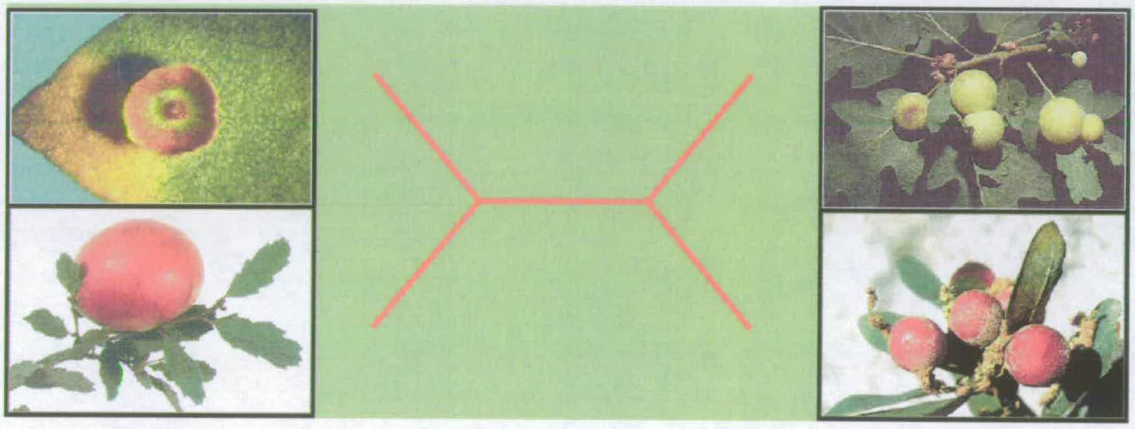
Tree-Thinking, Molecules and Gallwasps; Analysing Evolutionary Patterns in European Gallwasps Using a Molecular Phylogenetic Approach

Antonis Rokas



A thesis submitted to The University of Edinburgh in application for the degree of Doctor of Philosophy

November 2001

PREFACE

This thesis has been composed by me and is the result of my own work. Work done in collaboration is explicitly stated at the end of the appropriate chapters. The text does not exceed 100, 000 words. No part of this thesis has been or is currently being submitted to any other University in application for a higher degree.

ABSTRACT

In this study I have employed a suite of molecular markers and analysed them in a statistical phylogenetic framework to address patterns of evolutionary diversity found in a variety of taxonomic levels within the family of gallwasps (Hymenoptera: Cynipidae).

Gallwasps are a group of phytophagous parasitic wasps that induce gall formation in a variety of herb and tree species. They are distributed throughout the Holarctic and palaeontological evidence suggests that the family must have originated in the late Cretaceous, some 83 million years ago. A variety of lifecycle and ecological characteristics of gallwasps, such as presence of different reproductive modes, obligate association with their plant hosts, and wide geographical distribution pose interesting questions regarding their natural *history* that can be addressed by the combined use of molecular data and phylogenetics.

Selection of markers for molecular phylogenetic analysis requires matching the level of polymorphism to the desired taxonomic level of study. I have analysed data from nuclear and mitochondrial DNA sequence data in gallwasps and reviewed rare large-scale mutational changes in animals, in an attempt to understand their utility for phylogenetic analysis. Analysis of the sequence data led to the identification of fast and slow-evolving loci for insect phylogenetics. Review of published data on rare large-scale mutations (or Rare Genomic Changes – RGCs for short) suggested that RGCs are likely to be useful and low-homoplasy phylogenetic markers.

Using DNA sequence and allozyme data, I studied the phylogeography and post-Pleistocene range expansion of two widely distributed European oak gallwasp species, *Andricus quercustozae* and *Biorhiza pallida*. Patterns of genetic diversity in these species have been shaped by their association with their obligate hosts, the oaks (both species), and by the bacterial endosymbiont *Wolbachia* (*B. pallida*). These studies revealed the existence of multiple distinct European and Anatolian refuges and suggested that the species' post-glacial distribution range has been determined by the presence / absence of specific oak species in central and northern Europe.

Using mitochondrial DNA sequence data, I generated a phylogeny of the various European gallwasp genera, verifying the existence of various conserved

clades and suggesting that many species are non-monophyletic due to lineage sorting of ancestral polymorphisms and / or hybridisation. For a subset of this phylogeny – the genus *Andricus* – the evolution of shift(s) in the sexual generation oak hosts and host organs galled were analysed, using additional nuclear data.

Finally, I surveyed member species of most gallwasp tribes for infection with the bacterial endosymbiont *Wolbachia* and showed that: (i) the prevalence of *Wolbachia* differs between lineages with different reproductive modes; (ii) *Wolbachia* is not associated with thelytoky in the cyclical parthenogenetic gallwasp lineage (in contrast, thelytoky in herb and rose gallwasps is, most likely, *Wolbachia*-induced); and (iii) horizontal transfer of *Wolbachia* is likely to have occurred between gall-inducers and their associated inquilines.

ACKNOWLEDGEMENTS

Four years ago, when I first came across the word 'gall' I had to look into the dictionary to find its meaning. It is therefore not surprising that work described in this thesis would not have been produced, had I not been fortunate enough to have the help and collaboration of numerous people.

I would like to thank my supervisor Graham Stone for not only introducing me to gallwasps but also for being generous in letting me find and pursue my interests within the field. His great enthusiasm, knowledge and support for my work and well-being were very much appreciated.

The art of molecular techniques and the fascination for molecular evolution were acquired during my stay in Peter Holland' s lab in Reading, prior to the start of this thesis. I would like to thank him not only for accepting me in his lab, but also for his help after I had left (Ch. 2).

I would like to thank my second supervisor, Josephine Pemberton, for her encouragement, support, provision of full access in her lab as well as the chance to teach in 'Evolution in Action', the second year undergraduate course she is organising. I also wish to thank Richard Ennos for offering me the chance to teach in the undergraduate course 'Evolutionary and Ecological Genetics'. Both courses clarified endless things about molecular evolution (at least to my mind).

A number of collaborators offered their expertise and valuable help; these are Yoshihisa Abe (Ch. 8), James Cook (Ch. 6, 7), George Melika (Ch. 6), Jose-Luis Nieves-Aldrey (Ch. 6, 8), Johan Nylander (Ch. 3), Mark Pagel (Ch. 7), Fredrik Ronquist (Ch. 3) and Stu West (Ch. 4, 8). My warmest thanks to all of you.

Past members of the gallwasp group, Rachel Atkinson, Gordon Brown and Lucy Webster offered valuable help and advice. Rachel was especially helpful in discussing ideas, sharing data (Ch. 4, 5) and commenting on papers, while Gordon offered repeatedly his help during my frustrations with computers and also shared his data (Ch. 4). Lucy tolerated my supervision for her Honours project and allowed me to use some of her data as well (Ch. 5).

I was fortunate enough to participate twice (as a student the first and as a teaching assistant the second) in the 'Workshop on Molecular Evolution' in Woods

Hole, MA and to perform a large part of the analyses in Chapters 3 and 6 there. I would like to thank the course organiser Mike Cummings, the faculty, students and teaching assistants for making the course such good fun and for sharing their thoughts on my endless questions on trees and phylogenies.

A number of other labs offered their kind help. People in the Rowe lab – Alex, Ahmed and Ian – were very sympathetic to my overuse of their PCR machine, whereas the Blaxter lab offered access to computers and interesting discussions. Jill Lovell provided excellent service of the automated DNA sequencer.

My 'uneventful' life in the lab was occasionally (and pleasantly) disrupted by field- and work-related trips. Gallwasps took me to Italy (with Rachel), Greece (with Graham) and Hungary (with Graham and Gordon), whereas the data took me twice to Sweden (once with Graham, Rachel and James) and once in Cardiff. Thanks to my co-travellers for offering the rides but most importantly to ours hosts in the final destination; in Hungary, Gyuri Csóka, his wonderful family and the small 'army' of Hungarian foresters; in Sweden, Fredrik Ronquist and Johan Nylander; In Cardiff, Mark Jervis.

The department and even more so its Annexe, provided a friendly environment. Thanks to all those people who tolerated my mercurial moods, especially when sharing a lab or an office: Alex, Andy, Ashley, Becky, Boo, Culum, Dave, Dave, Felicity, Gordon, John, Kevin, Lucy, Martin, Nick, Rachel, Richard, Simon, Steve, Sue (I am sure I am forgetting some) and all those who joined Annexe coffee-time. Additionally, participation in the notorious 'Sporting ICAPB' football squad was always a joy.

Finally, I would like to thank all the non-ICAPB friends for their help and support through these years with special mention to Tracy, Nikiforos, Eva and Timm.

This Ph.D. was partially supported by a studentship from the Natural Environment Research Council. The James Rennie Bequest provided additional travel funding. However, I would never have been able to fulfil this work without the unending support of my family, and especially of my parents Kostas and Maria. This thesis is dedicated to them.

CONTENTS

CHAPTER 1

An introduction to 'tree-thinking', molecular phylogenetics and gallwasps

'As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications'

Charles Darwin (1859)

The tree is one of the most powerful images, not only in human thought and worship as a symbol of fertility, life and destiny among others, but also in evolutionary thought, as a pictorial representation of the historical associations among present and past organisms during the course of evolution (Hestmark, 2000). The idea of representing the course of history as a tree is not recent (Fig. 1.1A), although trees as today' s evolutionary biologists know and use them – mathematical structures describing the evolutionary associations of a group of taxa – find their root, as usual, in Darwin' s masterpiece (1859) (Fig. 1.1). Darwin was the first to draw a tree and include it as the only figure in the *Origin of Species* (Fig. 1.1B), commenting 'I believe this simile largely speaks the truth'; the word phylogeny was coined soon after (in 1866) by Ernst Haeckel and it was in his drawings that phylogenetic trees reached their artistic peak (Fig. 1.1C). From the very first days of evolutionary biology, phylogenetic trees have been influential for the acceptance and establishment of evolutionary theory.

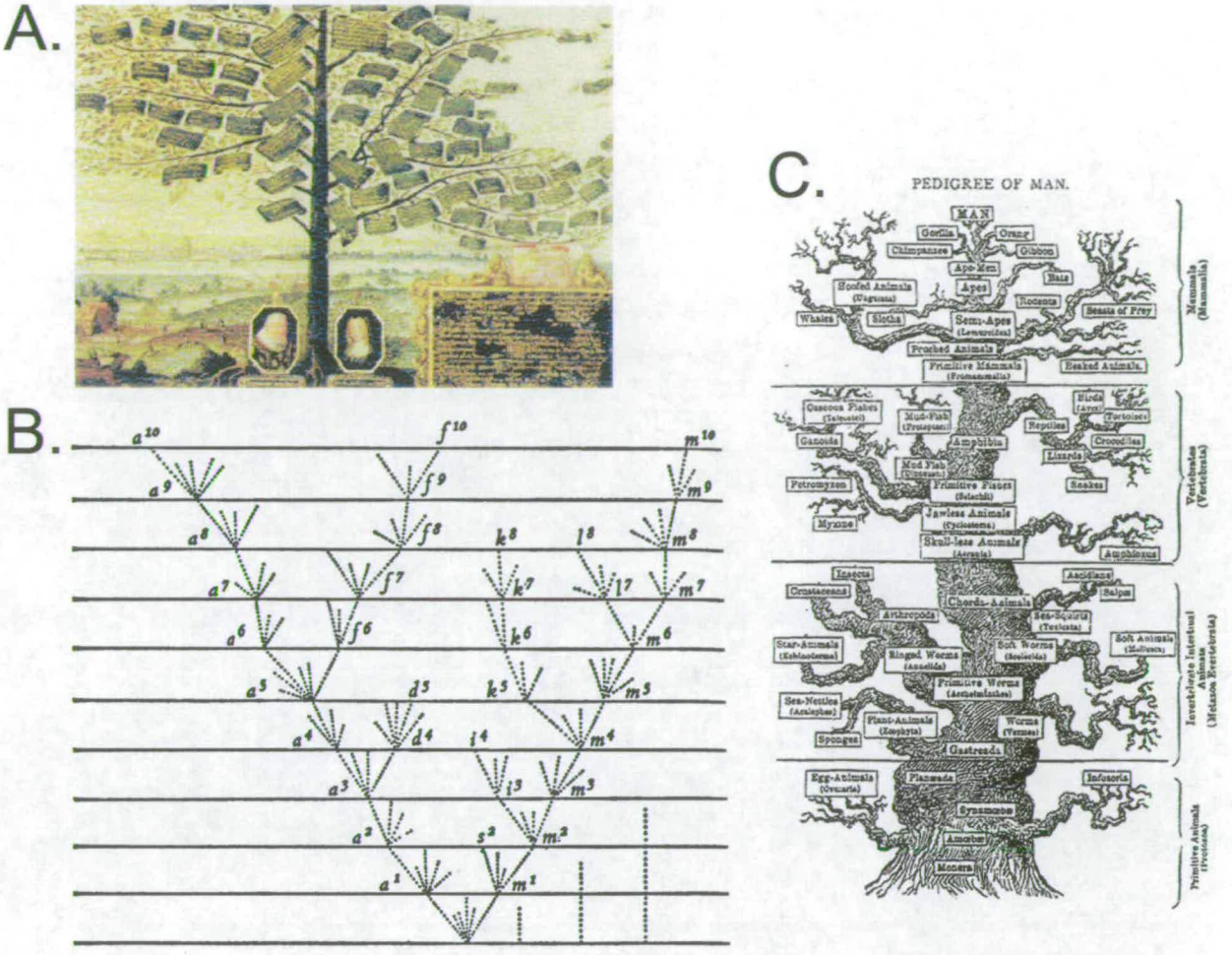


Figure 1.1. Phylogenetic trees through time. A. Genealogy from 17th-century Germany (from Hestmark, 2000). B. Part of the tree included in the *Origin* (from Darwin, 1859). C. Haeckel' s highly artistic phylogenetic tree of life. D. An example of a modern tree; the phylogeny of dinosaurs (from Sereno, P. C. 1999, Science, 284: 2137-47).

Two are the major themes of this thesis; the first is the attempt to identify molecular markers which are useful for phylogenetics (Chapters 2, 3). A plethora of articles have been devoted to comparison of methods of phylogeny reconstruction (e.g. Swofford, *et al.*, 1996; Kitching, *et al.*, 1998; Lewis, 1998b), with much less space being devoted on the usefulness of the markers themselves. The revolution in genomics and the progress in molecular biology techniques have initiated a surge of work focusing on the importance of the markers used for phylogenetic reconstruction, rather than on the methods. The second theme of this thesis is the application of 'tree-thinking' to address a variety of evolutionary questions in European oak gallwasps (Hymenoptera: Cynipidae: Cynipini) (Chapters 4 – 8), sometimes in relation with the bacterial endosymbiont *Wolbachia* (Chapters 4, 8).

In this Chapter I wish to introduce and review concepts and methods which I employed in this work to understand and explain patterns of the natural *history* of gallwasps; the concept of 'tree-thinking', molecular markers available for phylogenetics and methods of phylogenetic analysis. Finally, I will review aspects of the biology of the study system, gallwasps, which are relevant to the work presented in this thesis.

1.1 The concept of 'tree-thinking'

The shift from considering within-species variation as *random error* causing deviation from the 'true type' (essentialist or typological thinking), to understanding its reality (population thinking) revolutionised population genetics and evolutionary biology in first half of the last century (Mayr, 1982). A similar situation may be found in the phylogenetic literature. Not so long ago, species or taxa were considered as independent entities within a class ('group-thinking') (O'Hara, 1997) and even earlier as intermediate steps in the 'ontogeny' of evolutionary history going from (the simple) monad to (the complex) man ('developmental thinking') (O'Hara, 1997). References still found in articles regarding 'higher' and 'lower' taxa may be considered as a remnant of these years (Mogie, 2000). The realisation that species, and more generally taxa, are not independent replicates within a class as earlier thinkers tended to see them, but are instead interconnected parts of an evolutionary

tree has been termed 'tree-thinking' (O'Hara, 1988, 1997). Drawing parallels with the spread of 'population thinking', O'Hara has suggested that 'tree-thinking' spread as a new mode of thinking in phylogenetic systematics during the second half of the last century with the development of rigorous phylogenetic philosophies and had a similar effect to 'population thinking' on evolutionary studies.

The ramifications of the application of 'tree thinking' to evolutionary questions go beyond simple description of the historical relationships amongst a group of organisms and extend to inference of the evolutionary processes that have generated and shaped the patterns and trends we observe today (e.g. Harvey and Pagel, 1991; Pagel, 1997, 1999a).

1.2 Why molecules?

The 'molecules versus morphology' debate reverberated through the field of phylogenetics during the 1980s (e.g. Patterson, 1987) and although conflicts between hypotheses proposed from the two types of data still exist (e.g. Gatesy and O'Leary, 2001), the realisation that both types of data have their own merits has put aside arguments about superiority of either. Practical issues aside, there are a number of advantages in choosing molecules as markers for phylogenetic analysis (Hillis and Wiens, 2000; Nei and Kumar, 2000). These include the much larger amount of information contained in molecules, the much more tractable modelling of their evolution, and the existence of models able to assess the amount of data needed to resolve particular relationships (Lecointre, *et al.*, 1994; Philippe, *et al.*, 1994a). Additionally, comparisons between any living organisms can be made at the molecular level, be it bacteria and elephants or algae and plants, in other words between organisms with no morphological similarities. This may be accomplished because the different loci found in genomes exhibit a wide range of rates of evolution making them applicable to a very broad range of taxonomic levels.

For molecules, their genetic basis is, by definition, known, whereas for morphological characters it is assumed (the former are *genotypic* characters, while the latter are *phenotypic*) and furthermore, we currently know very little about morphological characters' underlying genetic structure and evolution, so weighing of

the importance of different morphological characters is not an easy task. For example, comparative developmental work on complex features such as abdominal and wing pigmentation and wing venation in Drosophilids has shown that their underlying genetic control is relatively simple (e.g. True, *et al.*, 1999; Kopp, *et al.*, 2000). Therefore, patterns seen in different taxa and formerly believed, under the assumption that their evolutionary emergence occurred in multiple steps, to have evolved only once might consequently have evolved independently relatively easily.

Delimitation of molecular characters is much more straightforward and objective than for morphological characters. A good example is offered by the complex eyespot patterns seen on the wings of butterflies (Brakefield, 2001). In this case the term 'character' can be applied to a single feature of the eyespot, a single eyespot, a subset of eyespots or the complete pattern. It is only after years of comparative genetical, developmental and morphological work (Brakefield, 2001, and references therein) that researchers have been able to show that the most appropriate use of the term 'character' is for the complete eyespot pattern. The ability to correctly identify characters that are independent and homologous (rather than homoplasious) can profoundly bias the result of the analysis (Poe and Wiens, 2000) and cases in which morphological characters are functionally or developmentally correlated are far from rare (Emerson and Hastings, 1998).

Despite these problems, morphological characters are still very useful since they allow wider taxonomic sampling and can be applied to fossil data. Furthermore, in almost all groups of organisms, species are identified and classified on the basis of morphology, and consequently molecular systematists rely on morphology to determine their sampling of species (Hillis and Wiens, 2000). Additionally, molecular data are not a panacea since they have their own problems (see below). It is perhaps an understatement to say that the use of molecular data to infer phylogenetic relationships will not displace morphology and embryology from the study of the evolutionary history of living (and extinct) organisms: after all, it is the history of morphological change that we wish to explain (Raff, *et al.*, 1989).

1.3 Characters in molecular phylogenetics

The origin of molecular phylogenetics can be traced in the start of the last century. For example, Nuttall in the beginning of last century (1904), conducted serological cross-reaction studies in primates and on the basis of their affinity constructed primate phylogenies. Another forgotten pioneer was the geneticist A. H. Sturtevant, one of the most influential students that emerged from T. H. Morgan' s famous 'fly room'. Alone and in collaboration, he produced a series of papers attempting to reconstruct the phylogenetic history of Drosophilids using chromosomal inversions (Sturtevant and Dobzhansky, 1936; Sturtevant and Tan, 1937; Sturtevant, 1942), a molecular marker which he discovered (Sturtevant, 1921) and which is still in use today (reviewed in Krimbas and Powell, 1999).

Despite the pioneering research described above, the emergence of molecular phylogenetics as an autonomous field had to wait for the molecular revolution that took place during the 1950s. For the students of molecular evolution and phylogenetics perhaps the most important first step was the discovery of protein sequencing by Fred Sanger, followed by the realisation from Harris and Lewontin and Hubby that electrophoresis of allozyme loci could be employed to address population genetics questions. These discoveries were to dominate experimental research on molecular evolution in the 1960s and 1970s. In the late 1970s, Sanger, and Maxam and Gilbert independently invented protocols for the sequencing of DNA and at the same period the first restriction enzyme variation studies (they later became known as RFLPs) on mitochondrial DNA appeared. The ability to read the information encoded in the genetic material coupled with the discoveries of Polymerase Chain Reaction by Mullis and co-workers and microsatellite sequence repeats by Jeffreys and colleagues in the mid-1980s and their quick application to address evolutionary and ecological questions, gave an unprecedented boost to the field. It was only a natural consequence of these discoveries in molecular biology that a few years later the sequences of the first complete genomes appeared.

