Inc., Massachusetts.

Csilléry K, Blum MGB, Gaggiotti OE, François O (2010) Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution*, **25**, 410–418.

Csilléry K, François O, Blum MGB (2012) abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, **3**, 475–479.

Dobzhansky T (1937) *Genetics and the origin of species*. Columbia University Press, New York.

Dufour L (1844) Anatomie générale des diptères. *Annales des Sciences Naturelles*, **1**, 244–264.

Felsenstein J (1981) Skepticism towards Santa Rosalia , or why are there so few kinds of animals? *Evolution*, **35**, 124–138.

Fu Y-X, Li W-H (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

Goudet J (2005) HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, **5**, 184–186.

Hedrick PW (2011) *Genetics of populations*. Jones and Bartlett Publishers, Sudbury, Massachusetts.

Höbel G, Gerhardt HC (2003) Reproductive character displacement in the acoustic communication system of green tree frogs (*Hyla cinerea*). *Evolution*, **57**, 894–904.

Hollander J, Smadja CM, Butlin RK, Reid DG (2013) Genital divergence in sympatric sister snails. *Journal of Evolutionary Biology*, **26**, 210–215.

Hudson RR (2002) Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Hudson RR, Slatkint M, Maddison WP (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics*, **132**, 583–589.

Inness-Campbell J, Stuckey M, Johnson MS (2003) Allozymic investigation of phylogeny of *Littoraria* (Gastropoda: Littorinidae). *Journal of Molluscan Studies*, **69**, 19–26.

Jiggins CD, Naisbit RE, Coe RL, Mallet J (2001) Reproductive isolation caused by colour pattern mimicry. *Nature*, **411**, 302–305.

Johnson MS, Black R (1998) Effects of isolation by distance and geographical discontinuity on genetic subdivision of *Littoraria cingulata*. *Marine Biology*, **132**, 295–303.
Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, **27**, 3070–3071.
Kelly JK, Noor MAF (1996) Speciation by reinforcement: a model derived from studies of Drosophila. *Genetics*, **143**, 1485–1497.
Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. *Genetics*, **49**, 725–738.
Kirkpatrick M (2001) Reinforcement during ecological speciation. *Proceedings of the Royal Society B*, **268**, 1259–1263.
Lemmon AR, Kirkpatrick M (2006) Reinforcement and the genetics of hybrid incompatibilities. *Genetics*, **173**, 1145–1155.
Mayr E (1942) *Systematics and the origin of species: from the viewpoint of a zoologist*. Columbia University Press, New York.
de Meeûs T, Goudet J (2007) A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. *Infection, Genetics and Evolution*, **7**, 731–735.
Mevik B-H, Wehrens R (2007) The pls package: Principle Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, **18**, 1–23.
Mori GM, Zucchi MI, Souza AP (2015) Multiple-geographic-scale genetic structure of two mangrove tree species: the roles of mating system, hybridization, limited dispersal and extrinsic factors. *PLoS ONE*, **10**, e0118710.
Muller HJ (1942) Isolating mechanisms, evolution, and temperature. *Biology Symposium*, **6**, 71–125.
Nadachowska-Brzyska K, Burri R, Olason PI *et al.* (2013) Demographic divergence history of Pied Flycatcher and Collared Flycatcher inferred from whole-genome re-sequencing data. *PLoS Genetics*, **9**, e1003942.
Nei M (1972) Genetic distance between populations. *The American Naturalist*, **106**, 283–292.
Nei M, Kumar S (2000) *Molecular evolution and phylogenetics*. Oxford University Press, Oxford.
Nosil P, Crespi BJ, Sandoval CP (2003) Reproductive isolation driven by the combined effects of ecological adaptation and reinforcement. *Proceedings of the Royal Society B*, **270**, 1911–1918.
Ortiz-Barrientos D, Grealy A, Nosil P (2009) The genetics and ecology of reinforcement: implications for the evolution of prezygotic isolation in sympatry and beyond. *Annals of the New York Academy of Sciences*, **1168**, 156–182.
Ortiz-Barrientos D, Engelstädter J, Rieseberg LH (2016) Recombination rate evolution and the origin of species. *Trends in Ecology and Evolution*, **31**, 226–236.
de Queiroz K (2007) Species concepts and species delimitation. *Systematic Biology*, **56**, 879–886.
Ramos-Onsins SE, Rozas J (2002) Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution*, **19**, 2092–2100.
Reid DG (1986) *The littorinid molluscs of mangrove forests in the Indo-Pacific region: the genus* Littoraria. British Museum (Natural History), London.
Reid DG (1996) *Systematics and evolution of* Littorina. The Ray Society, London.
Reid DG (2001) New data on the taxonomy and distribution of the genus *Littoraria* Griffith and Pidgeon, 1834 (Gastropoda: Littorinidae) in Indo-West Pacific mangrove forests. *The Nautilus*, **115**, 115–139.
Reid DG, Dyal P, Williams ST (2010) Global diversification of mangrove fauna: a molecular phylogeny of *Littoraria* (Gastropoda: Littorinidae). *Molecular Phylogenetics and Evolution*, **55**, 185–201.
Reid DG, Dyal P, Williams ST (2012) A global molecular phylogeny of 147 periwinkle species (Gastropoda, Littorininae). *Zoologica Scripta*, **41**, 125–136.
Ross-Ibarra J, Wright SI, Foxe JP *et al.* (2008) Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE*, **3**, e2411.
Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, **145**, 1219–1228.
Roux C, Tsagkogeorga G, Bierne N, Galtier N (2013) Crossing the species barrier: genomic hotspots of introgression between two highly divergent Ciona intestinalis species. *Molecular Biology and Evolution*, **30**, 1574–1587.
Saetre G-P, Moum T, Bures S *et al.* (1997) A sexually selected character displacement in flycatchers reinforces premating isolation. *Nature*, **387**, 589–592.
Servedio MR (2001) Beyond reinforcement: the evolution of premating isolation by direct selection on preferences and postmating, prezygotic incompatibilities. *Evolution*, **55**, 1909–1920.