Algorithms and models of evolution for phylogeny reconstruction exist for most molecular characters, including protein, DNA and RNA sequences, allozymes and microsatellites (e.g. Swofford, *et al.*, 1996; Hillis, 1998; Thorne, 2000).

However, the increasing technical easiness and automation of DNA sequencing on the one hand and the advances made in the mathematical modelling of its sequence evolution on the other have made DNA sequence data the most popular marker for molecular phylogenetics.

A number of other molecular markers also exist and have been used in molecular phylogenetics, especially for higher taxonomic levels; these include intron indels, gene order rearrangements, gene duplications, retroposon integrations, genetic code changes, signature sequences, genomic signatures, chromosome banding and presence/absence of genes in genomes among others. In Chapter 2 I present a review of these markers, which I collectively call *rare genomic changes* (RGCs) and suggest that they are valuable molecular markers for higher level phylogenetics (Rokas and Holland, 2000).

Given that the majority of this study deals with phylogenetic analysis of DNA sequence data, below I will highlight some of the technical problems associated with their use in phylogenetics as well as briefly review the methods that exist for their analysis.

1.3.1 Problems associated with DNA sequencing

In the majority of phylogenetic studies which employ DNA sequence data as the marker of choice, the standard protocol includes amplification of the desired region by PCR using degenerate primers, followed by isolation of the amplified fragment and subsequent sequencing of it. Most of the technical problems / artefacts associated with this protocol arise from problems in the PCR step. What follows is a list of such problems that researchers in molecular phylogenetics should know, and occasionally worry, about.

A number of problems are associated with the use of degenerate primers, which frequently lead to amplification of sequences related to the target one. One such example is the amplification of pseudogenes and paralogous loci (Perna and Kocher, 1996; Zhang and Hewitt, 1996a, b; Telford and Holland, 1997; Bensasson, *et al.*, 2000; Mirol, *et al.*, 2000; Bensasson, *et al.*, 2001). In most reported cases, the identification and exclusion of pseudogenes is straightforward due to their mutational

degeneration (presence of indels and stop codons) resulting from loss of selective pressure. However, in an increasing number of cases pseudogenes have been isolated which bear close resemblance to the 'true' copy, making identification of the latter difficult without resort to more elaborate molecular techniques (Chapter 5). An example from oaks (genus *Quercus*), where two different and incongruent phylogenetic hypotheses have been proposed (Samuel, *et al.*, 1998; Manos, *et al.*, 1999), is illuminating in this respect. Mayol and Roselló (2001), after re-analysis of both datasets, have suggested that the incongruence between the two phylogenies is due to inclusion, in one of the studies, of nonfunctional paralogous loci. Given that the locus used was the internal transcriber region (ITS) of the ribosomal RNA array, the identification of non-functional pseudogenes is not as straightforward as for protein-coding loci. Pseudogenes resembling mitochondrial loci and which are probably located in the nucleus have also been found in gallwasps (Chapters 5, 6 , Appendix I, Stone, *et al.*, 2001)(see also http://www.pseudogene.net).

Wagner *et al.* (1994) have argued that amplification of pseudogenes or paralogous copies may be accomplished through two major processes, PCR selection and PCR drift. PCR selection occurs when the reaction preferentially amplifies certain paralogues / pseudogenes and the result is replicable. A potentially major contributor to PCR selection is differential primer affinity due to differences in primary or secondary structure of DNA at potential target sites. An example from the literature is Buckler' s *et al.*(1997) report of differential PCR amplification of ITS paralogues. In contrast, PCR drift is the result of random events occurring in the early cycles of the reaction, and the bias will not be repeatable. A sufficient number of examples exist (Buckler IV, *et al.*, 1997; Telford and Holland, 1997; Bensasson, *et al.*, 2000; Mayol and Rossello, 2001) to suggest that molecular phylogenetic studies are not immune to either of these processes.

One of the most common PCR-related problems is contamination. Although standard laboratory protocols and precautions may guard against exogenous sources of DNA contamination, they are not effective at excluding endogenous sources of DNA, such as from symbiotic microbes. The significance of symbionts and their potential for introducing error in molecular studies is not generally recognized, is rarely accounted for and has not been generally tested.

Incomplete chain extension during PCR may result in the generation of recombinant products (Pääbo, *et al.*, 1990; Bradley and Hillis, 1997). Although the majority of such cases has been observed in connection with large amount of damaged DNA, especially obtained from ancient samples (Handt, *et al.*, 1994), recombination during amplification of multi-gene families (Scharf, *et al.*, 1988) or from heterozygous individuals from a single locus (Bradley and Hillis, 1997) have also been observed. It is unknown how common this phenomenon is.

Another source of problems are the nucleotide misincorporations which occasionally occur leading to amplification errors in the synthesized DNA, and which have been shown to be dependent on the specific DNA polymerase used (Cline, *et al.*, 1996). Additionally, PCR artefacts such as deletions have been reported (Cariello, *et al.*, 1991), probably associated with the formation of intra-strand hairpins (e.g. Mayol and Rossello, 2001).

1.4 Methods of analysis in molecular phylogenetics

As mentioned earlier, since most of the work described here and the majority of published articles utilise DNA sequence data, only methods developed for their analysis will be discussed.

The standard recipe for the reconstruction of a phylogenetic tree requires two ingredients: an *optimality* criterion which allows assessment of the *optimal* (*best*) tree between alternatives and a *search strategy* which specifies the way the space of possible trees is explored in our quest for the *optimal* tree. The latter is necessary for the simple reason that even for a small number of taxa, the number of alternative topologies is very large, making the exhaustive search of 'tree space' an impossibility for large datasets.

Optimality criteria can be further classified by the type of data they operate on; pairwise distance datasets or discrete character datasets (conversion from the latter to the former is possible but not vice versa). The two optimality criteria that operate on pairwise distances are minimum evolution (ME) and least squares (LS). The LS criterion was proposed by Fitch and Margoliash (1967) to select the tree that minimises the sum of squares, in other words the tree with the best-fitting branch

lengths. The ME criterion was proposed by Cavalli-Sforza and Edwards (1967) and its principle is to select the topology that minimises the total branch length.

There are three optimality criteria that operate on discrete characters, parsimony (P – sometimes, incorrectly, referred to as maximum parsimony), maximum likelihood (ML) and Bayesian inference (BI). The philosophy of parsimony goes back to the philosopher William of Ockham and his rule, which has been known as Ockham' s razor: it is vain to postulate more causes when fewer suffice to explain (Sober, 1988). For phylogenetic purposes, the criterion of parsimony was first advocated by Hennig (1966) and suggests that the topology minimising the number of required character state changes is the best. Likelihood as a statistical method was invented by the great statistician and theoretical biologist R. A. Fisher in 1912 and by 1922 improved to its current status (the history of likelihood is reviewed in Appendix 2 of Edwards, 1992). Given some data D and a hypothesis H, the likelihood of obtaining the data D *given* the hypothesis H is given by $L_D = Pr (D \mid H)$. The likelihood of a hypothesis should not be confused with its probability, the latter being Pr (H | D). For phylogenetic purposes, ML was first used in 1964 by Edwards and Cavalli-Sforza (, 1964) on gene frequency data, but it was not until 1981 that Felsenstein described a likelihood method for DNA sequences (Felsenstein, 1981). ML' s principle is that the topology of choice is the one that maximises the probability of the observed data. Bayesian inference uses a slightly modified (and somewhat controversial in statistical circles) definition of probability from the standard one, in that it allows (posterior) probability to be estimated from a combination of data and prior information / belief (Shoemaker, *et al.*, 1999). In contrast, standard probability allows estimation of support for a hypothesis only from the data; prior belief in non-informative. Although Bayes' rule was proposed more than 200 years ago by Reverend Thomas Bayes (Bayes and Price, 1763), it has only recently been applied in phylogenetics due to problems associated with the evaluation of posterior probabilities (e.g. Mau and Newton, 1997; Yang and Rannala, 1997; Larget and Simon, 1999; Huelsenbeck, *et al.*, 2000). BI 's principle in phylogenetics is to select the topology with the highest (posterior) probability.

Many alternative search strategies in tree space exist and choice between them is usually dictated by available computational power, rather than by the user' s

philosophical preferences, the latter being most often the reason for selection of optimality criteria. Search strategies may be classified by the type of strategy they follow to find the optimal tree into *algorithmic*, *heuristic* and *exact*. Search methods in the first category find the optimal tree by following a specified algorithm, with the characteristics of high speed and very limited searching of total tree space. Two well-known algorithmic search methods are *star decomposition* (from an initial star tree of all taxa, successive joining of the most closely related taxa occurs) and *stepwise addition* (addition of taxa in the topology is done in a stepwise fashion, starting with those most closely related). Heuristic search strategies are computationally much more intensive than algorithmic ones but search a larger part of the total tree space. This is achieved by re-arrangement of clades for each topology searched, a process also known as branch-swapping. A number of different branch-swapping strategies exist, usually known by their acronyms (Nearest Neighbour Interchange – NNI, Subtree Pruning and Regrafting – SPR, and Tree Bisection and Reconnection – TBR). Exact methods are the most computationally intensive of all three types of search strategy, but are also the only ones that search the whole tree space and guarantee to find the optimal tree.

1.4.1 Justification of a statistical and model-based framework for phylogenetic analysis

As outlined above, many optimality criteria are available for reconstructing DNA sequence phylogenies. The two features that make maximum likelihood and Bayesian inference the preferred methods for the majority of phylogenetic analysis conducted in this work (Chapters 3-7) is that they are *model-based* and *statistical*.

A statistical framework is necessary for almost every field of science (although few disciplines have had as much trouble in establishing it as had phylogenetics, Felsenstein, 2001). Maximum likelihood (ML) is the most well-known and frequently used method of statistical phylogenetic inference. Its primary advantage over other methods is that it allows incorporation of acquired knowledge about substitution processes and phylogenies in a non-arbitrary way, through rigorous hypothesis testing (Huelsenbeck and Rannala, 1997; Lewis, 1998b). Assumptions

like a constant rate of molecular evolution (molecular clock), rate heterogeneity among nucleotide sites, independence among nucleotide sites, whether two tree topologies are significantly different or which model of sequence evolution adequately describes the data at hand may be tested using ML.

Bayesian phylogenetic inference (BI) represents a very recent advance in phylogenetic theory; it is perhaps an indication of the rapid advancement in the field that the major computer packages offering Bayesian phylogenetic inference have appeared during the course of this work (BAMBE by Larget and Simon, 1999; MRBAYES by Huelsenbeck and Ronquist, 2001) or have yet to be published (MCMC by M. Pagel). BI offers the same advantages as maximum likelihood and in near future may prove superior to ML for two reasons: (a) support for clades and topologies is given in terms of posterior probabilities, which are easier to interpret than bootstraps, and perhaps more importantly (b) the algorithms currently employed in BI are significantly faster than existing ML algorithms. However, the robustness and consistency of BI have not been, as yet, well studied and comparisons of BI with other methods have yet to be made.

1.5 Extrapolating from gene trees to species trees

'*Although all of us have seen diagrams of an electron zooming around the nucleus of an atom like a discrete little satellite, physics now tells us that the electron is diffuse. It is not a matter of uncertainty about where the electron is. Rather, in a real sense the electron is in more than one place at once. Likewise, phylogenetic history is in more than one place at once; it is a composite of all the varied histories of all the genes, some of which might place species A next to species B, others might place A next to C, etc. Just as an electron can be depicted as a cloud, we might want to view phylogeny as a diffuse cloud of gene histories*'

Wayne Maddison (1997)

In molecular phylogenetics, the assumption that a gene tree will mirror the species tree is a very important one (e.g. Slowinski and Page, 1999; Edwards and Beerli, 2000). Lineage sorting of ancestral polymorphisms (deep coalescence), gene

duplication and horizontal transfer are the three processes which may cause a discordance between a gene phylogeny and a species phylogeny (assuming that estimation of the former is correct) (Maddison, 1997). Persistence of ancestral polymorphisms through speciation events may generate gene trees that are different from the species tree (Neigel and Avise, 1986; Maddison, 1997). Moore (1995) has shown that phylogenies based on mitochondrial loci (Chapter 6), and more generally phylogenies based on genes with low effective population size, will be less prone to distortion from lineage sorting, given that time for coalescence crucially depends on the effective population size of the ancestral population (Hoelzer, 1997; Moore, 1997). Given the dependence of the gene tree – species tree discordance on effective population size, lineage sorting will be more of an issue in phylogenies of closely related species than in phylogenies of distantly related taxa (see also Chapter 6). Problems arising from gene duplication and comparison of paralogous, instead of orthologous, loci are more difficult to assess. Unless the genome sequences for the taxa under study are known, it is impossible to be certain that our comparisons are between orthologous loci. Examples of lineage-specific gene duplications are generally known and common (e.g. Rokas, *et al.*, 1999, see Appendix I), including duplications in loci frequently used in phylogenetics (Telford and Holland, 1997; Danforth and Ji, 1998). Horizontal (or lateral) transfer of genes across species is more of a problem for prokaryote phylogenetics than it is for multicellular eukaryote phylogenetics, especially of those organisms with sequestered germ lines (Doolittle, 1999). However, one special case of horizontal transfer, that of introgression of genes between species that are hybridising is a real problem in animal populations (e.g. Moore, 1995, Chapter 6).

Another problem emerging from the gene tree – species tree distinction is the realisation that molecular and species polytomies or multifurcations (Slowinski, 2001) should not be considered the same. Most molecular polytomies are due to lack of data (and hence 'soft polytomies') rather than due to simultaneous divergence of three or more genes ('hard polytomies'); demonstration of species polytomies requires multiple unlinked gene trees (Slowinski, 2001).

1.6 Gallwasps (Hymenoptera: Cynipidae) as a study system

The early history of gallwasp research is intimately linked with the study of plant galls (*cecidology*) and may be traced back to ancient times. The majority of references to galls from this era focused on the, mainly medical, uses of galls and on how the galls were generated. A number of explanations were offered to account for gall formation as well as for the emergence of insects from their interior. These include rather colourful explanations such as that gall formation is achieved by the corruption of the plant's essential life force by some noxious compound (in Csóka, 1997), or that insects within galls are the result of spontaneous generation as well as more intriguing (ecologically at least) explanations such as that the insect eggs were laid on the ground, taken up by the roots of the plant and transported in the sap to the site of gall development (in Askew, 1984). The link between insect oviposition and gall formation was not established until the 17^{TH} century, when Antony van Leeuwenhoek and Marcello Malpighi (who also was Pope Innocent XII' s personal doctor) independently provided evidence demonstrating the causal role of the insect in the generation of the gall. Subsequent work on gallwasps during the second half of the 19^{TH} century lead to the confirmation, by Adler and by Beijerinck, that the insect larva is the causative agent of gall formation (in Askew, 1984) and to the discovery, by Bassett, of cyclical parthenogenesis in oak gallwasps (Wehrmaker, 1998). A very prolific worker on gallwasps and the first to focus on the evolutionary history of the group was Alfred Kinsey (of the 'The Kinsey Review' fame). Kinsey worked for almost twenty years on the systematics and evolution of gallwasps producing three major monographs on the genera *Neuroterus* and *Cynips* (Kinsey, 1923, 1930, 1936) and drawing from his gallwasp data he reached more general conclusions regarding the evolution of species and taxonomic categories (e.g. Kinsey, 1937), before shifting the focus of his research to (allegedly) more interesting animals. In the past, galls were not only attractive to naturalists and researchers; they have been commercially exploited by the dying and tanning industries (through extraction and subsequent use of the high quantities of tannins found in galls) as well as for the manufacturing of writing inks.

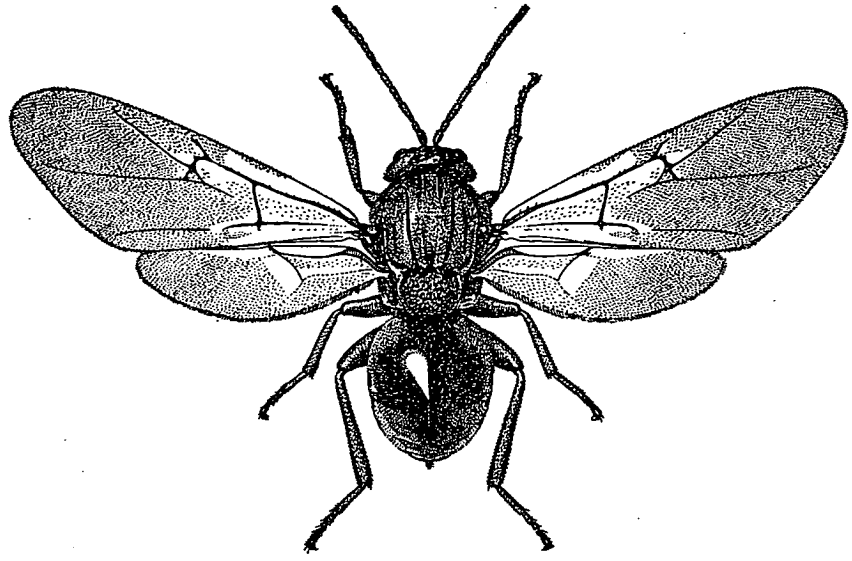


Figure 1.2. A parthenogenetic generation female of *Andricus kollari*. From Eady and Quinlan (1963).

1.6.1 General biology of gallwasps

Gallwasps (Hymenoptera: Cynipidae) are typically small winged insects of black, brown or dull yellow colour with a laterally compressed abdomen (Fig. 1.2). The majority of gallwasp species are characterised by their ability to induce highly characteristic outgrowths on herb and woody plants which are known as galls. Eggs are laid by the females into meristematic plant tissue (tissue that is capable of growth and differentiation) and when the larva hatches from the egg it induces gall formation. Each species has precise preferences in the location of its eggs, selecting a single plant species (see also Chapter 5) and a particular organ on that plant (see Chapter 7). The internal anatomy of cynipid galls is remarkably similar, despite the diversity in outer form and structure (Rohfritsch, 1992), with the most prominent characteristic being the presence of concentric layers of differentiated cells around the larval chamber. Three successive layers of parenchymatic, schlerenchymatic and parenchymatic (again) cells separate the larval chamber from the outer world (Rohfritsch, 1992). The chamber is additionally lined with a thin layer of nutritive

tissue, through which feeding of the larva occurs (Rohfritsch, 1992). The number of larval chambers per gall varies between species. Certain species form single-chambered galls (*unilocular* or *monothalamous*), whereas others form multi-chambered galls (*multilocular* or *polythalamous*) (Askew, 1984).

In addition to the gall-inducers, the gallwasp family includes a number of inquiline species. These are species that have lost the ability to induce galls themselves *de novo*; instead, their phytophagous larvae develop inside galls induced by other cynipids (some inquilines have retained some ability to modify the form and structure of their host gall) (Ronquist, 1994).

Gallwasps are either *univoltine* (one generation per season, usually in late summer / autumn) or *bivoltine* (two generations per season – one in late summer / autumn and one in spring). Certain gallwasps have complex lifecycles, involving the strict alteration of a sexual with a parthenogenetic generation (*cyclical parthenogenesis* or *heterogony*). In cyclical parthenogenetic species the wasps of each generation are remarkably different in morphology; in the past they have frequently been classified as separate species (Wehrmaker, 1998) and research is still continuing regarding pairing of generations (e.g. Lund, *et al.*, 1998, see also Chapter 7). In general, wasps of the parthenogenetic generation are larger and with a higher fecundity compared to the sexual ones. Females from each generation also induce different types of gall, with the parthenogenetic generation gall being more conspicuous than the sexual generation gall. In the parthenogenetic generation, the larva spends usually the whole winter inside the gall, feeding on nutritive tissues supplied by the host plant, only to emerge as an adult wasp, although many species emerge as adults before winter, and over-winter outside the gall. The sexual generation, in many bivoltine species, is developing quickly having induced a small and inconspicuous gall.

1.6.2 Cynipid phylogenetic relationships

Gallwasps are members of the Cynipoidea, a major lineage of predominantly parasitoid wasps within Hymenoptera (Ronquist, 1999), which are thought to have originated some 80 million years ago in the late Cretaceous from ancestors whose larvae were internal parasitoids of gall-inducing insects (Ronquist, 1999). The family is distributed throughout the Palaearctic and currently includes 1369 described cynipid species (Liljeblad and Ronquist, 1998). Nordlander (1984) estimated the total number of species to be in the range of 2,500 – 6,000, but the current rate of discovery of new species suggests that this estimate is exaggerated with more realistic estimates suggesting a number of some 1,800 species (Nieves-Aldrey, 1994). For historical reasons (Kinsey, 1920), the Cynipidae are not classified into subfamilies but into tribes. The first study of higher-level relationships within Cynipidae was performed by Kinsey (1920). Through study of three morphological (two regarding variation in wing venation and another regarding variation in the second abdominal segment) and four biological characters (host plant use, gall structure, reproduction and alternation of generations) at the intra- and inter-specific level he formulated a phylogenetic hypothesis based on intuitive evaluation of the results, distinguishing three tribes. According to Kinsey' s phylogenetic scheme, tribe Cynipini includes gall-inducers on oaks (genus *Quercus*) and allied Fagaceae, tribe Rhoditini includes gall-inducers on roses (Rhoditini has been recently found to be a junior synonym of Diplolepidini, the current naming of the rose gallwasp tribe, Ronquist, 1999) and tribe Aylacini includes the gall-makers on all other plants. The inquiline wasps (nowadays placed in a separate tribe, Synergini) were not dealt with (Kinsey, 1920). Subsequent work, partly based on the morphology and partly on the biology of gallwasps, by Nieves-Aldrey (1994) and Ronquist (1994; Ronquist, 1995) lead to the separation of gallwasps into six tribes. The tribe with the highest species-richness is that of oak gallwasps (tribe Cynipini), followed by the gallwasps inducing galls on herbs (tribe 'Aylacini', the inverted commas indicating the tribe' s paraphyletic status – see below). Another speciose tribe is that of the inquiline gallwasps (tribe Synergini), numbering 170 species. Tribes Pediaspidini (galls *Acer* trees), Diplolepidini (rose gallwasps, make galls on roses) and Eschatocerini (they

gall *Acacia* and *Prosopis* plants – Fabaceae) have a much lower species richness. Specific information on species richness, host plants and number of genera found within each tribe is listed in Table 1.1.

Table 1.1. Number of genera and species belonging to the six tribes of Cynipidae and the host plants each tribe uses. After Liljeblad and Ronquist (1998).

Tribe	Genera	Species	Host
'Aylacini' herb gallers	21	156	Asteraceae, Rosaceae, Lamiaceae, Papaveraceae, Apiaceae, Valerianaceae, Brassicaceae.
Cynipini Oak gallers	44	974	Fagaceae (mostly *Quercus*).
Eschatocerini	1	3	*Acacia*, *Prosopis* (Fabaceae).
Pediaspidini	2	2	*Acer* (Aceraceae).
Diplolepidini Rose gallwasps	2	63	*Rosa* (Rosaceae).
Synergini inquilines	7	170	Inquilines in galls induced by *Diastrophus*, *Diplolepis* and Cynipini.

Ronquist (1994) and Liljeblad and Ronquist (1998) are the only authors who have considered the phylogenetic relationships among cynipid tribes, using modern phylogenetic techniques on a rather large morphological character dataset for thirty-seven gallwasp species from thirty-one genera representing all six tribes (Fig. 1.3). According to their phylogenetic hypothesis, tribe 'Aylacini' form a paraphyletic assemblage of two distinct clades. One of the clades is the sister group of tribe Synergini, whereas the other 'Aylacini' clade is the outgroup of the woody rosid gallers lineage (tribes Cynipini, Eschatocerini, Pediaspidini and Diplolepidini; they all gall woody plants, Fig. 1.3). Within the woody rosid gallers, the Cynipini are the sister group of Pediaspidini, and the Diplolepidini are the sister group of Eschatocerini. No molecular phylogenetic study has tested these relationships. In Chapter 3 (see also, Rokas, *et al.*, 2001b) I have made a first attempt to provide a

molecular phylogeny for the six gallwasp tribes, by generating more than 5,000 base pairs of DNA sequence data (representing 8 loci) for 8 exemplar species representing four out of the six tribes. The results obtained do not allow comparison with the morphology-based phylogeny, due to poor resolution of the molecular phylogenies. The only consistent result is the recovery of the oak gallwasp clade. However, the study has allowed identification of a range of molecular markers for a range of taxonomic levels within gallwasps, which may form a robust basis for subsequent phylogenetic work (see Chapters 6 and 7).

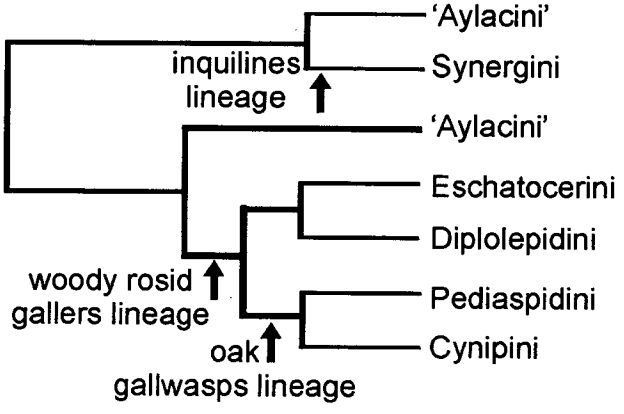


Figure 1.3. Phylogenetic relationships among the six tribes of cynipid wasps according to the only available morphology-based phylogeny (Liljeblad and Ronquist, 1998).

Moving towards lower-level gallwasp phylogenetics, the knowledge of phylogenetic relationships within each tribe is also limited. Prior to work described here, the only available molecular phylogenetic studies were done as part of an attempt to understand the evolution of gall structure in oak gallwasps (Stone and Cook, 1998) and rose gallwasps (Plantard, *et al.*, 1998b). Molecular phylogenetic work has also been done on the American oak gallwasp fauna revealing extensive paraphyletic relationships between the various genera (Drown and Brown, 1998), perhaps highlighting the poor taxonomic status of the Nearctic fauna (e.g. Abrahamson, *et al.*, 1998). In Chapter 7, as part of examining the evolution of host shifts and host organ shifts in the genus *Andricus*, I have generated a phylogeny

using long-wavelength opsin, a nuclear locus recently proposed to be useful for insect phylogenetics (Mardulyn and Cameron, 1999; Ascher, *et al.*, 2001; Rokas, *et al.*, 2001b, Chapter 3). The phylogenetic pattern suggested by long-wavelength opsin is concordant with the cytochrome *b* phylogeny of Stone and Cook (1998). In Chapter 6, I have expanded the molecular phylogeny of the genus *Andricus* and the European oak gallwasps of Stone and Cook (1998), by sequencing additional oak gallwasp species for the mitochondrial locus cytochrome *b* (reaching a total of 62 species). Clades within the genus *Andricus* identified by Stone and Cook (1998) are robustly supported in this extended dataset. However, the relationships between the various oak gallwasp genera are not well resolved, partly due to the substitutional saturation of the molecular marker and partly due to the high ratio of taxa to characters. Additionally, in an attempt to identify how intra-specific variation affects inter-specific relationships, I have added to the extended dataset more than one haplotype for 16 oak gallwasp species (reaching a total of 85 taxa). Interestingly, analysis of the 85-taxon dataset suggests the non-monophyly for the majority of the 16 taxa. Underlying causes of the observed non-monophyly, such as lineage sorting of ancestral polymorphisms or hybridisation between oak gallwasp species, are discussed in Chapter 6.

1.6.3 Modes of reproduction in gallwasps

All hymenopteran insects are haplodiploids, that is females have two sets of chromosomes (diploid) whereas males only one (haploid). The most common reproductive mode of haplodiploid insects is *arrhenotoky*; females are the product of fertilised eggs (and hence with a diploid genome), while males are the product of unfertilised eggs (and hence haploid). Arrhenotoky is likely to be the ancestral mode of reproduction for the gallwasp family (Liljeblad and Ronquist, 1998), and is the predominant reproductive mode in inquiline gallwasps, herb gallers and rose gallwasps (tribes 'Aylacini', Synergini, Eschatocerini and Diplolepidini). Apart from arrhenotoky, two other reproductive modes exist within Cynipidae; cyclical parthenogenesis and parthenogenesis.

Cyclical parthenogenesis (the alternation of an arrhenotokous and a

parthenogenetic generation) is found only in oak gallwasps (tribes Cynipini and Pediaspidini). Cytological work on the parthenogenetic generation of seven cyclically parthenogenetic species (Doncaster, 1910, 1911, 1916; Dodds, 1939) suggests that the mechanism of parthenogenesis is apomixis (neither chromosome reduction nor fusion of nuclei nor any of the phenomena related to automixis or sexual reproduction occur, resulting in a population with extremely high heterozygosity, Suomalainen, *et al.*, 1987). This finding is corroborated by population genetic data from sexual generation wasps of the oak gallwasp *Biorhiza pallida* suggesting that reproduction in the parthenogenetic generation is through apomixis (Atkinson, 2000).

Parthenogenetic species may be found throughout the Cynipidae (Askew, 1984). The cytogenetics of parthenogenesis have been worked out for very few species. For example, parthenogenesis in the rose gallwasp *Diplolepis rosae* occurs through automixis (Stille and Davring, 1980). Automixis is the name given to many different cytological types of parthenogenesis that most often lead to the production of homozygous individuals. The specific mechanism via which automictic parthenogenesis occurs in *Diplolepis* is gamete duplication; the two identical sets of chromosomes do not separate during the anaphase of the first mitotic division and the net result of the first mitotic cycle is a single nucleus containing two copies of the same set of chromosomes (Suomalainen, *et al.*, 1987). Recent studies indicate that parthenogenesis in rose gallwasps and in herb gallers is induced by the bacterial endosymbiont *Wolbachia* (Plantard, 1997; Plantard, *et al.*, 1998a; Plantard, *et al.*, 1999).

The situation is different for parthenogenetic species in the oak gallwasp tribes (Cynipini and Pediaspidini). Due to the difficulties associated with the pairing of sexual and parthenogenetic generations and with the inconspicuous nature of many sexual generation galls, many species of oak gallwasps have been described only from their parthenogenetic generation, and until recently have been believed to be exclusively parthenogenetic. Recent work by Atkinson and colleagues (Atkinson, 2000; Atkinson, *et al.*, 2001, Atkinson, *et al.*, in preparation, see also Chapter 7) has focused on using indirect techniques (based on the use of wide array of molecular markers and population genetic and phylogenetics software) to establish whether

certain oak gallwasp species are exclusively parthenogenetic or not. For all the parthenogenetic species examined so far, the molecular evidence is strongly suggesting that these species are actually cyclical parthenogens with, as yet, unidentified sexual generations. In parallel with the work of Atkinson and colleagues, in Chapter 8 I investigated whether the bacterial endosymbiont *Wolbachia*, which is known to induce parthenogenesis in certain haplodiploid insect groups (including rose gallwasps) (Stouthamer, 1997; Plantard, *et al.*, 1998a; Plantard, *et al.*, 1999), is implicated in the generation of parthenogenesis observed in oak gallwasps. Almost all the oak gallwasp species which are known only from their parthenogenetic generations were not found to be infected with *Wolbachia*. Additionally, if *Wolbachia* was the causative agent of reversal from a cyclical parthenogenetic mode to a parthenogenetic mode, one would not expect the complete loss of the sexual generation, but its reversal to a second parthenogenetic generation. What is observed for parthenogenetic species in the oak gallwasp clade is the complete loss of the sexual generation, leading to a parthenogenetic species with just one generation per season (and not two as would be predicted if infection with *Wolbachia* was the causative agent of parthenogenesis).

1.6.4 Oak gall wasp communities

The presence of highly nutritious plant tissues in cynipid galls attracts numerous phytophagous insects and these in turn entomophagous species. For example, Askew (1961) found some oak galls supporting as many as 22 species, whereas the oak gall induced by *Biorhiza pallida* (see Chapter 4) has over 75 species associated with it (Mani, 1964). Thus cynipid galls commonly house a small micro-community (Wiebes-Rijks and Shorthouse, 1992). Given the ease of sampling and the diverse assemblage of inhabitants they contain, it is not surprising that some of the best known work on insect-based food webs and community structure concerns the micro-community of insect galls (Godfray, 1994), among which perhaps the best studied group is that of oak gallwasps. Galls of oak gallwasps, in addition to the gall-inducer, often harbour inquiline gallwasps (tribe Synergini) and a characteristic assemblage of chalcid parasitoids (Askew, 1961). The communities harboured in a

single gall range from a single insect to hundreds of insects, of up to twenty species, on 5 trophic levels (Askew, 1984). The parasitoid fauna attacking cynipid galls is highly diverse, with parasitoids attaching Palaearctic cynipid galls belonging to 27 genera in 8 hymenopteran families (Askew, 1984). Furthermore, the majority of oak gallwasp species (see above) have two generations per season, one sexual and one parthenogenetic, with wasps of each generation developing in very different galls with very different associated communities (Stone, *et al.*, 1995).

The rich and diverse insect community associated with gallwasps is an interesting model for addressing questions regarding the pathways via which the endosymbiont *Wolbachia* is transmitted in natural populations. Few assemblages have been analyzed to date, with most studies inferring frequent horizontal transfer of *Wolbachia* strains between different insect hosts on the basis of the incongruence observed between phylogenies of hosts and their *Wolbachia* strains (e.g. Werren, *et al.*, 1995b; Schilthuizen and Stouthamer, 1997). Direct evidence for the possibility of horizontal transfer comes from the experimental demonstration that *Wolbachia* can be transmitted by infection (Huigens, *et al.*, 2000). However, the route by which this horizontal transfer occurs is less clear. One possibility is that parasitoids may be the agents of transfer (Werren, *et al.*, 1995b), although an early attempt to test this hypothesis on a leaf-miner and an aphid food web provided negative results (West, *et al.*, 1998). More recent work on a *Drosophila*-Hymenoptera web shows cases of sharing of the same *Wolbachia* strains between parasitoids and their insect hosts (Vavre, *et al.*, 1999) (although crucially the two studies differed in the type of marker used, making the contrast between the two studies more ambiguous). The detailed natural history knowledge of the host-parasitoid system found within the family of gallwasps (consisting of the gall-inducers, their inquilines as well as other parasitoid taxa) represents an attractive ecological assemblage for the study of *Wolbachia* transmission (e.g. Askew, 1961; Askew, 1984). In Chapter 8, I present the results of screening of 64 gallwasp species from three different orders at two trophic levels (gall-inducers and their inquilines) for infection with *Wolbachia*. My data suggest that there have been frequent incidences of horizontal transfer of *Wolbachia* strains from one species to another, including transfers from inquilines to gall-inducers (or vice versa).

1.6.5 The intricate relationship between oak gallwasps and their hosts, the oaks

Generally, the distribution of parasites is constrained by the distribution of their hosts. The Cynipini (oak gallwasps) are the most speciose tribe within the gallwasp family and the one whose species have been most studied in this thesis. A number of Chapters in this thesis (4, 5 and 7) directly or indirectly address issues regarding the distribution of particular oak gallwasp species and how they are constrained by the distribution of their oak host (s), so a brief introduction to the world of oaks is necessary for adequate understanding of patterns of evolution seen in oak gallwasps.

Oaks (genus *Quercus*) are thought to have originated in the New World some 40 million years ago and the genus is currently comprised of approximately 500 species (Manos, *et al.*, 1999). The taxonomic difficulties associated with the taxonomy of the genus are notorious (Burger, 1975; van Valen, 1976) and are partly due to the extensive hybridisation observed between the various species (e.g. Toumi and Lumaret, 1998). Despite these problems, the genus *Quercus* is currently divided into two subgenera *Cyclobalanopsis* and *Quercus*, the latter being comprised of four sections (Manos, *et al.*, 1999); sections *Lobatae* (red oaks) and *Protobalanus* (golden cup or intermediate oaks) have a Nearctic distribution, whereas the section *Quercus sensu stricto* (*s. s.*) (white oaks) may be found both in the Nearctic and the Palaearctic. The fourth section (section *Cerris*) has a Palaearctic distribution.

European oak gallwasp species occur on both oak sections found in Europe (sections *Quercus s. s.* and *Cerris*). Oak gallwasp species are very selective on the oak species and section that induce their galls on (Abrahamson, *et al.*, 1998). For example, all parthenogenetic generations of species in the genus *Andricus* induce galls on white oaks (section *Quercus s. s.*). In contrast, the sexual generation wasps of *Andricus* species induce their galls either on white oaks or on section *Cerris* oaks, but never on both. In Chapter 7 I investigate whether the switch from a sexual generation gall on a white oak to a sexual generation gall on a section *Cerris* oak has occurred multiple times in the evolution of the genus *Andricus*. The phylogenies produced by two independent molecular markers strongly favour a single switch

from a sexual generation gall on white oaks to a sexual generation gall on section *Cerris* oaks.

The ultimate dependence of gallwasp distributions to their oak hosts' distributions is best observed at the intra-specific level, in studies on the effect of the Pleistocene ice ages on the distributions of both gallwasps and their oak hosts. More specifically, during the last 2.5 million years of earth history (also known as the Pleistocene or the Quaternary), Earth' s climate has been dominated by a series of major glacial periods with a periodicity of about 100 kyr. These ice ages have been separated by shorter, milder interglacials, during which the climate was more similar to that of the present day (Hewitt, 1996, 1999, 2000). The ice age cycles lead to repeated cycles of range expansion and contraction in many plant and animal taxa, including gallwasps and their oak hosts. In Chapters 4 and 5 I have studied the phylogeography of two oak gallwasp species, *Biorhiza pallida* (Chapter 4) and *Andricus quercustozae* (Chapter 5) and their differential response regarding range expansion into central and northern Europe in the current interglacial, which is associated with differences in the constraints imposed by their specific oak hosts in these regions (see also Stone, *et al.*, 2001, Appendix I).

Oak gallwasps are not only taxon-specific but also organ-specific. Each generation of oak gallwasps, with few exceptions, attacks only one of buds, shoots, roots, leaves, catkins or acorns. In Chapter 7 I show, by focusing on the genus *Andricus*, that evolutionary changes in gall location occur at a significantly higher rate than evolutionary changes in host plant usage, perhaps suggesting that gall induction is more dependent on the oak genotype than on the characteristics of a particular host tissue.

SECTION I: MOLECULAR MARKERS

CHAPTER 2

Rare genomic changes (RGCs) as a tool for phylogenetics

'The present crisis of molecular phylogeny will be overcome only if one recognises that the most important step in the analytical process is the search for good characters'

Patrick Forterre and Hervé Philippe (1999)

2.1 Abstract

DNA sequence data have offered valuable insights into the relationships between living organisms. However, most phylogenetic analyses of DNA sequences rely primarily on single nucleotide substitutions, which might not be perfect phylogenetic markers. Rare genomic changes (RGCs), such as intron indels, retroposon integrations, signature sequences, mitochondrial and chloroplast gene order changes, gene duplications and genetic code changes, provide a suite of complementary markers with enormous potential for molecular systematics. Recent exploitation of RGCs has already started to yield exciting phylogenetic information.

2.2 Introduction

In recent years, considerable progress has been made in the field of molecular phylogenetics. A significant driving force has been the increasing technical ease of DNA sequencing, which has led to the dominance of primary sequence data as indicators of the historical relationships between taxa. Important advances have also occurred in the computational analysis of DNA sequence data (Hillis, *et al.*, 1996), such as improved methods for modelling patterns of nucleotide substitutions (Yang,

et al., 1994). However, the task of phylogenetic reconstruction using molecular sequences is not without problems. To a large extent, these stem from the fact that the dominant methods for molecular phylogeny reconstruction exploit nucleotide substitutions (plus, in some cases, single-site insertions or deletions) as indicators of divergence or common descent. Convergent evolution of nucleotide bases, differing substitution rates among sites and lineages (Aguinaldo, *et al.*, 1997), saturation of mutations at variable sites (Philippe and Laurent, 1998), non independent substitutions among sites (Averof, *et al.*, 2000) and functional constraints at the molecular level (Lee, 1999) are just a small sample of the potential caveats when using these types of data. As a result, phylogenetic hypotheses based on primary sequence data can sometimes be equivocal (Huelsenbeck, 1998; Forterre and Philippe, 1999), whereas other hypotheses can simply be incorrect (Philippe, 1997; Naylor and Brown, 1998). The advent of the genomic era has brought the opportunity to consider other types of data embedded in DNA sequences. Here, we consider the phylogenetic use of large-scale mutations, which occur relatively infrequently, rare genomic changes (RGCs). Researchers have already started using RGCs for inferring relationships between living organisms.