Servedio MR, Noor M (2003) The role of reinforcement in speciation: theory and data. *Annual Review of Ecology, Evolution, and Systematics*, **34**, 339–364.
Smadja CM, Loire E, Caminade P *et al.* (2015) Seeking signatures of reinforcement at the genetic level: a hitchhiking mapping and candidate gene approach in the house mouse. *Molecular Ecology*, **24**, 4222–4237.
Stuckey M (2003) Genetic divergence along Australia's tropical coastline, as seen in littorine snails: effects of geological history and ocean currents. Ph.D. Thesis, University of Western Australia.
Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
Trickett AJ, Butlin RK (1994) Recombination suppressors and the evolution of new species. *Heredity*, **73**, 339–345.
Turney CSM, Haberle S, Fink D *et al.* (2006) Integration of ice-core, marine and terrestrial records for the Australian Last Glacial Maximum and Termination: a contribution from the OZ INTIMATE group. *Journal of Quaternary Science*, **21**, 751–761.
Verhoeven KJF, Simonsen KL, McIntyre LM (2005) Implementing false discovery rate control: increasing your power. *Oikos*, **108**, 643–647.
Wahlund S (1928) Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas*, **11**, 65–106.
Wang J, Tsang LM, Dong Y-W (2015) Causations of phylogeographic barrier of some rocky shore species along the Chinese coastline. *BMC Evolutionary Biology*, **15**, 114.
Wang RJ, White MA, Payseur BA (2015) The pace of hybrid incompatibility evolution in house mice. *Genetics*, **201**, 229–242.
Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.
Wee AKS, Takayama K, Chua JL *et al.* (2015) Genetic differentiation and phylogeography of partially sympatric species complex *Rhizophora mucronata* Lam. and *R. stylosa* Griff. using SSR markers. *BMC Evolutionary Biology*, **15**, 57.
Wilding CS, Butlin RK, Grahame J (2001) Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *Journal of Evolutionary Biology*, **14**, 611–619.

Concluding remarks

The present study characterised the morphological and genetic variation of species in the *Littorina saxatilis* complex from previously unexamined sites. The results indicate that shell shape is associated with environmental variables and that certain ecotypes of *L. saxatilis*, namely *neglecta* and *tenebrosa*, have independent origins in different regions. This was also supported by the genetic data, which indicated that individuals from the same regions were more related to each other regardless of ecotype identity. The genetic data also confirmed previous reports of complex phylogeographic structure in the area around the British Isles.

The morphological and molecular data generated for *L. saxatilis* can be used for genome-wide association studies. The data can also be used for identifying putative loci affected by natural selection, which is especially valuable in a dataset with a known phylogeographic structure. This provides an advantage for speciation studies because the parallel divergence could be more readily tested once the phylogeographic history of the species is known. The molecular data set for *L. saxatilis* also constitutes a source of information to model and test specific colonisation history scenarios of the British Isles using an approximate Bayesian computation approach.

The study on the two sister species of *Littoraria* revealed no gene flow and therefore the case for reinforcement could not be made. However, neither hypothesis can be ruled out until further investigations into this matter. For instance, validation of the results should be performed. Also, additional models that allow independent estimations of gene flow and ancestral population sizes are needed.

In conclusion, the present study emphasises the importance of sea snails in the subfamily Littorininae as model systems to address a wide range of evolutionary questions. It also provides a wealth of data for many potential follow-up studies.

Acknowledgements

I am most thankful to my supervisor Roger Butlin for his constant support, guidance, patience, inspiration, and motivation over the whole duration of my Ph.D. His guidance always helped me in the research and writing of this thesis. His comments on many previous versions of this work greatly improved its quality. However, errors that remain in this thesis are entirely my own.

I also express my deepest gratitude to Anja Westram for her precious support, insightful comments and detailed explanations of difficult matters.

My sincere thanks also go to Johan Hollander, who provided me an opportunity to collaborate on a very exciting project.

I am grateful to Ludovic Duvaux for his patience, advice and encouragement to become more autodidactic.

I thank my examiners Terry Burke and Angus Davison for agreeing to assess my thesis and for providing very useful comments on an earlier version of it.

I also thank the members of my research team for stimulating discussions.

I also thank both Mexican institutions Consejo Nacional de Ciencia y Tecnología (CONACYT) and Secretaría de Educación Pública (SEP) for the funding provided towards my tuition fees and living expenses.

Last but not least, I would like to thank my family for their unconditional support not only throughout this challenging task that I chose to undertake, but also all the way through my life in general.