2.3 Rare genomic changes

We define RGCs as large-scale mutational changes that have occurred in the genomes of particular clades (for definitions of terms see Table 2.1). Examples of RGCs (Table 2.2) include intron indels, retroposon integrations, signature sequences in genes, changes to organelle gene order, gene duplications and genetic code variants. Most RGCs represent changes caused by single (or few) mutational events; in our discussion of RGCs we do not include genomic characteristics that are, most probably, the end result of multiple processes (e.g. genomic compositional contrasts, Campbell, *et al.*, 1999). Until recently, many studies mapped RGCs onto existing phylogenies to gain insight into their mode of evolution. The consensus that has emerged is that RGCs are often evolutionarily conserved and phylogenetically informative. We believe the time has come to turn the question around: what can RGCs tell us about phylogenies themselves?

RGCs provide an independent source of phylogenetic information, largely immune from some of the problems that affect primary sequence data. A major difficulty with this approach is the identification of these rare mutations in the clades of interest. However, the increasing automation of molecular techniques has brought us to the dawn of the genomic era where tremendous amounts of information, freely available in primary literature and public databases, are generated. Additionally, protocols have been developed for the targeted identification of many, if not all, RGCs. Here, we argue that the application of RGCs to phylogenetics can offer new insights into evolutionary history. Furthermore, whenever primary sequence data are generating conflicting or equivocal results, RGCs offer an independent way of evaluating alternative phylogenetic scenarios.

Table 2.1. Glossary of terms used in this chapter.

Term	Definition
Bilateria	the bilaterally symmetrical animals
Homoplasy	a general term denoting that the acquisition of the same character state in two taxa is not because of common descent. This can arise by parallel evolution (independent acquisition from the same ancestral state), convergent evolution (independent acquisition from different ancestral conditions) or secondary loss (reversion from the derived to the ancestral condition)
Indel	an insertion or deletion event
LINEs	(long interspersed elements) a class of retroposons that are capable of self-transposition
Orthology	the relationship between two homologous loci derived from a speciation event
Paralogy	the relationship between two homologous loci derived from a duplication event
Polyphyly	when a group does not include the most recent common ancestor of all its members

Protein domain	a well defined region within the protein. It can be distinguished on the basis of function or structure. For example, the homeodomain is a 60-amino acid domain shared by proteins encoded by homeobox genes
Protein motif	any stretch of contiguous sequence within a protein that has been evolutionarily conserved
RGC	(rare genomic change) a large-scale mutational change that has occurred in the genome of a particular clade
Retroposons	the class of transposable elements that relocate in the genome via an RNA intermediate using the enzyme reverse transcriptase
Signature sequences	shared conserved insertions or deletions in proteins or RNAs
SINEs	(short interspersed elements) a particular class of retroposons that have lost the ability to transpose themselves (to transpose they use another class of mobile elements, LINEs)
Synapomorphy	a shared derived character state that suggests a monophyletic grouping

2.4 RGCs as 'Hennigian' markers

The field of phylogenetics has been strongly influenced by the founder of the cladistic methodology, the German entomologist Willi Hennig. Hennig argued that only shared derived characters (synapomorphies; Table 2.1) should be used as indicators of common descent. Plotting the distribution of synapomorphies is the essence of cladistic reconstruction. The principal hindrance to this task is homoplasy (see Table 2.1 for definition). In general, character states that arise rarely will not be prone to extensive convergent or parallel evolution, which should contribute to low levels of homoplasy. Although the precise frequency of occurrence of most RGCs has not been robustly estimated, large-scale mutations are generally rare. Additionally, precise secondary loss of the character (homoplasy because of reversion) is likely to be extremely rare for most large-scale mutations and has been

demonstrated to be so in some cases (e.g. short interspersed element insertions - Shedlock and Okada, 2000; Table 2.2). Therefore, with respect to homoplasy, RGCs might constitute good markers of common descent. In Table 2.2 we provide a summary of the characteristics and phylogenetic applicability of various categories of RGC and in the next section we expand on a few examples published recently to demonstrate their potential use.

2.5 Of fish and flies: intron indels as clade markers

The power and robustness of RGCs is well demonstrated by the study of Venkatesh *et al.* (1999), in which intron indels (Table 2.1) were used to investigate fish phylogeny. Venkatesh *et al.* identified seven intron positions (in five genes) that are present in the pufferfish *Takifugu rubripes* but not in the homologous genes of mammals. Four introns were also found in the rhodopsin gene that were present in the ancestral chordate rhodopsin gene (as inferred by their presence in basal chordates, such as lampreys and skates, and in the more apical lineage of mammals) but were absent in the pufferfish. Several ray-finned fish species (class Actinopterygii) were screened for the presence or absence of these eleven intron indels, and these data were used to reconstruct evolutionary relationships (Venkatesh, *et al.*, 1999). From all indels, only one showed considerable homoplasy and an unclear phylogenetic signal, whereas all the others were unique synapomorphies able to resolve phylogenetic relationships. It is noteworthy that some of the relationships resolved, such as the basal position of bichirs (*Polypterus* spp.) within the Actinopterygii, have proved contentious using primary sequence comparisons.

Marker	Taxonomic resolution	Homoplasy	Taxa in which applicable
Intron indels	Wide ranging	Low	Eukaryotes
Retroposons (SINEs and LINEs)	Within Orders	Zero to very low	Animals
Signature sequences	Wide ranging	Unknown but recognisable	All life
mtDNA genetic code variants	Phyla to Classes	Low to moderate	Eukaryotes
Nuclear DNA genetic code variants	Phyla	Low to moderate	All life
mtDNA gene order	Wide ranging (Phyla to Families)	Low to moderate in animals. High in plants, fungi and protists	Eukaryotes
cpDNA gene order	Families	Low	Plants
Gene duplications	Wide ranging	Unknown	All life
Comparative cytogenetics	Within Phyla	Unknown	All life (lateral gene transfer is prevalent in prokaryotes)

Table 2.2. Summary of useful characteristics of rare genomic change (RGC) markers for phylogenetic purposes. For more detailed information, see references cited in the text. Abbreviations: SINEs, short interspersed elements; LINEs, long interspersed elements.

Another recent use of an intron indel as a phylogenetic character deals with the placement of the insect order Strepsiptera within holometabolous insects (Rokas, *et al.*, 1999, Appendix I). Strepsipteran forewings resemble the hindwing balancing organs of flies (order Diptera), which are known as halteres. Among other phylogenetic scenarios, an affinity of Strepsiptera to Coleoptera has been widely discussed, based primarily on the use of hindwings for flight in both orders. An alternative proposal is a sister group relationship with Diptera (Whiting and Wheeler, 1994; Whiting, *et al.*, 1997). In this case, halteres could be homologous, but a radical homeotic mutation might have reversed their position in Strepsiptera (Whiting and Wheeler, 1994). Evidence from morphology is equivocal (Kristensen, 1999) and 18S rDNA sequence data have generated a lively debate between researchers favouring different phylogenetic reconstruction methods (Whiting, *et al.*, 1997; Huelsenbeck, 1998). Rokas *et al.* noted a unique intron insertion in the homeobox of the *engrailed* gene of Diptera and Lepidoptera, which is absent from other insects and all outgroups. Possession of the intron in Strepsiptera would support a sister group relationship with Diptera, whereas its absence would argue against this affinity. Cloning of the Strepsipteran homologue of *engrailed* showed that the intron is absent in Strepsiptera, thus suggesting that the halteres of Strepsiptera and Diptera might not represent a rare case of natural homeotic transformation but might instead represent a remarkable case of convergent evolution (Rokas, *et al.*, 1999, Appendix I) (Fig. 2.1).

2.6 Of SINEs and LINEs

Retroposons (Table 2.1) belong to the group of transposable elements that use an RNA-mediated mode of transposition (Shedlock and Okada, 2000). Retroposon integrations, especially from the class of SINEs (retroposons that lack the ability for self-amplification), have been used successfully as phylogenetic markers; an application pioneered by Okada and colleagues in the 1990s (Kido, *et al.*, 1991; Murata, *et al.*, 1993). It has been argued that SINE integrations come close to being 'perfect' markers of common descent because integration is apparently random and irreversible, and because most eukaryotic genomes have an abundance of SINE elements (Shedlock and Okada, 2000). Their presence or primitive absence can also

be readily detected by PCR amplification across integration sites. Successful applications of SINEs include the generation of convincing support for a sister group relationship between whales and hippopotamuses (Nikaido, *et al.*, 1999), also known as the 'whippo' hypothesis (Fig. 2.1), and detailed insight into salmonid fish phylogeny (Murata, *et al.*, 1993). Criticisms of SINEs include: the non-independence of SINE insertions (Miyamoto, 1999) (several can be integrated at the same time, although at different sites), incomplete lineage sorting (Hillis, 1999; Miyamoto, 1999) (although this applies to all characters), the considerable amount of work needed for their development (Shedlock and Okada, 2000) and practical limits to detection beyond circa 30% difference in sequences flanking orthologous elements (Shedlock and Okada, 2000). In our view, these are simply factors to be considered when designing or interpreting phylogenetic studies; they do not detract significantly from the robustness of SINE markers. Another class of retroposons are long interspersed elements (LINEs; Table 2.1) - the main difference from SINEs being their ability for self-amplification. LINEs have been used not only for determining the cladogenetic pattern, but also for dating speciation events. Verneau *et al.* (Verneau, *et al.*, 1998) exploited the fact that LINE elements belonging to the L1 family rapidly generate defective copies, which are retained in the genome and mutate at the neutral rate, to resolve and date the phylogenetic history of the rodent genus *Rattus*.

2.7 Animals, archaebacteria and archezoa: the use of signature sequences

The complementary use of primary sequence data and RGCs for phylogenetic purposes is shown by attempts to reconstruct the interphyletic relationships of animals. Recent studies using 18S rDNA sequences have suggested a three-branched Bilateria (Table 2.1) tree comprising of the Deuterostomia, the Lophotrochozoa and the Ecdysozoa (Aguinaldo, *et al.*, 1997). Lophotrochozoans include spiral cleaving phyla, such as molluscs, annelids, platyhelminths and nemerteans, plus the lophophorates; whereas Ecdysozoa include arthropods, onychophorans, priapulids and nematodes (all sharing moulting). This proposal was controversial because it

represented radical restructuring of the classic tree of animal phyla. It would also mean abandoning several well known super phyletic groupings, such as Acoelomata (animals without a coelom, which are traditionally basal in the animal tree) and Articulata (segmented protostomes), and implies that some apparently 'simple' animals, such as flatworms, are actually highly degenerate. Radical hypotheses often require independent support before they are accepted; in this case, complementary supporting data have come from RGCs within the Hox gene clusters (Balavoine, 1998; de Rosa, *et al.*, 1999). The proteins encoded by many Hox genes possess specific sequence motifs near the homeodomain, which is a domain common to all Hox genes. These sequence motifs have helped distinguish orthologous (Table 2.1) and paralogous (Table 2.1) Hox genes. Each of the three major clades has its own unique Hox genes that do not have identifiable orthologues in the others. In other words, gene duplications have yielded distinct genes in each lineage and these have acquired unique signature protein motifs (Table 2.1). For example, the lophotrochozoans share *Lox2*, *Lox4*, *Lox5*, *Post-1* and *Post-2*, whereas the ecdysozoans share *Ubx* and *Abd-B* (de Rosa, *et al.*, 1999).

As well as providing independent support for the controversial Lophotrochozoa and Ecdysozoa clades, this approach has been used to investigate the affinities of a particularly enigmatic animal phylum: the dicyemid mesozoa (Kobayashi, *et al.*, 1999). These are microscopic parasites of squid and octopus, with an amazingly simple body plan consisting of a solitary axial cell surrounded by a single layer of 10-40 ciliated outer cells. Morphology and 18S rDNA sequence data have previously failed to adequately resolve their phylogenetic position. Recently, Kobayashi *et al.* (Kobayashi, *et al.*, 1999) cloned the *Lox5* gene from a dicyemid, including the diagnostic *Lox5* peptide, thus demonstrating that these animals are almost certainly highly degenerate members of the Lophotrochozoa clade (Fig. 2.1). Indeed, dicyemids represent one of the most extreme cases of secondary simplification of morphology in the animal kingdom.

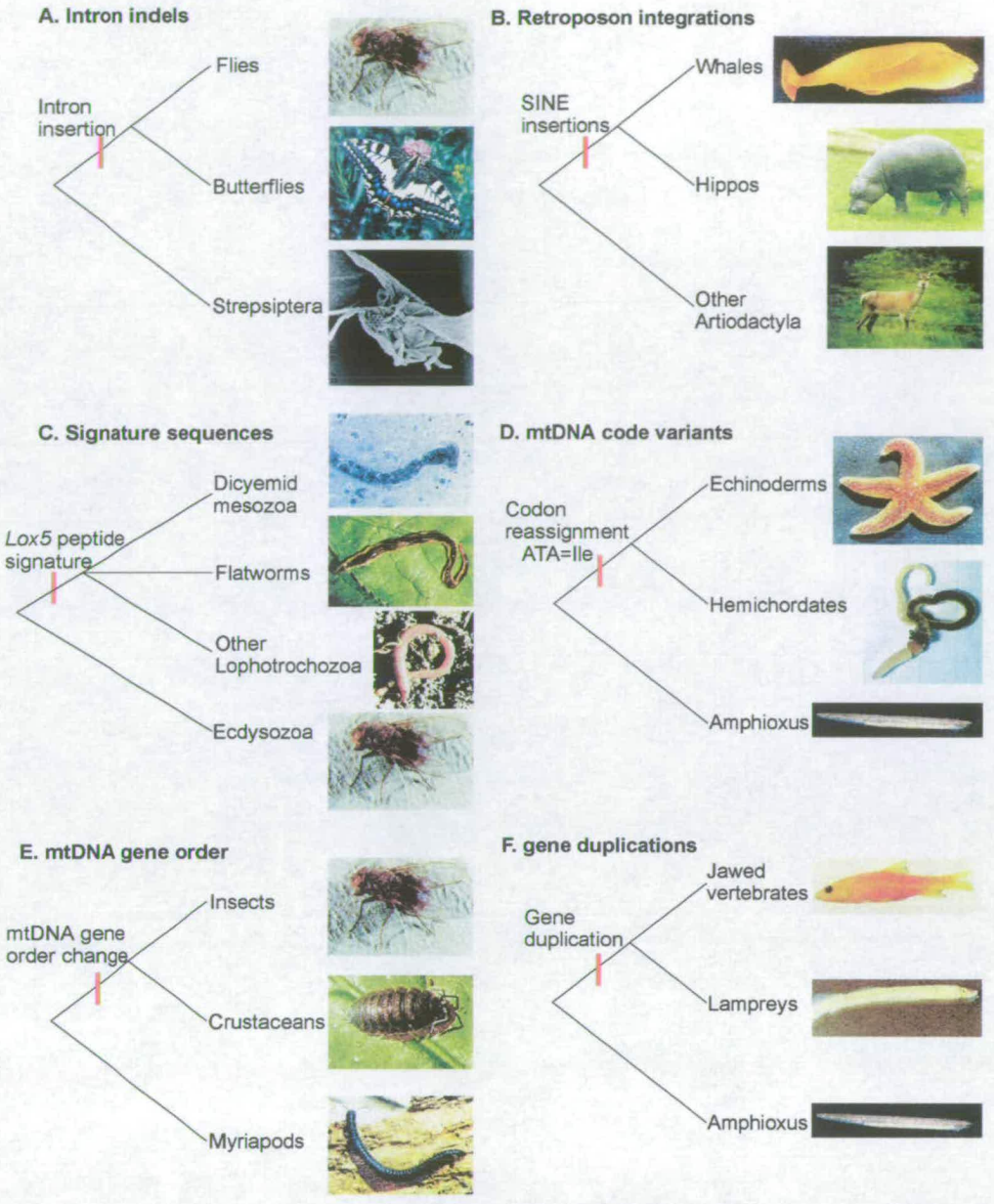


Figure 2.1. Examples of rare genomic changes (RGCs) as phylogenetic markers. A. An intron insertion in the gene *engrailed* suggests that Strepsiptera are not a sister group to Diptera (flies) (Rokas *et al.*, 1999). B. Retroposon integrations using short interspersed elements (SINEs) have established the sister group relationship of whales and hippos, to the exclusion of other Artiodactyla (Nikaido *et al.*, 1999). C. Hox gene signature sequences have robustly supported the division of protostome invertebrates into the lophotrochozoan and ecdysozoan clades (de Rosa *et al.*, 1999), and have identified dicyemid mesozoa as lophotrochozoans (Kobayashi *et al.*, 1999). D. A codon reassignment in the mitochondrial genetic code suggests that echinoderms and hemichordates are sister groups (Castresana *et al.*, 1998), a result supported by sequence data (Bromham and Degnan, 1999). E. mtDNA gene order in mitochondrial genomes supports the common grouping of insects and crustaceans, with myriapods as an outgroup (Boore *et al.*, 1998). F. Repeated events of gene duplication have occurred in the lineage leading to vertebrates [Garcia-Fernandez and Holland, 1994). Photographs reprinted with copyright permission from G. Brown (flatworm / earthworm / crustacean / myriapod / fish / fly / whale / echinoderm / lamprey), S. Ferguson (hippo), P. Holland (butterfly / amphioxus) M. Kobayashi (mesozoa), the Marine Biological Laboratory and D. Remsen (hemichordate), A. Rokas (Strepsiptera) and J. Pemberton (deer).

The Hox gene data represent a special case of 'signature sequences' (Table 2.1); the latter being defined as shared, conserved indels in proteins and RNAs. In Hox genes, the existence of distinct signature sequences in different genes, and in different clades, suggests that these motifs (such as the *Lox5* peptide) have biochemical functions. In other examples, insertions might have little functional significance; nonetheless, they can be used as RGCs for phylogenetic reconstruction; for example, there is an ongoing debate in prokaryote phylogenetics about whether archaebacteria are monophyletic. Recently, the paraphyly of archaebacteria has been supported by several signature sequences (Rivera and Lake, 1992; Gupta, 1998). However, it should be noted that historical associations within prokaryotes are still incompletely resolved owing to extensive later gene transfer. No molecular marker is immune from this all-pervading complication (Doolittle, 1999); for example, a signature sequence in the gene *hsp70* used to support the paraphyly of archaebacteria (Gupta, 1998) is also present in one of the three copies of the gene in *Escherichia coli*, suggesting a possible recent transfer (Philippe and Laurent, 1998). Other important studies using the signature-sequence approach include a confirmation that the archezoa are true eukaryotes that have lost mitochondria (Hashimoto, *et al.*, 1998) and an investigation of branchiopod crustacean phylogeny (Crease and Taylor, 1998). In branchiopod crustaceans, three unique helices in 18S rDNA were used to distinguish cladocerans from other branchiopods, demonstrating that useful sequences can be found in RNA as well as in protein sequences.

2.8 Deviant codes and shuffled genes

Several organisms use genetic codes that deviate from the standard 'universal' code. These 'deviant' codes can be useful markers for higher level phylogenetics. Keeling and Doolittle (1997) showed that a genetic code in which TAA and TAG codons encode glutamine rather than termination is used by almost all diplomonads, with the exception of the genus *Giardia*, which employs the standard genetic code. This argues for an early divergence of *Giardia* in the evolution of diplomonads and is in agreement with phylogenies from primary sequence data (Keeling and Doolittle, 1997). The diplomonad deviant code has also been found in certain green algae and

in ciliates, showing that homoplasious changes can occur. Mitochondrial genomes have the widest variety of deviant codes, while plastids show no deviation from the universal (so far!) (Keeling and Doolittle, 1997; Palmer and Delwiche, 1998). Variant mtDNA codes in animals have been studied in some detail, aided by complete sequences of mtDNA from a wide range of animals. For example, a sister group relationship between echinoderms and hemichordates is supported by the assignment of the ATA codon to the amino acid isoleucine (Castresana, *et al.*, 1998), as well as by sequence analyses (Bromham and Degnan, 1999) (Fig. 2.1), although the same reassignment has occurred independently in Cnidaria.

Gene order changes, particularly in circular genomes, such as mitochondria and chloroplasts, comprise another type of RGC that has already proved useful in phylogenetics (Boore, 1999). These arrangements, effected by inversions, translocations and duplications, generally affect several adjacent genes. They are unlikely to be reversed precisely because of their complexity; therefore, they satisfy one of the principal criteria demanded of the perfect phylogenetic marker. The second criterion, low levels of homoplasy, is also predicted to be true because convergence or parallelism would imply bias towards particular gene rearrangements or gene orders. Isolated cases of convergence have been detected (Mindell, *et al.*, 1998), suggesting bias in some taxa; however, this does not seem to be a widespread problem. Some key phylogenetic problems have been tackled using mtDNA order, with definitive results. For example, until recently, it was widely accepted that insects and myriapods were close relatives within the arthropods; indeed, these two primarily terrestrial taxa share many derived morphological characters. Several lines of evidence, including developmental data and primary sequence comparisons, have challenged this relationship, raising the alternative possibility of a crustacean-insect clade (Averof and Akam, 1995). This suggestion is effectively confirmed by the shared presence of a rare tRNA translocation within insects and crustacean mtDNA, which is not seen in myriapods, chelicerates, tardigrades, onychophorans or outgroups (Boore, *et al.*, 1998) (Fig. 2.1). In most animal taxa, changes to mtDNA gene order are rare, making these markers useful for higher level phylogenetics (Boore, *et al.*, 1998; Boore, *et al.*, 1999); although one exception might be the gastropod molluscs, where mtDNA gene order is extremely variable (Kurabayashi

and Ueshima, 2000). Similarly, plant, fungi and protist mtDNAs display rapid genome reshaping, making gene order a more appropriate marker for lower-level phylogenetics (Lang, *et al.*, 1999).

Chloroplast DNA (cpDNA) gene order has been exploited in a similar way to mtDNA gene order. For example, in 1987 Jansen and Palmer used a cpDNA inversion within the sunflower family to propose the basal position of the Barnadesiinae, with implications for biogeography and morphological evolution in this group (Jansen and Palmer, 1987). More recently, Doyle *et al.* (1996) surveyed 132 legume genera for the occurrence of a 50kb inversion, finding evidence that at least two tribes within the legumes were polyphyletic. A qualitatively different sort of rearrangement from those discussed above is deletion. For example, monophyly of the conifers is supported by loss of one copy of an inverted repeat found in cpDNA (Raubeson and Jansen, 1992).

2.9 Other potential RGCs

The list of RGCs we have described so far is not exhaustive; several other categories of large-scale mutation exist, some of which have potential for phylogenetics. For example, gene duplications have not yet been widely exploited. One difficulty is technical: unless a family of genes is arranged in a tandem array, discerning whether a duplicated copy of a gene exists is difficult because absence of evidence does not equate with evidence for absence. This problem cannot be definitively overcome except by the acquisition of complete genome sequences. Until this becomes faster, easier and cheaper, gene duplications represent a potentially untapped source of markers.

Gene families that have proved most amenable for tracing gene duplications include Hox genes and globin genes. It is no coincidence that these form stereotyped clustered arrangements that permit extra genes to be readily cloned. In the case of Hox genes, duplication of the entire gene cluster is deduced to have occurred on the vertebrate lineage, after divergence from the cephalochordate (amphioxus) lineage (Garcia-Fernandez and Holland, 1994). This dispels the view that cephalochordates are degenerate vertebrates (Fig. 2.1), although in reality this notion has had few

supporters in the past century. There is also good evidence that the Hox gene clusters underwent additional duplications somewhere within the ray-finned fish lineage; however, more taxa need to be surveyed before this event can be used as a phylogenetic marker (Stellwag, 1999). Although gene duplications are sufficiently widespread to be used as phylogenetic markers, there is still the potential for homoplasy. If new genes can be exploited for new roles (or to refine old roles), convergent duplication and retention is an ever present possibility. Somewhat paradoxically, gene losses might prove more useful markers than gene duplications. Although homoplasy is still a real possibility, at least reversion is virtually impossible. For example, after an additional round of Hox gene cluster duplications in ray-finned fish, approximately 21 individual gene losses (plus one cluster loss) must have occurred in the lineage leading to zebrafish (Stellwag, 1999). This large number of independent events provides great scope for refinement of ray-finned fish phylogeny.

The study of the differences in chromosome structure and appearance between species has given rise to the field of comparative cytogenetics (O'Brien, *et al.*, 1999; O'Brien and Stanyon, 1999): another source of RGCs potentially useful for phylogenetics. By comparing chromosomes, a phylogeny can be constructed based on the minimum number of rearrangements required or the maximum number of shared segments. Existing data are limited and come mainly from mammals, but there are exciting prospects (O'Brien, *et al.*, 1999). As partial and complete genome sequences are obtained from an ever-growing number of species, the resolution of this approach can be greatly refined. Inversions, translocations and duplications, at the scale of one to a few genes, have occurred extensively in eukaryotic nuclear genomes and should provide a plethora of phylogenetic markers in the future.

2.10 A concluding mix of caution and optimism

One obstacle that makes some researchers feel uneasy about the use of RGCs is the absence of statistical evaluation (Philippe and Laurent, 1998; Hillis, 1999). This concern stems primarily from analogy with primarily sequence comparisons. Understanding the forces that shape sequence evolution is a necessary prerequisite to

using sequence data for phylogenetics and for evaluating the statistical robustness of trees. To reach the same level of sophistication in the analysis of RGCs demands greater knowledge about the mechanisms that generate RGCs because this will affect their rate of production, character independence, mutational biases and reversibility. Some of these parameters are reasonably well understood for some RGCs (notably SINE insertions and gene losses), but there is much more to be learnt. Nonetheless, the usefulness of several categories of RGC has been tested by comparison with phylogenies inferred by other methods (morphological and molecular). With few exceptions, RGCs have performed exceptionally well. Therefore, we feel that 'psychological constraints' (Philippe and Laurent, 1998) about statistical evaluation should be put aside, while applicability and robustness are tested further. Additionally, we stress that the 'Hennigian' framework is not the only one that can be employed when attempting to reconstruct a phylogeny based on RGCs. A statistical approach is also possible (e.g. maximum likelihood, Bayesian analysis) and will surely be of help, especially as more 'messy' datasets are obtained.

It is an inescapable (if uncomfortable) fact that a few good characters might contain more phylogenetic 'truth' than many poor ones. We do not suggest that all RGCs are necessarily 'good' markers; we certainly do not propose that they are a panacea for phylogenetics. Indeed, we have already alluded to cases of convergence, parallelism and reversion. However, we do believe that the suite of characters that we refer to as RGCs harbours enormous potential. They have already contributed some robust insights into important phylogenetic debates, such as the origin of whales, arthropod relationships, deuterostome phylogeny and diversification of the protostome invertebrates. Each of these 'new' phylogenies has wider implications, not only for evolutionary biology but also for biogeography, developmental biology and other areas. In a time when the acquisition of molecular data is outpacing analysis, it is worth recalling Darwin's comment (1859): '...we possess no pedigrees or armorial bearings; and we have to discover and trace the many diverging lines of descent in our natural genealogies, by any characters which have long been inherited'.

2.11 Acknowledgements

A version of this chapter has already been published: Rokas, A. and Peter W.H. Holland, 2000. Rare genomic changes as a tool for phylogenetics. *Trends in Ecology & Evolution* **15:** 454-459 (Appendix I).

This review would not have been written without the help and insight of Peter Holland, in whose writings I first came across the ideas expressed in this chapter. N. Okada and two anonymous referees made valuable comments on an earlier version of the manuscript. Special thanks for providing photographs to G. Brown, S. Ferguson, M. Kobayashi and J. Pemberton. P. Preston very kindly allowed the use of specimens from the Natural History collections of the University of Edinburgh for photography; we acknowledge Marine Biological Laboratory and D. Remsen for permitting reproduction of the hemichordate photograph.

CHAPTER 3

A maximum likelihood analysis of eight phylogenetic markers in gallwasps (Hymenoptera: Cynipidae); implications for insect phylogenetic studies

3.1 Abstract

We assessed the utility of eight DNA sequence markers (5.8S rDNA, 18S rDNA, 28S rDNA, ITS regions, long-wavelength opsin, elongation factor 1-α, cytochrome *b* and cytochrome oxidase I) in reconstructing phylogenetic relationships at various levels of divergence in gallwasps (Hymenoptera: Cynipidae), using a set of eight exemplar taxa. We report sequence divergence values, saturation levels, and compare phylogenetic results of these sequences analysed both separately and combined to a well-corroborated morphological phylogeny. Likelihood ratio tests were used to find the best evolutionary model fitting each of the markers. The likelihood model best explaining the data is, for most loci, parameter-rich, with strong A-T bias for mitochondrial loci, and strong rate heterogeneity for the majority of loci. Our data suggest that 28S rDNA, elongation factor 1-α and long-wavelength opsin may be potentially useful markers for resolving cynipid and other insect within-family divergences (circa 50-100 mya old), whereas mitochondrial loci and ITS regions are most useful for lower-level phylogenetics. In contrast, the 18S rDNA marker is likely to be useful for resolving above-family-level relationships.

3.2 Introduction

Gallwasps are a family of wasps (Hymenoptera: Cynipidae) that parasitise herb and tree

species in the Palaearctic and Nearctic regions, inducing gall formation (Table 3.1) (Askew, 1984). A number of species within the family (members of the tribe Synergini) have lost the ability to induce galls themselves but instead develop inside the galls of other cynipids; these are called inquilines (Table 3.1) (Ronquist, 1994). Little is known about cynipid phylogenetic relationships and few studies have dealt with the issue (Ronquist, 1994; Liljeblad and Ronquist, 1998; Stone and Cook, 1998). There are 1369 described species within the Cynipidae, currently divided, on the basis of morphological data, into six tribes, listed in Table 3.1 (Liljeblad and Ronquist, 1998). Biogeographic and fossil evidence suggest that cynipids originated at least as long ago as the mid Cretaceous (83 mya) (Ronquist, 1999). Fossil data also suggest that major groups in one tribe, the inquiline Synergini, diverged at least 45 mya (Fig. 3.1) (Ronquist, 1999, Z. Liu, pers. comm.). This dating also suggests that the woody rosid gallers (Fig. 3.1), as obligate hosts to the inquilines, must also have diverged at least 45 mya. The evolutionary age of Cynipidae makes them an appropriate model taxon for testing models of sequence evolution and the utility of molecular markers for family-level insect phylogenetics.

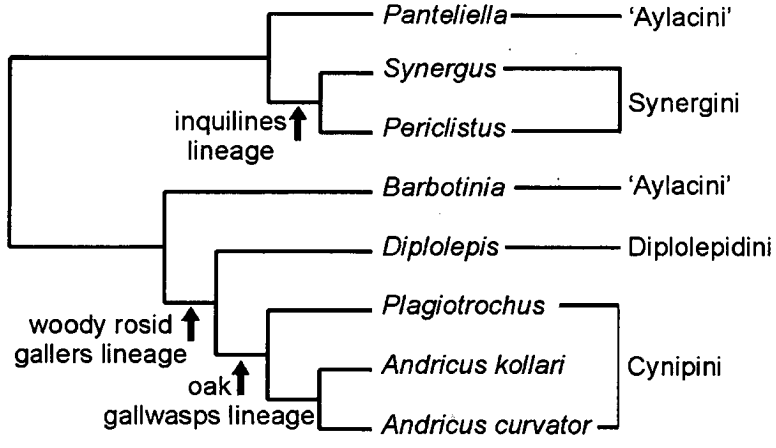


Figure 3.1. The morphology-based phylogeny of the eight exemplar species used in the present analysis. The topology was generated using parsimony (Liljeblad and Ronquist, 1998). The tribes in which the taxa belong are also mentioned. Arrows indicate lineages discussed in the text.

Models of nucleotide substitution are important for estimation of evolutionary trees and for understanding of the evolutionary processes of DNA sequences (Yang, *et al.*, 1994; Swofford, *et al.*, 1996). While even the best currently available models do not describe the evolution of DNA sequences perfectly (Goldman, 1993), it is well documented that better models lead to more accurate estimates of the evolutionary history of the species concerned and to a better understanding of the forces and mechanisms that affected the evolution of the sequences (e.g. Yang, *et al.*, 1994; Swofford, *et al.*, 1996; Huelsenbeck and Rannala, 1997; Lewis, 1998). Therefore, correct estimation of the parameters involved in the construction of models of sequence evolution (such as rate heterogeneity among sites, base composition and types of substitution) is an important task and statistical methods such as maximum likelihood (ML) allow explicit evaluation of parameters involved in phylogenetic estimation (Swofford, *et al.*, 1996; Lewis, 1998). On a more general note, the estimation of a molecular phylogeny depends on achieving a match between the mutation rate of the marker selected, and the time for which the selected lineages have been diverging. A marker with a low mutation rate may evolve too slowly to resolve relationships in rapidly diversifying lineages. Similarly, a rapidly evolving marker will become mutationally saturated over larger time scales (due to the effect of multiple substitutions) and so will be a poor estimator of a phylogeny.

Table 3.1. Overview of the diversity of gallwasps (Hymenoptera: Cynipidae). From Ronquist (1999). [1] inquilines are cynipids that have lost the ability to induce gall formation on their own but develop inside the galls of other cynipids.

Tribe	Genera	Species	Biology	Taxa used in this study
Synergini	7	171	Phytophagous inquilines[1] in galls of other cynipids	*Synergus gallaepomiformis* *Periclistus brandtii*
Aylacini	21	156	Mostly gallers on eudicot herbs. This tribe is	*Barbotinia oraniensis* *Panteliella bicolor*

			paraphyletic (indicated by the inverted commas in Fig. 1).	
Diplolepidini	2	63	Gallers on *Rosa*	*Diplolepis rosae*
Eschatocerini	1	3	Gallers on *Acacia* and *Prosopsis* (Fabaceae)	-
Pediaspidini	2	2	Gallers on *Acer*	-
Cynipini	44	974	Gallers on Fagaceae and Nothofagaceae, mostly on *Quercus*	*Plagiotrochus quercusilicis* *Andricus kollari* *Andricus curvator*

In groups in which sequence and genome evolution have been well studied (e.g. hominoids, *Drosophila*), the task of selecting loci for phylogeny reconstruction is relatively straightforward, since the evolutionary 'behaviour' of a number of loci is well characterised (e.g. Johns and Avise, 1998). Additionally, primer design for the exploration of new markers is easy, cost- and information-wise, since data from genome projects are already available. However, for the vast majority of less-studied taxonomic groups (such as gallwasps and arthropods more generally) such information is unavailable. When reconstructing a molecular phylogeny for such a group, it has become common practice to use one (or two) of a set of gene regions already applied with success in similar circumstances (for a review see Caterino, *et al.*, 2000). This practice is generally justified because of its success in most cases (e.g. Hillis, *et al.*, 1996), owing to the existence of a large suite of conserved primers (Brower and DeSalle, 1994; Simon, *et al.*, 1994; Palumbi, 1996) and the standardisation of techniques associated with PCR amplification and sequencing (Hillis, *et al.*, 1996). In insect phylogenetics, the most frequently used loci have been the cytochrome oxidase subunits (e.g. Beckenbach, *et al.*, 1993; Crespi, *et al.*, 1998), 16S ribosomal DNA (e.g. Whitfield and Cameron, 1998) and cytochrome *b* (e.g. Jermiin and Crozier, 1994; Stone and Cook, 1998) from the mitochondrial genome, and the ribosomal DNA array (Hillis and Dixon,

1991) and elongation factor 1-α (Cho, *et al.*, 1995; Danforth and Ji, 1998) from the nuclear genome. Additionally, a number of researchers have identified further loci useful for arthropod systematics (Friedlander, *et al.*, 1992, 1994; Cho, *et al.*, 1995; Mardulyn and Cameron, 1999).

Despite the success of this general approach, few studies have directly compared the utility of commonly used loci across a range of taxonomic levels in a single clade (e.g. Hillis and Dixon, 1991; Zardoya and Meyer, 1996). In this chapter we attempt to identify which of the currently popular markers would be the best for the task of reconstruction of the gallwasp tree at various taxonomic levels. For this task we have sequenced eight markers for eight exemplar species. Our primary aim is not to create a robust phylogeny *per se*, so much as to identify which markers are potentially useful at a particular taxonomic level. This study may serve as a test case (i.e. which genes should be tried first at a particular taxonomic level) for less-studied groups of insects in general (such as gallwasps). We suggest our results may be of interest to those workers on parasitoid wasps in the sister groups of cynipids, themselves the object of increasing numbers of evolutionary studies (e.g. Dowton and Austin, 1995; Mardulyn and Whitfield, 1999). For reasons outlined above, we have decided to adopt a statistical approach (maximum likelihood) for the analysis of our dataset. This is not the only way; for example, Graybeal (1994) employed a cladistic approach in an attempt to find genes informative about deep divergences in the vertebrate lineage. Caveats of our approach are also discussed.

All but one of the six nuclear and two mitochondrial loci employed here have been used frequently to reveal phylogenetic relationships for a wide variety of taxonomic groups (for reviews see Brower and DeSalle, 1994; Simon, *et al.*, 1994; Caterino, *et al.*, 2000). Specifically, we use the nuclear loci elongation factor 1-α and the ribosomal DNA array which includes 5.8S rDNA, 18S rDNA, 28S rDNA and Internal Transcribed Spacer regions (ITS1 and ITS2, treated as a single locus), as well as the mitochondrial loci cytochrome *b* and cytochrome oxidase I. Additionally, we have sequenced a nuclear locus recently proposed to be useful in insect phylogenetics, long-wavelength opsin

(Mardulyn and Cameron, 1999). In total, we have generated more than 7500 base pairs of aligned nucleotide sequence from eight gallwasp species which were selected for their position in the existing wider morphological and molecular phylogenies of Cynipidae (Liljeblad and Ronquist, 1998; Stone and Cook, 1998). In this paper we attempt: (1) to identify the best-fit evolutionary model for each locus using a statistical framework (maximum likelihood) and compare the evolutionary 'behaviour' of each locus, and, (2) to assess the phylogenetic utility of markers by describing sequence divergence across different time scales and by comparing the phylogenetic results (obtained from the markers) to a morphology-based phylogeny.

3.3 Materials and Methods

3.3.1 Selection of species for molecular analysis

Our selection of cynipid species for molecular analysis was primarily based on the morphological character-based phylogeny of Liljeblad and Ronquist (1998), with additional information from the mitochondrial cytochrome *b* phylogeny of oak gallwasps from Stone and Cook (1998). Our criteria for selection of taxa were: (1) selected taxa should represent all the major clades of the gallwasp family, and (2) taxa should exhibit a range of phylogenetic distances from each other. The selection of species is shown in Table 3.1 and their phylogeny, according to morphology, is shown in Figure 3.1. We did not use members of two tribes (Eschatocerini and Pediaspidini) because of their low species richness (see Table 3.1). We did not use an outgroup of Cynipidae since the focus of this paper is more on the quality of the markers rather than in obtaining the most correct phylogeny for this set of taxa.

3.3.2 Selection of genes, amplification and sequencing

Amplification and sequencing of all loci was done by PCR using primers either

previously published and/or of our design. Details of the primers used, PCR conditions and fragments amplified are listed in the Appendix 3.1. In short, we amplified fragments of 18S rDNA, 28S rDNA, 5.8S rDNA, Internal Transcribed Spacer regions (ITS1 and ITS2), elongation factor 1-α F1 (EF1α F1), long-wavelength opsin (LW *Rh*), cytochrome *b* (Cyt*b*) and cytochrome oxidase I (COI). Sequencing was either performed directly on the PCR product or the PCR fragment was first cloned and subsequently sequenced, using previously described methodology (Rokas, *et al.*, 2001a, Chapter 4, Appendix I; Stone, *et al.*, 2001). Intra-individual variation for most genes was non-existent. In the case of the ITS region, intra-individual variation was observed, but detailed analyses in a single gallwasp species suggest that this will not obscure inter-specific relationships (Rokas, *et al.*, 2001a, Chapter 4, Appendix I). All sequencing reactions were done twice to minimise PCR artefacts, ambiguities and base-calling errors. Sequencing was carried out using the Perkin-Elmer BigDye Terminator chemistry and an ABI 377 sequencer.

3.3.3 Sequence analysis

Each locus was aligned using CLUSTALW (Thompson, *et al.*, 1994) with the default settings options. The alignment generated a total sequence of 7685 base pairs (bp) per specimen. Because it was impossible to obtain an accurate alignment for certain parts of the ITS1 and ITS2 region, a dataset was manually assembled for the ITS1-ITS2 fragment consisting only of the conserved regions. We will refer to this fragment as ITS. After this adjustment, the complete aligned data set consisted of 5908 bp. All the alignments used in this study are available electronically from TreeBASE (http://www.herbaria.harvard.edu/treebase/, TreeBASE Study Acc. No.: S645 - see Appendix 3.1 for matrix accession numbers of individual datasets).

Phylogenetic analysis was performed using maximum likelihood (ML) algorithms, which allow hypothesis-testing in a statistical framework (e.g. Huelsenbeck and Rannala, 1997; Lewis, 1998) and description of important aspects of sequence evolution,

such as rate heterogeneity, transition/transversion rate ratios and compositional bias. These are important parameters widely used in phylogenetic reconstructions based on molecular data (for a review see Lewis, 1998).

Each locus was analysed separately. Comparisons between mtDNA loci (COI and Cyt*b* combined), ribosomal DNA loci (rDNA array: 18S, 28S, 5.8S and ITS regions combined), nuclear loci (rDNA array, LW *Rh* and EF1α F1) and all data sets combined (all loci) were also made. Gaps were coded as missing. The best-fit ML model for each locus as well as for the complete data set was identified using MODELTEST 3.0 (Posada and Crandall, 1998). The parameters allowed to vary in model-fitting were base composition, substitution rates (which includes variation in transition/transversion - ti/tv - ratio) and rate heterogeneity across sites (both by the invariable-sites model and the gamma-distributed rates model). MODELTEST utilises likelihood ratio tests (Huelsenbeck and Rannala, 1997; Lewis, 1998) to identify the ML model of sequence evolution on an initial, approximate tree (Posada and Crandall, 1998), since it has been shown that estimation of ML parameters is not very sensitive in regard to the tree topology on which they are estimated (Yang, *et al.*, 1994). Tree reconstruction was performed using ML as implemented in the package PAUP* (Swofford, 2000). ML searches were performed using the branch and bound algorithm on 100 bootstrap replicates using the ML values suggested by MODELTEST. All subsequent measures of sequence divergence were estimated using the same parameters.

The ML topologies generated from the 'mtDNA loci' (COI and Cyt*b* combined), 'nuclear loci' (all the nuclear loci combined) and 'all loci' (all the loci combined) datasets were compared with the morphological phylogeny using the SOWH test (Goldman, *et al.*, 2000). The SOWH test is a likelihood-based test for comparison of tree topologies which are not specified *a priori* (as is the case for an alternative, the Kishino-Hasegawa test) (for discussion see Goldman, *et al.*, 2000). The principle behind the SOWH test is the generation of a null distribution for the difference in likelihood scores between the two topologies (using parametric bootstrap analysis) and the testing of the observed data against this distribution (Goldman, *et al.*, 2000). SOWH tests were

performed using the software packages PAUP* (Swofford, 2000) and SEQGEN (Rambaut and Grassly, 1997). Null distributions were generated using 100 simulations.

Mutational saturation was investigated using the simple visual method proposed by Philippe *et al.* (1994b), which consists of plotting the proportion of the observed (uncorrected) differences between pairs of species as a function of the estimated (in this case the estimation made using ML) proportion of differences for the same species pairs. Observed differences initially increase linearly with estimated differences, but as the sequences under study become saturated, observed difference approaches an asymptote and changes little with increasing estimated difference. The transition from linear increase to asymptote indicates the onset of saturation; while additional substitutions are happening (as indicated by the raise in the estimated proportion) they are not actually observed (as indicated by the asymptote approached by the observed proportion). Beyond this point, estimated differences continue to increase substantially while observed differences increase very little.

3.4 Results

3.4.1 Aspects of molecular evolution for the 8 loci

3.4.1.1 Rate heterogeneity

Failure to account for rate heterogeneity (testing whether substitution rates over all nucleotide sites are constant or not) can have serious effects on phylogenetic estimation (Yang, *et al.*, 1994; Yang, 1996). The two most commonly used methods for explicitly dealing with rate heterogeneity are the invariable-sites model (Palumbi, 1989), in which some proportion of sites is assumed to never change, with all variable sites assumed to evolve at the same rate, and the gamma-distributed-rates model, in which the distribution of relative rates over sites is assumed to follow a gamma distribution (Yang, 1994) whose shape parameter α determines the strength of rate heterogeneity. Both models were tested (individually and combined) for statistically significant improvement

in the likelihood score. Only the 5.8S locus showed no rate heterogeneity (Table 3.2). For most loci (with the exceptions of 18S, the rDNA array and all data, which required both models) the gamma-distributed-rates model adequately explained rate heterogeneity along a locus and addition of the invariable-sites model did not significantly improve the ML model (Table 3.2).

3.4.1.2 Variation in base composition across loci

Individual loci of the rDNA array (18S, 28S, 5.8S and ITS) are the only ones in which a model assuming equal base frequencies is supported. For all the other loci (including the whole rDNA array data set), models allowing unequal base frequencies provided a significantly better fit to the data. At the other extreme, mtDNA loci verify the general observation from insects and especially Hymenopterans of a strong A-T bias (Crozier and Crozier, 1993; Dowton and Austin, 1995; Whitfield and Cameron, 1998). The two mtDNA loci of Cynipidae are 74.85% A-T rich (32.87% A, 12.20% C, 12.96% G, 41.97% T). These frequencies agree with the cytochrome oxidase I data of Dowton and Austin (1995), showing a higher A-T content in parasitic wasps (Apocrita; A-T content: 74.0 ± 0.7 % - gallwasps belong here) compared with non-parasitic wasps (Higher Symphyta; A-T content: 70.7 ± 0.7%).

locus	ML model	Base frequencies	No. of substitution rates[a]	Rate heterogeneity models		Sequence length
				a. Invariable-sites	b. gamma-distributed-rates	
18S rDNA	TrNef+I+G	equal	3 – 2 ti, 1 tv	0.9195	0.7167	1798
28S rDNA	K80+G	equal	2 – 1 ti, 1 tv	0	0.0768	1073
5.8S rDNA	JC69	equal	1	0	∞	122
ITS	TrNef+G	equal	3 – 2 ti, 1 tv	0	0.4491	557[b]
Ef1α F1	TrN+G	unequal	3 – 2ti, 1 tv	0	0.27	367
LW *Rh*	GTR+G	unequal	6 – 2 ti, 4 tv	0	0.4158	481
COI	TVM+G	unequal	5 – 1 ti, 4 tv	0	0.2917	1077
Cyt*b*	TVM+G	unequal	5 – 1 ti, 4 tv	0	0.3844	433
rDNA array	TrN+I+G	unequal	3 – 2 ti, 1 tv	0.7227	0.7664	3550
mtDNA loci	TVM+G	unequal	5 – 1 ti, 4 tv	0	0.3131	1510
Nuclear loci	GTR+I+G	unequal	6 – 2 ti, 4 tv	0.7024	0.6077	4398
All loci	GTR+I+G	unequal	6 – 2 ti, 4 tv	0.4928	0.5281	5908

Table 3.2. Best fitting ML models and their estimated parameter values for 8 loci and their combined datasets. The sequence length of each locus is also shown. Short name descriptions of the ML models are according to Posada and Crandall (1998). [a] ti: number of transition types, tv: number of transversion types. [b] This fragment contains only the alignable regions.

Lineage (→) Locus (↓)	Cynipidae (observed / estimated)	oak gallwasps (observed / estimated)	woody rosid gallers (observed / estimated)	Inquilines (observed / estimated)
18S rDNA	0.06-1.17 / 0.06-1.73	0.06-0.17 / 0.06-0.18	0.06-0.7 / 0.06-0.9	0.8 / 1.1
28S rDNA	0.28-8.89 / 0.29-22.99	0.28-4.18 / 0.29-7.13	0.28-8.89 / 0.29-22.99	4.83 / 7.97
5.8S rDNA	0-3.39 / 0-3.52	0 / 0	0-1.69 / 0-1.75	2.46 / 2.53
ITS	1.33-16.69 / 1.36-26.94	1.33-7.22 / 1.36-8.74	1.33-16.88 / 1.36-25.52	14.46 / 22.47
Ef1α F1	1.91-12.81 / 2.07-21.28	1.91-6.54 / 2.07-8.87	1.91-11.72 / 2.07-18.28	8.17 / 11.23
LW *Rh*	3.53-17.82 / 3.89-29.87	3.53-10.6 / 3.89-14.26	3.53-17.41 / 3.89-28.6	9.77 / 12.77
COI	6.68-20.15 / 8.68-43.89	6.68-13.11 / 8.68-21.94	6.68-18.89 / 8.68-39.23	19.77 / 42.87
Cyt*b*	6.7-22.97 / 8.31-53.90	6.7-13.16 / 8.31-19.72	6.7-21.36 / 8.31-46.07	20.55 / 39.87
rDNA array	0.31-5.72 / 0.32-8.65	0.31-2.43 / 0.32-2.93	0.31-5.51 / 0.32-8.00	4.12 / 5.61
mtDNA loci	6.69-20.77 / 8.55-44.76	6.69-13.12 / 8.55-21.28	6.69-19.38 / 8.55-40.86	20.0 / 41.83
Nuclear loci	0.80-7.45 / 0.85-12.63	0.80-3.68 / 0.85-4.85	0.80-7.45 / 0.85-12.63	5.09 / 7.68
All loci	2.32-10.54 / 2.6-17.62	2.32-5.94 / 2.6-7.89	2.32-10.54 / 2.6-17.39	8.94 / 13.8

Table 3.3. Percent sequence divergence within Cynipidae. Estimation of sequence divergences was performed using ML with the parameter values specified in Table 3.2.

3.4.1.3 Rates of different substitution types

In the simplest case (one-parameter model of Jukes and Cantor) all nucleotide substitutions occur at the same rate (see review by Swofford, *et al.*, 1996). From the one-parameter model, more complex models can be constructed if rates of substitution are free to vary (e.g. by allowing different rates for transitions and transversions, etc.). In the most general (parameter-rich) case (represented by the General Time Reversible model - GTR) all possible 6 substitution types – two transitions (A↔G, T↔C) and four transversions (A↔T, A↔C, C↔G, G↔T) – are free to occur at a different rate. The model and number of different substitution types utilised by each of the loci are shown in Table 3.2. Only the 5.8S rDNA data are best explained by a single substitution rate; a result perhaps due to small length of the locus. Most loci are not adequately represented by the simplest model, and require different substitution rates (Table 3.2). Note that in our analysis we assumed that substitution rates are stable among evolutionary lineages (but see Yang and Yoder, 1999).

3.4.1.4 Mutation rates and saturation rates across loci

The graphs depicting the proportion of observed versus estimated substitutions for each locus (using the ML parameters estimated from MODELTEST) are shown in Fig. 3.2. In this data set certain loci become saturated more quickly than others, as shown by the divergence ranges in Table 3.3. The most rapidly saturated loci are the mitochondrial ones (mtDNA – see Fig. 3.2c), which also show the highest divergences across the 8 sampled taxa (the percentage for the most divergent pair of sequences is 44% and 54% for COI and Cyt*b* respectively – see Table 3.3). In contrast, the nuclear loci (Fig. 3.2b, d-f) saturate more slowly and show a range of rates of divergence, with the rDNA array being the slowest (8.65%) and LW *Rh* the fastest (29.87% - see Table 3.3).

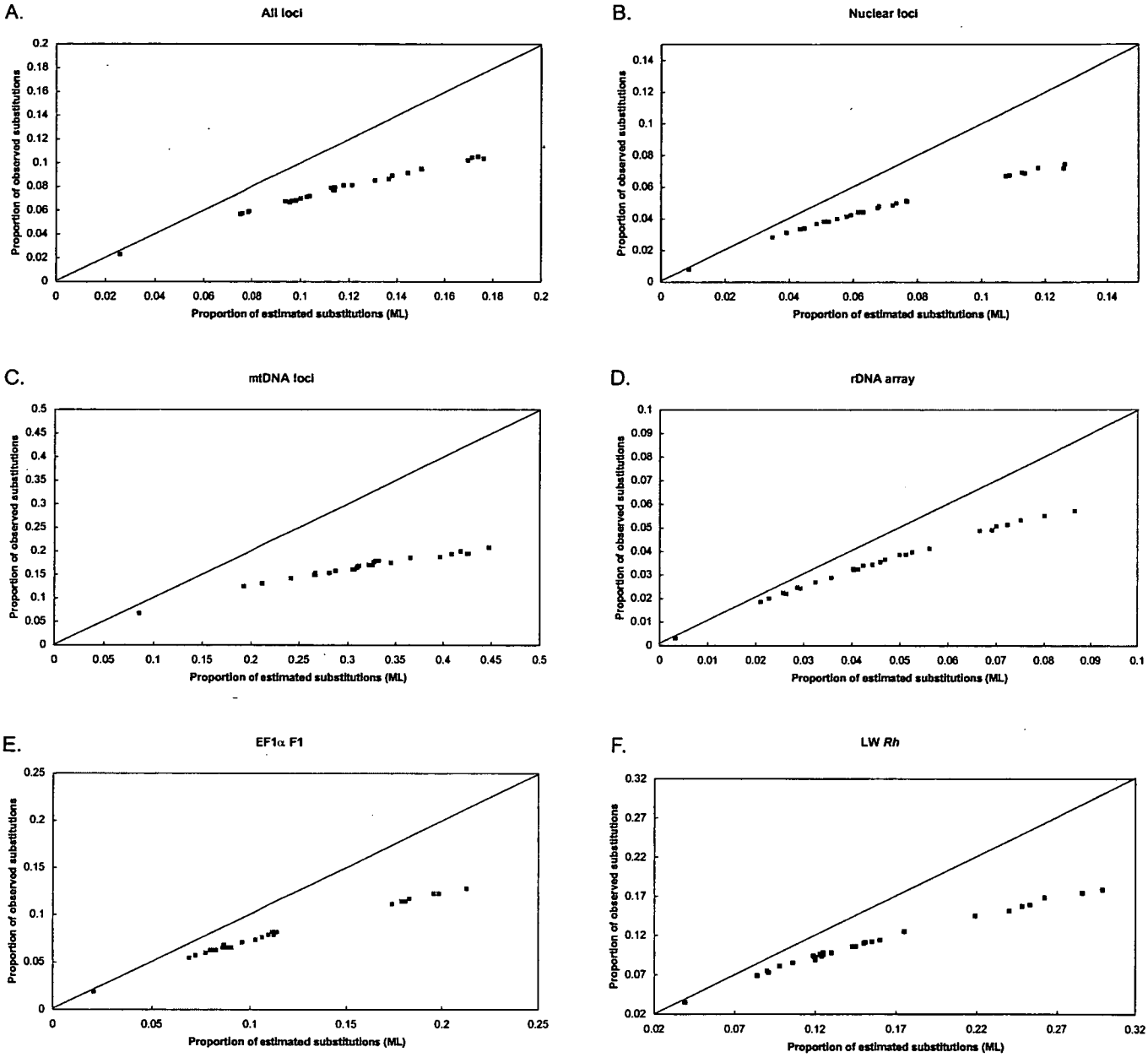


Figure 3.2. Saturation plots of the proportion of uncorrected (observed) sequence divergence on the Y-axis versus the proportion of the estimated sequence divergence on the X-axis, for a. All loci, b. Nuclear loci, c. mtDNA, d. rDNA array, e. Ef1α F1 and f. LW *Rh*. Mitochondrial loci (COI and Cyt*b*) are shown combined since the plots for each locus look very similar; for the same reason, saturation levels of the rDNA array loci are also shown in a single plot. Estimated sequence divergence values were calculated in the phylogenetic package PAUP* (Swofford, 2000), using the ML model suggested by MODELTEST (Posada and Crandall, 1998).

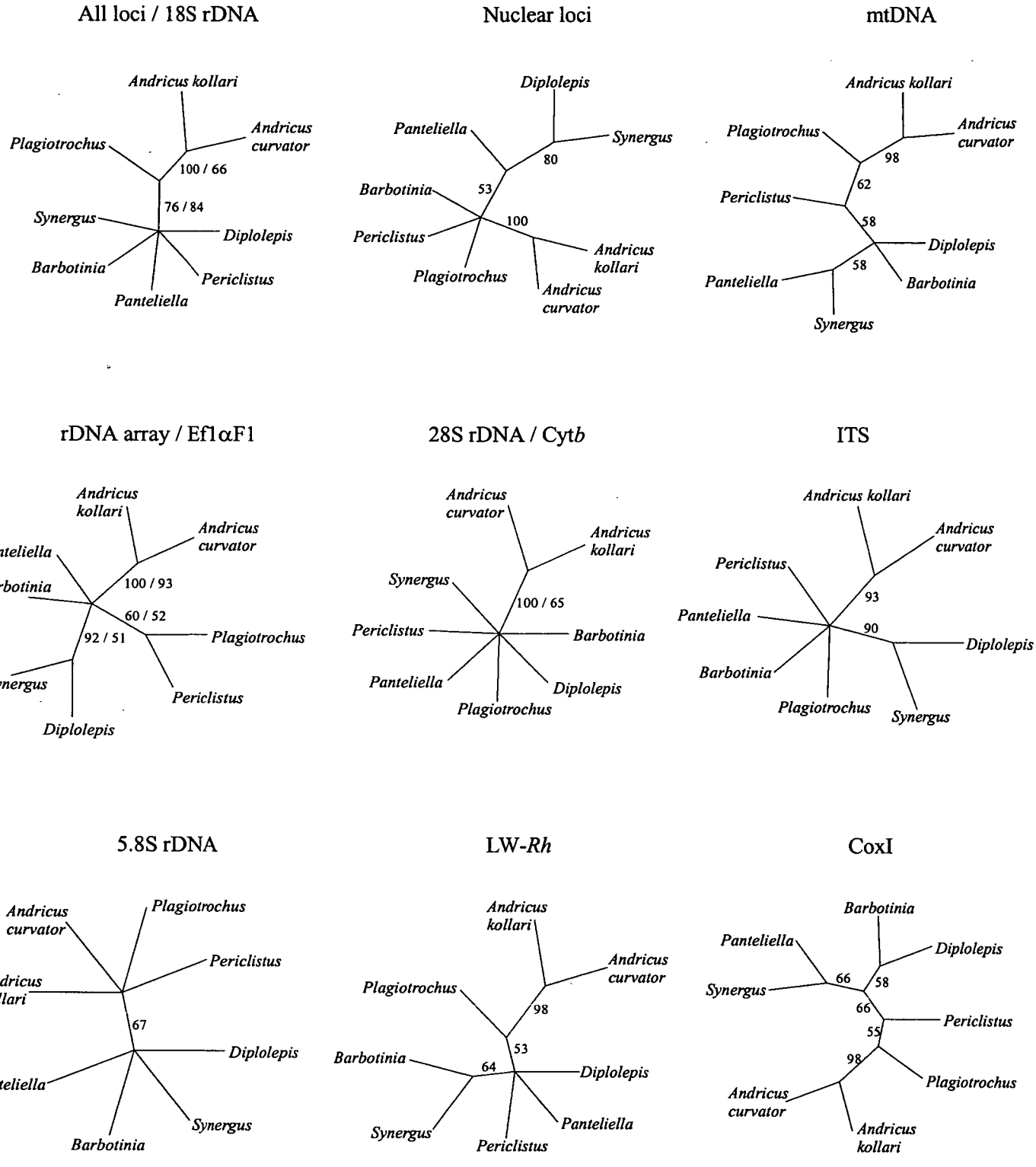


Figure 3.3. 50% majority rule consensus unrooted trees for all data sets. Numbers in branches denote bootstrap support (100 pseudo-replicates using ML and a branch & bound algorithm).

3.4.2 Gene and species phylogenies

The consensus topologies obtained from ML reconstruction of 100 bootstrap replicates are shown in Fig. 3.3. Comparison of the topologies generated by the different loci and combined datasets shows that the generated trees are substantially different. The only clade supported by all datasets, and in agreement with morphology, is that uniting the two *Andricus* species. Most datasets also support the oak gallwasp lineage (*Plagiotrochus* and the two *Andricus* species) (Figs. 3.1, 3.3). However, the molecular datasets offer no support for monophyly of morphologically well-established clades (according to Liljeblad and Ronquist (1998)) such as the woody rosid gallers (*Diplolepis* and oak gallwasps), the inquilines (*Periclistus* and *Synergus*) or the *Panteliella-Periclistus-Synergus* lineage (Figs. 3.1, 3.3).

Statistical significance of differences among our (molecule-generated) topologies and the morphological phylogeny of Liljeblad and Ronquist (1998) was tested with the SOWH test (Goldman, *et al.*, 2000). Due to computational limitations, comparisons were only made between the morphological phylogeny and the three following datasets: mtDNA (all the mitochondrial loci), nuclear loci (all the nuclear loci) and all loci (all the data sets combined). Given the all loci data set, the morphological phylogeny and the all loci ML topology gave significantly different likelihood scores (Δ=28.9312, p<0.01), with the morphological topology providing a significantly worse fit for this data set. Significant differences were also found when the morphological topology was compared with ML topologies generated under the mtDNA and the nuclear loci data sets (Δ=14.1253, p<0.01, and Δ=18.6775, p<0.01, respectively).

3.4.3 Which gene for what taxonomic level?

The percent sequence divergence within Cynipidae and within various morphologically well-established clades within the family (oak gallwasps, woody rosid gallers, inquilines) are shown in Table 3.3.

3.5 Discussion

3.5.1 Aspects of molecular evolution for the 8 loci

3.5.1.1 Rate heterogeneity

Apart from the 5.8S locus, in which all sites are evolving at the same rate (very probably because of its small length), best-fit models for all other loci require a parameter allowing rate heterogeneity (Table 3.2). Rate heterogeneity in most loci is adequately explained by the gamma-distributed-sites model. In the few cases in which the invariable-sites model is used, it is in addition to the gamma-distributed-sites model (Table 3.2). This result is intuitive since while both models are designed to explain rate heterogeneity, the gamma-distributed-rates model is more parameter-rich, and so allows a better fit to the data than invariable-sites alone. However, this comes at the cost of increased computational effort. The shape parameter α of the gamma-distributed-rates model is small in most cases denoting strong rate heterogeneity (Table 3.2). This is concordant with results from other studies (Yang, *et al.*, 1994; for a review see Yang, 1996). We note however that our conclusion should be taken with some caution given some recent results suggesting that in certain cases estimates of rate heterogeneity might be sensitive to taxon sampling (Sullivan, *et al.*, 1999).

3.5.1.2 Variation in base composition across loci

Variation in base composition is important in modelling sequence evolution because of its effect in reducing the number of character-states for a given site. For example, the extreme A-T bias observed in insect mitochondrial genomes reduces many sites from four-state (A, C, G or T) to two-state characters (A or T), the consequence being that these sites become saturated more quickly than others. Variation such as this observed in the base composition of mtDNA genes among Hymenoptera seriously reduces the utility of mtDNA genes (e.g. Whitfield and Cameron, 1998; Dowton and Austin, 1999), at least

for higher-level phylogenetics, due to fast mutational saturation (see below). In contrast, the nuclear loci show smaller deviations from equal base frequencies (even those that show variation in base composition) and are less affected by the problem of character-state reduction.

3.5.1.3 Rates of different substitution types

It is interesting to note that certain widely-used substitution models such as Kimura's (1980), which assumes one rate for all types of transitions and one rate for all types of transversions, are not utilised frequently by the loci analysed here (only the 28S rDNA data set fits the assumptions of Kimura's model). From Table 3.2 it is evident that genes are idiosyncratic in terms of substitution types and that there is considerable variation within transitions or transversions. For example, the rDNA array is best explained by a ML model with one transversion rate and two transition rates, in contrast to the mtDNA loci which are best explained by a model containing one transition rate and four transversion rates (see Table 3.2 for more details).

3.5.1.4 Mutation rates and saturation rates across loci

The plotting of observed versus estimated values of sequence divergence can be used as a way to select genes that may give the correct topology with most methods, by simply selecting these loci that exhibit low saturation levels at the level of divergence exhibited by the sampled taxa. However, saturation plots cannot be taken as *prima facie* predictors of data quality. First, there must be adequate sequence divergence for the phylogeny to be resolved. Second, at the other end of the mutation rate continuum, the phylogenetic performance of apparently saturated loci, or partitions of a locus (e.g. codon positions), is dependent on taxonomic sampling (cf. Hillis, 1996; Björklund, 1999). Nevertheless, in a study like this (i.e. with limited taxonomic sampling), this approach gives valuable information on the relative evolutionary rates for the compared loci. For example, it is clear that mitochondrial loci become saturated most rapidly - perhaps partly due to their

strong A-T bias - and hence they might not be the best first candidate loci for higher-level phylogenetics or that nuclear loci are less saturated.

3.5.2 Gene and species phylogenies

The significant differences in likelihood scores observed between the morphological topology and the three molecular topologies (all loci, nuclear loci and mtDNA loci) indicate a conflict between morphology and molecules. However, given that most of the consensus trees for the loci (Fig. 3.3) agree (partially or completely) with the morphological phylogeny and that our molecular analyses have used relatively few species (especially compared with the morphological analysis), we don't want to emphasise this conflict, until more data are generated.

An interesting feature of the consensus trees of most loci presented in Fig. 3.3 is their poor resolution, which may be explained in two ways. The first is that the topologies might represent what has really happened, i.e. an adaptive radiation. If the rate of speciation for a given time window is relatively high (emergence of many new species in a small amount of time), inter-specific differentiation will be low. This has as a consequence a poorly-resolved phylogeny, which however is an accurate depiction of the historical associations between the taxa analysed. Liljeblad and Ronquist (1998) alluded to this scenario for gallwasps, after having difficulty in resolving basal cynipid relationships. A rapid early radiation of gallwasps represents an attractive scenario, since the radiation might have been correlated with the transition to a new adaptive zone (the evolution of the gall-induction mechanism being the key adaptation). Such a scenario would predict not only poorly-resolved topologies, but also similar sequence divergence estimates between the major cynipid clades. Data shown in Table 3.3 suggest that distances for woody rosid gallers and inquilines are similar (although major exceptions are LW *Rh* and 28S) to each other as well as to estimates for the whole family of Cynipidae (these comparisons are valid assuming that the assemblages of the inquilines, the woody rosid gallers and the Cynipidae are monophyletic). Although more data are

needed, these preliminary results are intriguing. A point that should also be considered is that when we are attempting to resolve the phylogenetic history of an adaptive radiation, where divergence times between taxa in a well sample data matrix range between millions of years, it is likely that a single locus will not be effective at resolving all the nodes. A combined approach, employing several loci which diverge at various rates, is probably the only way to obtain well supported phylogenetic hypotheses (Hillis, *et al.*, 1996).

A second explanation for lack of resolution might be the absence of data (be it characters or taxa). Certain clades are resolved in most cases (e.g. the oak gallwasps) although this is not true for the inquiline lineage or for higher clades (Fig. 3.3), and the addition of more taxa and/or characters might resolve the currently observed polytomies.

3.5.3 Which gene for what taxonomic level?

Selecting a gene for phylogenetic analysis requires matching the level of sequence variation to the desired taxonomic level of study (i.e. adequate sequence diversity to resolve taxonomic affinities, but minimal artefacts due to saturation). Because of the arbitrariness of taxonomic categories (there is no guarantee that a genus of beetles is the same age as a genus of aphids and even less that it is the same age as a genus of fish), generalisations about the taxonomic rank at which particular genes might be useful should be made with caution. Families across insect orders probably show a wide range of genetic divergences (although to our knowledge there is no review available on this topic) as shown in other groups (Johns and Avise, 1998; Avise and Johns, 1999). Reasons for this diversity at a given taxonomic level include variation among lineages in age, rate of evolution (molecular or morphological) or simply in the alpha-taxonomist's philosophy (the 'splitters versus lumpers' debate) (Johns and Avise, 1998). Furthermore, a given protein gene may vary in rate from one taxon to the next (Gillespie, 1986; Crozier, *et al.*, 1989; Jermiin and Crozier, 1994). Nevertheless, generalisations can be made about the comparative evolutionary rate of specific genes. Divergence time, if it is

known (e.g. from fossils), is another predictor of degree of genetic divergence (usually better than taxonomic rank) (Simon, *et al.*, 1994) and will also be briefly discussed. What follows is a preliminary comparison (the subject is too large to be fully discussed here) of published findings that have employed the same markers for other hymenopterans and, more generally, insect taxa.

3.5.3.1 rDNA array

18S is the standard marker for insect phylogenetics, especially for higher-level categories (Caterino, *et al.*, 2000). Levels of divergence of 18S rDNA within Cynipidae are very small (Table 3.3), and the largest part of the molecule is resistant to substitutions, as indicated by the fact that the best-fit model of evolution for this locus estimates 91.95% of the sites to be invariable (Table 3.2). However, it is one of the least saturated genes (Fig. 3.2d) and the generated topology, although not very resolved, is concordant with morphological phylogeny (Fig. 3.1). 18S will probably be more useful as a marker for inter-family and inter-order insect phylogenies (i.e. for divergences >85mya, considering that gallwasps originated at least 85 mya). Wiegmann *et al.* (2000) recently suggested that 18S will be useful for resolving insect phylogenetic splits of Mesozoic age (~ 65 – 250 mya). Our result suggests that, at least in Cynipidae, 65 mya may be too recent for good resolution.

There is an extensive literature on the use of 28S rDNA in insect phylogenetics, and compared with 18S, 28S is more frequently used in hymenopteran systematics (for a compilation of published work see Caterino, *et al.*, 2000). Levels of divergence observed within Cynipidae (0.28-8.89 % - uncorrected - Table 3.3) generally agree with levels observed within Microgastrinae (Hymenoptera: Braconidae) (0.7-12% - uncorrected) (Mardulyn and Whitfield, 1999) and in therevid flies (Diptera: Therevidae, with divergence percentages ranging from 0.26 – 6.90%) (Yang, *et al.*, 2000).

Table 3.4. Variation in nucleotide length in the ITS region (ITS1 and ITS2) for the eight taxa used in this study.

Species (tribe)	Nucleotide length (unaligned)
Panteliella bicolor (Aylacini)	1560
Synergus gallaepomiformis (Synergini)	1330
Periclistus brandti (Synergini)	1516
Barbotinia oraniensis (Aylacini)	1636
Diplolepis rosae (Diplolepidini)	1758
Plagiotrochus quercusilicis (Cynipini)	1657
Andricus kollari (Cynipini)	1447
Andricus curvator (Cynipini)	1335

The ITS regions might be poor markers for higher-level phylogenetics since a large amount of data is ambiguous due to alignment problems. Table 3.4 shows the variation in nucleotide length in the ITS regions observed for each of the eight taxa in this study. However, alignment ambiguities might be dependent on density of taxon sampling. Notwithstanding alignment problems, the usefulness of ITS regions in resolving intra-generic and intra-specific relationships has already been demonstrated (e.g. Beebe, *et al.*, 1999; Rokas, *et al.*, 2001a, Chapter 4, Appendix I). 5.8S rDNA has similar properties to 18S and 28S but its small length is a disadvantage due to the small number of variable positions.

3.5.3.2 Long-wavelength opsin

This is the first study after Mardulyn and Cameron (1999) to use LW *Rh* in insect, and more specifically hymenopteran, phylogenetics. Levels of genetic distance are similar in both studies (within the bee family Apidae the uncorrected pairwise divergence ranges between 1.93 – 19.83% whereas within Cynipidae the values are between 3.53 and 17.82%). Direct comparison shows similar average base frequencies for the two groups with a slight A-T bias in both data sets (Bees: A – 25.52%, C – 21.86%, G – 22.42%

and T – 30.2%; Gallwasps: A – 26.08%, C – 20.45%, G – 22.32%, T – 31.15%). We agree with Mardulyn and Cameron (1999) that LW *Rh* represents a promising candidate gene for insect phylogenetics and stress the importance of studies both within and outside Hymenoptera. Preliminary data within the gallwasp genus *Andricus* indicate that LW *Rh* might be also useful for the resolution of intra-generic relationships (Chapter 7).

3.5.3.3 Ef1α F1

Ef1α is a marker that has proved very useful in resolving within-family relationships (Cho, *et al.*, 1995; Belshaw and Quicke, 1997; Mitchell, *et al.*, 2000). Current evidence suggests that Hymenoptera possess two copies of Ef1α, F1 and F2 (Danforth and Ji, 1998). The Ef1α primers used in this study (see Appendix 3.1) seem to amplify the F1 copy preferentially (the one analysed here). However, in one case we did amplify the F2 copy (in *Diplolepis*, data not shown) but the high sequence divergence allowed easy discrimination between the paralogous and the orthologous copy by comparison with the *Apis* (bee) F1 and F2 copies. This suggests that – for lower-level phylogenetics at least – the discrimination between paralogy and orthology in the Ef1α is not a serious obstacle. The uncorrected distances reported for gallwasps (1.91-12.81%) roughly agree with observed distances within the butterfly superfamily Noctuoidea (1 – 10.7%) (Mitchell, *et al.*, 2000), whereas the upper limit is higher within the dipteran family Therevidae (2-17%) (Yang, *et al.*, 2000). The Ef1α data presented here agree with published findings about the usefulness of the Ef1α marker for lower-level phylogenetics.

3.5.3.4 mtDNA loci

The two mitochondrial loci in this study (COI and Cyt*b*), as well as their combined dataset (mtDNA loci), are best explained by the same ML model (Table 3.2). Mitochondrial loci essentially have a single history (since they do not recombine) and although studies using many mtDNA markers (e.g. Crespi, *et al.*, 1998; Naylor and Brown, 1998; Koulianos and Schmid-Hempel, 2000) may avoid problems associated

with low amounts of character data (given the large size of a typical eukaryote mitochondrial genome) (Cummings, *et al.*, 1995), their conclusions are based on data from what is essentially a single locus. Mitochondrial markers are also more susceptible to loss of variation due to small population sizes than nuclear loci (e.g. Rokas, *et al.*, 2001a, Chapter 4, Appendix I).

Cytochrome oxidase subunits and cytochrome *b* are the most conserved insect mtDNA genes (Simon, *et al.*, 1994). They are nonetheless the fastest-evolving (with the exception of the ITS region) among the set of loci in this study, showing high levels of divergence (in the range of 40-50%). This has been found for other hymenopterans, thus explaining their success in recovering inter- and intra-generic phylogenies within the Hymenoptera (for example Stone and Cook, 1998; Nyman, *et al.*, 2000) and even intra-specific phylogenies (Rokas, *et al.*, 2001a, Chapter 4, Appendix I; Stone, *et al.*, 2001). They have been less useful for recovering higher-level phylogenies both within Hymenoptera (i.e. lineages that have diverged more than 50 mya) (Belshaw and Quicke, 1997, this study) and other lineages of insects (Howland and Hewitt, 1995). Additionally, Crozier and Crozier (1993) noted a rate acceleration in the mtDNA of *Apis* (Hymenoptera) resulting in *Drosophila* proteins being more similar in amino acid composition to those of *Locusta* (Orthoptera) than to those of *Apis*, despite Hymenoptera being phylogenetically closer to Diptera than to Orthoptera (the phenomenon of long branch attraction). This rate acceleration (if widespread in Hymenoptera) may represent a serious constraint for the use of mitochondrial data for higher-level phylogenetics.

3.6 Conclusion

These results suggest a rough match between particular loci and certain taxonomic ranks. mtDNA loci and ITS regions appear promising for genus-level or even within species phylogenies and indeed the first results are encouraging (Stone and Cook, 1998; Rokas, *et al.*, 2001a, Chapter 4, Appendix I; Stone, *et al.*, 2001). However, their fast substitution rate and other peculiarities (high A-T bias for mtDNA, gaps and un-

alignable parts for ITS regions) make them less suitable for higher-level phylogenetics. In contrast to the mtDNA and ITS regions, the rest of the loci analysed here show lower substitution rates. At one extreme, 18S rDNA is evolving too slowly to be of use for cynipid phylogenetics or even for within-family insect phylogenetics in general. However, 28S rDNA, EF1α F1 and LW *Rh* are very promising candidates for the resolution of cynipid relationships, with low saturation levels (Fig. 3.2) and in a useful range of divergence. At least in gallwasps, LW *Rh* is faster-evolving than 28S and EF1α F1, making it a more appropriate marker for within-tribe phylogenetics.

The analysis presented here has enabled us to identify a number of useful loci for various taxonomic ranks within gallwasps (and possibly within insects in general). More extensive sampling of species will be essential in our attempt to resolve phylogenetic relationships among insects and test evolutionary hypotheses.

3.7 Acknowledgements

A version of this chapter is in press for publication: Rokas, A., J. A. A. Nylander, F. Ronquist and G. N. Stone, 2001. A maximum likelihood analysis of eight phylogenetic markers in gallwasps (Hymenoptera: Cynipidae): implications for insect phylogenetic studies. *Molecular Phylogenetics and Evolution.*

This study was performed in collaboration with Johan Nylander and Fredrik Ronquist of the University of Uppsala in Sweden. Johan provided the sequences for the 28S rDNA, LW *Rh* and COI loci and contributed in the analysis of the data. Johan, Fredrik and Graham provided extensive comments on earlier versions of the manuscript.

Appendix 3.1. Protocols for the amplification of the loci

18S rDNA

Primers used: 18e from Palumbi (1996) and the reverse of ITS5 from White *et al.* (1990). For the direct sequencing, two internal primers were designed from gallwasp sequences. Their sequences are:

18SF2: 5' - CTA CCA CAT CCA AGG AAG GCA G – 3' (22 nucleotides)

18SR2: 5' - AGA GTC TCG TTC GTT ATC GGA – 3' (21 nucleotides)

Sequencing was performed directly from the PCR product.

Concentration of PCR ingredients: PCRs were performed in 25µl volumes and they consisted of 1µl of DNA sample, 2.5µl of 10x PARR Buffer (HYBAID), 1.5µl of $MgCl_2$ (25 mM), 0.5µl of dNTPs (10mM), 0.35µl of each primer (20mM), 0.25µl of Taq (Promega) and 18.55µl of distilled, deionized H_2O.

PCR program: one step at 94 °C for 3 min followed by 35 cycles of 94 °C for 30 sec, 55 °C for 60 sec, 72 °C for 2 min and a final extension step at 72 °C for 10 min.

Genbank Acc. No.: AF395142-AF395149. TreeBASE Matrix Acc. No.: M1002.

28S rDNA

Primers used: 28SF and 28Sbout from the Wheeler/DeSalle lab at the American Museum of Natural history, New York.

Their sequences are:

28Sbout 5' - CCC ACA GCG CCA GTT CTG CTT ACC – 3' (24 nucleotides).

28SF 5' - AGT CGT GTT GCT TTG ATA GTG CAG – 3' (24 nucleotides).

These primers were also used for the direct sequencing, together with two internal primers designed from gallwasp sequences. Their sequences are:

28SFAf: 5' - GGT ACT TTC AGG ACC CGT CTT – 3' (21 nucleotides)

28Sin1: 5' - ACC TTC ACT TTC ATT AYG CCT TTA – 3' (26 nucleotides)

Sequencing was performed directly from the PCR product.

Concentration of PCR ingredients: PCRs were performed using pre-mixed, pre-dispensed reactions (Ready-to-Go™ PCR Beads, Amersham Pharmacia Biotech, cat. 27-9553-01). Each reaction contained 5 μl of DNA sample, 4μl of each primer (final concentration 1.6 μM), 1.5 units of Taq DNA polymerase, 10 mM Tris-HCl, 1.5 mM $MgCl_2$, 50 mM KCl, and 200 μM of dNTPs and stabilisers, including BSA. Sterile distilled water was added to give a final reaction volume of 25μl.
PCR program: one step at 94 °C for 5 min followed by 38 cycles of 94 °C for 60 sec, 60 °C for 60 sec, 72 °C for 60 sec and a final extension step at 72 °C for 5 min.
Genbank Acc. No.: AF395150-AF395157. TreeBASE Matrix Acc. No.: M1003.

ITS1-5.8S rDNA-ITS2

Primers used: ITS5 and ITS4 from White *et al.* (1990). For the direct sequencing, two internal primers were designed from gallwasp sequences. Their sequences are:
ITS5.8F: 5' - GTC CAC GGA TAC AAT TCC CGG ACC A – 3' (25 nucleotides) and its reverse complement ITS5.8R: 5' - TGG TCC GGG AAT TGT ATC CGT GGA C – 3' (25 nucleotides). Sequencing was performed from clones.
Concentration of PCR ingredients: PCRs were performed in 25μl volumes and they consisted of 1μl of DNA sample, 2.5μl of 10x PARR Buffer (HYBAID), 1.5μl of $MgCl_2$ (25 mM), 0.5μl of dNTPs (10mM), 0.35μl of each primer (20mM), 0.25μl of Taq (Promega) and 18.55μl of distilled, deionized H_2O.
PCR program: one cycle of 94 °C for 2 min, 55 °C for 60 sec and 72 °C for 2 min, followed by 35 cycles of 94 °C for 30 sec, 55 °C for 60 sec, 72 °C for 2 min and a final extension step at 72 °C for 10 min.
Genbank Acc. No.: AF395158-AF395165. TreeBASE Matrix Acc. No.: M1004 (for 5.8S rDNA) and M1008 (for ITS1 and ITS2).

Elongation factor 1α F1

There are two paralogs of EF1α in hymenopterans (Danforth and Ji, 1998). In this analysis we use only EF1α F1. Primers used: M44-1 and rc51-1 from Cho *et al.* (1995)

(these primers can amplify both F1 and F2 paralogs). Sequencing was performed from cloned fragments.

Concentration of PCR ingredients: PCRs were performed in 25µl volumes and they consisted of 1µl of DNA sample, 2.5µl of 10x PARR Buffer (HYBAID), 1.5µl of $MgCl_2$ (25 mM), 0.5µl of dNTPs (10mM), 0.35µl of each primer (20mM), 0.25µl of Taq (Promega) and 18.55µl of distilled, deionized H_2O.

PCR program: one step at 94 °C for 3 min followed by 35 cycles of 94 °C for 30 sec, 55 °C for 60 sec, 72 °C for 2 min and a final extension step at 72 °C for 10 min.

Genbank Acc. No.: AF395166-AF395173. TreeBASE Matrix Acc. No.: M1007.

Long-wavelength opsin

Primers used: LWRhF and LWRhR from Mardulyn and Cameron (1999). Sequencing was performed from cloned fragments.

Concentration of PCR ingredients: PCRs were performed using pre-mixed, pre-dispensed reactions (Ready-to-Go™ PCR Beads, Amersham Pharmacia Biotech, cat. 27-9553-01). Each reaction contained 8 µl of DNA sample, 3µl of each primer (final concentration 1.2 µM), 1.5 units of Taq DNA polymerase, 10 mM Tris-HCl, 1.5 mM $MgCl_2$, 50 mM KCl, and 200 µM of dNTPs and stabilisers, including BSA. Sterile distilled water was added to give a final reaction volume of 25µl.

PCR program: one step at 94 °C for 5 min followed by 36 cycles of 94 °C for 60 sec, 59 °C for 60 sec, 72 °C for 60 sec and a final extension step at 72 °C for 5 min.

Genbank Acc. No.: AF395182-AF395189. TreeBASE Matrix Acc. No.: M1009.

Cytochrome *b*

Primers used: CB1 and CB2 (Jermiin and Crozier, 1994; Stone and Cook, 1998). The PCR protocol for cytochrome *b* used in this study has been described in detail elsewhere (Stone and Cook, 1998).

Genbank Acc. No.: AF395136-AF395141. TreeBASE Matrix Acc. No.: M1006.

Cytochrome Oxidase I

Primers used: lco and hcoexternb from the Wheeler/DeSalle lab at the American Museum of Natural history, New York. The amplified fragment corresponds to positions 1835 to 2911 in the *Apis mellifera* mtDNA sequence (Genbank Acc. No.: L06178, Crozier and Crozier, 1993). Their sequences are:

lco 5' - TCW ACM AAT CAT AAA RAT ATT GG – 3' (23 nucleotides).

hcoexternb 5' - CCT ATT GAW ARA ACA TAR TGA AAA TG – 3' (26 nucleotides).

Sequencing was performed from cloned fragments.

Concentration of PCR ingredients: PCRs were performed using pre-mixed, pre-dispensed reactions (Ready-to-Go™ PCR Beads, Amersham Pharmacia Biotech, cat. 27-9553-01). Each reaction contained 8 μl of DNA sample, 4μl of each primer (final concentration 1.6 μM), 1.5 units of Taq DNA polymerase, 10 mM Tris-HCl, 1.5 mM $MgCl_2$, 50 mM KCl, and 200 μM of each dNTP and stabilizers, including BSA. Sterile distilled water was added to give a final reaction volume of 25μl.

PCR program: one step at 94 °C for 5 min followed by 38 cycles of 94 °C for 30 sec, 46 °C for 75 sec, 72 °C for 75 sec and a final extension step at 72 °C for 5 min.

Genbank Acc. No.: AF395174-AF395181. TreeBASE Matrix Acc. No.: M1005.

SECTION II: APPLICATIONS OF 'TREE-THINKING' AT THE INTRA-SPECIFIC LEVEL

CHAPTER 4

Understanding patterns of genetic diversity in the oak gallwasp *Biorhiza pallida*: demographic history or a *Wolbachia* selective sweep?

4.1 Abstract

The endosymbiont *Wolbachia* can be responsible for selective sweeps on mitochondrial DNA variability within species. Similar signals can also result from demographic processes, although crucially the latter affect nuclear as well as mitochondrial loci. Here we present data on *Wolbachia*-infection status and phylogeographic patterning for a widely distributed insect host, the oak gallwasp *Biorhiza pallida* (Hymenoptera: Cynipidae). 218 females from 8 European countries were screened for *Wolbachia*. All individuals from Hungary, Italy, France, U.K., Ireland, Switzerland, Sweden, northern and southern Spain were infected with a single group A strain of *Wolbachia*, while populations in central Spain were not infected. A mitochondrial marker (cytochrome *b*) shows low variation and departure from neutrality in infected populations, but greater variation and no deviation from neutrality in *Wolbachia*-free populations. This pattern is compatible with a *Wolbachia*-induced selective sweep. However, we also find parallel differences between infected and uninfected populations for nuclear markers (sequence data for ITS1 and ITS2). All markers support the existence of a deep split between populations in Spain (some free of *Wolbachia*), and those in the rest of Europe (all infected). Allelic variation for 5 allozyme loci is also consistent with the Spain – rest of Europe split. Concordant patterns for nuclear and mitochondrial markers suggest

that differences in the nature and extent of genetic diversity between these two regions are best explained by differing demographic histories (perhaps associated with range expansion from Pleistocene glacial refuges), rather than a *Wolbachia*-associated selective sweep.

4.2 Introduction

Bacteria of the genus *Wolbachia* are intracellular microorganisms that infect the reproductive tissues of arthropods and nematodes (O'Neill, *et al.*, 1997; Werren, 1997; Stouthamer, *et al.*, 1999). They are inherited cytoplasmically (i.e. passed from mother to daughter) and alter reproduction in their arthropod hosts in a number of ways, including cytoplasmic incompatibility, male killing, feminisation and imposition of parthenogenesis (see O'Neill, *et al.*, 1997, and chapters therein). *Wolbachia* are extremely common, infecting 16-22% of insects (Werren, *et al.*, 1995; West, *et al.*, 1998; Werren and Windsor, 2000) with a recent study indicating that the percentage might be even higher (Jeyaprakash and Hoy, 2000). In addition to immediate reproductive modifications, *Wolbachia* infection has a range of longer term evolutionary impacts on host taxa (O'Neill, *et al.*, 1997; Werren, 1997; Stouthamer, *et al.*, 1999).

Most research on *Wolbachia*, particularly for insect hosts, has concentrated on understanding the phylogenetic distribution and extent of infection (Werren, *et al.*, 1995; West, *et al.*, 1998; Werren and Windsor, 2000), using a small number of individuals for each species. The few studies examining geographic variation in levels of *Wolbachia* infection of a single host have revealed spatial patterns in the presence/absence of the bacterium and the occurrence of multiple infections (Turelli, *et al.*, 1992; Solignac, *et al.*, 1994; Plantard, *et al.*, 1998a; Malloch, *et al.*, 2000). Studies that combine analyses of spatial variation in *Wolbachia* infection, and of genetic diversity in the host, are of particular relevance because of the potential impact of *Wolbachia* on the host's genetic structure (described below). In this Chapter we address the potential impact of *Wolbachia* infection on large scale genetic patterning in an insect host, the oak gallwasp *Biorhiza pallida*.

When an advantageous mutation is driven through a population to fixation (a process known as a selective sweep), much of the neutral variation at linked loci is eliminated during the process (Maynard Smith and Haigh, 1974). The genetic variants at the linked loci which are initially paired with the advantageous mutation, though neutral in themselves, 'hitch-hike' to fixation. Spread of a specific *Wolbachia* strain through a host population can have a directly analogous effect on variation in other cytoplasmically inherited markers, such as mitochondrial DNA. Evidence from *Drosophila* suggests that a single mtDNA haplotype may become widespread in the host population through hitch-hiking with a successful *Wolbachia* strain (Turelli, *et al.*, 1992; Ballard, *et al.*, 1996). Selective sweeps on mtDNA not only reduce haplotype diversity but also cause the remaining set of host haplotypes to deviate from predictions based on neutrality (Johnstone and Hurst, 1996).

These potential impacts of *Wolbachia* infection on mitochondrial markers are particularly important because the latter are often employed as phylogeographic markers under the assumption of neutral evolution (Johnstone and Hurst, 1996). Low variability in mtDNA, and departure from neutrality, can also be caused by demographic processes such as range expansions or population bottlenecks (Avise, 2000). Demographic and *Wolbachia*-related explanations for spatial patterning in mtDNA variation can be distinguished by comparing patterns seen for nuclear and mitochondrial markers. Unlike mitochondrial markers, nuclear markers are typically inherited in a Mendelian fashion and are not expected to show any change in nucleotide diversity or deviation from neutrality in response to *Wolbachia* infection, although this depends on the phenotypic effect that *Wolbachia* has on its host. More specifically, only if sexual reproduction is involved may this assumption hold. If the phenotypic effect of *Wolbachia* on its host is the induction of parthenogenesis (and assuming that genetic exchange with sexual relatives is rare), concordance between nuclear and mitochondrial markers of the host may be expected.

In contrast to a *Wolbachia*-induced selective sweep, demographic processes cause changes in the extent and nature of variation for both mitochondrial and nuclear markers, although mitochondrial markers are expected to show stronger responses due to their lower effective population size (Avise, 2000). This difference is the basis for demonstrating *Wolbachia*-associated selective sweeps using

comparative studies of sequence variation in mitochondrial and nuclear markers (for example, Turelli, *et al.*, 1992; Ballard, *et al.*, 1996). For hosts infected with *Wolbachia*, concordance in spatial patterning of genetic diversity in nuclear and mitochondrial markers argues against a significant causative role for *Wolbachia* infection.

Undetected selective sweeps on mtDNA variation due to *Wolbachia* infection can thus generate at least two types of artefact if interpreted in a phylogeographic context. First, loss of mtDNA diversity in part of the host species' range could be attributed to a demographic effect having a similar impact, such as a population bottleneck. Here a process that in fact affects only cytoplasmically inherited markers could be mistaken for a demographic process affecting both mitochondrial and nuclear markers. Second, patterning in mtDNA variation generated by spatial patterning in *Wolbachia* infection could be mistakenly interpreted as indicative of the phylogeographic history of the host. These potential pitfalls are highly relevant to studies of insect phylogeography, because of the high proportion of insect species infected with *Wolbachia* (Werren, *et al.*, 1995; West, *et al.*, 1998; Jeyaprakash and Hoy, 2000; Werren and Windsor, 2000) and the widespread use of mitochondrial markers in phylogeographic reconstruction (Avise, 2000).

In this paper we analyse spatial patterns of variation in *Wolbachia* infection, and mitochondrial and nuclear markers for a widespread European phytophagous insect, the oak apple gallwasp *Biorhiza pallida* (Olivier 1791). *B. pallida* induces galls on oaks in the section *Quercus* (Csóka, 1997). *B. pallida* is extremely widely distributed in the western Palaearctic, extending from Morocco in the west to Georgia in the east, and as far north as Sweden. *Wolbachia* infection was detected in central European populations of *B. pallida* as part of a broader phylogenetic survey of *Wolbachia* in oak gallwasps (Chapter 8). These two features of *B. pallida* make it a suitable taxon within which to examine large scale spatial patterns in *Wolbachia* infection, and any associated impacts on variation at both mitochondrial and nuclear markers.

Oak gallwasps are obligate parasites of oak trees, and their spatial patterns of genetic variation are expected to reflect to an extent the phylogeographic history of their oak hosts. During the Pleistocene (1.8 mya until 0.01 mya), cycles of glacials

and interglacials led to repeated range contraction and expansion of many taxa across Europe, with many species surviving the glacials in refugial areas in the southern of Europe (Italy, the Balkans and Spain) (Huntley and Webb, 1989; Hewitt, 1999). The retreat of the last ice sheet across Europe at the end of the Pleistocene was followed by northern range expansion by many organisms from one or more of these refuges, resulting in the distribution pattern we see today. A growing body of work on oak gallwasps, covering 12 species in the genera *Andricus* and *Cynips*, shows that their current distributions and geographic patterns of genetic variation are determined largely by two factors: (a) the number and location of regions that acted as glacial refuges, and (b) the extent to which alternative refuges have contributed colonists to post-glacial range expansion (Stone and Sunnucks, 1993; Sunnucks and Stone, 1996; Atkinson, 2000) (Stone, *et al.*, 2001, see also Chapter 5). Consistent features of all species studied include (a) the existence of refuge-specific allozyme alleles and mtDNA haplotypes, and (b) a decline in genetic diversity with increasing latitude and distance from refuges. Where both have been studied, patterns of genetic variation northern of glacial refuges are similar for mitochondrial and nuclear markers (Stone, *et al.*, 2001), and are consistent with genetic sub-sampling of neutral variation associated with the range expansion process (Stone and Sunnucks, 1993; Sunnucks and Stone, 1996; Atkinson, 2000). Preliminary sampling of the species whose phylogeography has been studied also suggests that, unlike *B. pallida,* they are free of *Wolbachia* infection. Though precise phylogeographic scenarios vary among species, these studies indicate that in the absence of any impact of *Wolbachia* infection similar concordance in spatial patterning in nuclear and mitochondrial markers represents a qualitative null expectation for *B. pallida*. In contrast, discordant patterning in nuclear and mitochondrial markers, and concordance in patterning of *Wolbachia* infection and mtDNA diversity would suggest a significant impact of *Wolbachia*.

We established spatial patterns in *Wolbachia* infection by using a PCR-based screening technique for 218 *B. pallida* individuals sampled from 46 localities in 8 countries across Europe (Fig. 4.1). Strain diversity of *Wolbachia* was assessed by sequencing a fragment of the *wsp* gene. Host genetic diversity was analysed using a mitochondrial sequence (a fragment of the cytochrome *b* gene) and two nuclear

sequences (the internal transcriber regions ITS1 and ITS2). We extended the diversity of nuclear markers sampled by screening 270 individuals for five polymorphic allozyme loci. We use these data to address the following questions: (a) How many strains of *Wolbachia* are present in this host, and how many infection events have occurred? (b) Is infection with *Wolbachia* associated with lower sequence diversity, and departure from neutrality, for the mitochondrial marker? (c) Do the mitochondrial and nuclear markers show similar or discordant patterns of variation? (d) Can we discriminate between demographic processes and a *Wolbachia*-associated selective sweep as possible causes of observed variation in host mitochondrial DNA?

4.3 Materials and methods

4.3.1 Collection and DNA extraction

Sexual generation galls of *B. pallida* were collected from 46 localities in 8 European countries (Fig. 4.1, Table 4.1). *B. pallida* galls are multilocular (more than one offspring emerges from a single gall). To minimise screening of siblings we used one female from each gall, except where there was just a single gall from a particular location, when two females were screened. DNA was extracted from 218 female wasps as in Stone and Cook (1998). To avoid contamination, prior to DNA extraction each female wasp was soaked in 5% bleach and then serially rinsed in drops of sterile water. With each DNA extraction three control extractions were performed using a *Nasonia Wolbachia*-positive strain, a *Nasonia Wolbachia*-negative strain and a no-DNA sample.

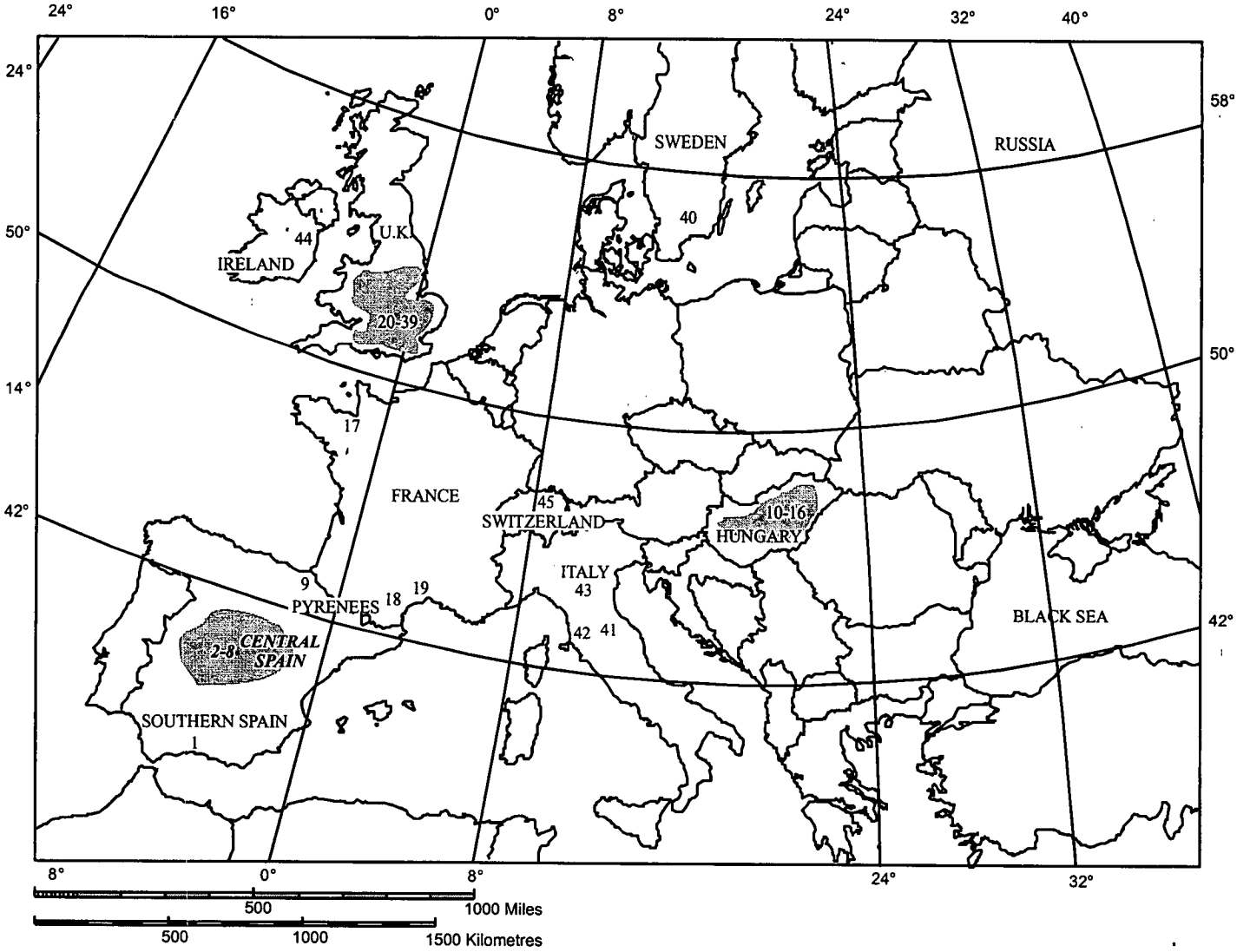


Figure 4.1. Map of collection sites of populations of *Biorhiza pallida*. Spain: 1. Prado del Rey, 2-8. Guadalix de la Sierra, Los Molinos, Zarzalejo, Avila-El Escorial, Cercedilla, Soto del Real, Villaviciosa. Northern Spain: 9. Puerto de Velate. Hungary: 10-16. Tiszaigar, Mátrafüred, Szentendre, Karcag, Bajna, Szeghalom, Visegrad. France: 17. Rennes, 18. St Jean pied de Porte, 19. Clermont-l'Hérault. U.K.: 20-39. Cambridge, Hertford, Birnwood Forest, Oxford, Isle of Wight, Hampstead Heath, Fakenham, Broughton, SW-Lincoln, Thetford, Elsfield, London, Ascot, Cawood, Chatham, West Shropshire, Grace Dieu Wood, Shorne, Hertingfordbury, Stoughton. Sweden: 40. Uppsala. Italy: 41. Chianti, 42. Volterra, 43. Casina. Ireland: 44. Dublin. Switzerland: 45. Luin. Populations in ***bold-italics*** in central Spain indicate absence of *Wolbachia* infection.

Table 4.1. Sampling regions of *Biorhiza pallida*. + indicates infection with *Wolbachia*, - indicates absence of infection.

Geographical region	Number of individuals tested	*Wolbachia*
Central Spain	34	–
Southern Spain	5	+
Northern Spain	5	+
Hungary	26	+
France	17	+
U.K.	95	+
Sweden	2	+
Italy	12	+
Ireland	5	+
Switzerland	5	+

4.3.2 Wolbachia *screening*

Screening for *Wolbachia* was performed by PCR using *Wolbachia*-specific primers for the *ftsZ* cell-cycle gene (Werren, *et al.*, 1995). These and all other PCRs were performed in a PTC-200 DNA engine (MJ Research). The forward primer was ftsZF1 (Werren, *et al.*, 1995) and a new reverse primer WOLG-R was designed based on sequences available in Genbank. The sequence of WOLG-R (26 nucleotides) is 5'- GCA GVA TCA ACY TCA AAY ARA GTC AT -3' (V=G/A/C, Y=C/T, R=A/G). The ftsZF1 – WOLG-R pair amplifies A, B and C groups of *Wolbachia*. Screening PCRs were attempted for sample DNA extractions at dilutions ranging from one tenth to one hundredth. Control PCRs were always performed. The PCR cycle for *ftsZ* was: one cycle of 94 °C for 3 min, 55 °C for 90 sec and 72 °C for 5 min, followed by 35 cycles of 94 °C for 30 sec, 55 °C for 90 sec, 72 °C for 5 min and a final extension step at 72 °C for 5 min. The PCRs were performed in 25µl volumes and they consisted of 1µl of DNA sample, 2.5µl of 10x PARR Buffer (HYBAID), 1µl of $MgCl_2$ (25 mM), 0.5µl of dNTPs (10mM), 0.35µl of each primer (20mM), 0.25µl of Taq (Promega) and 19.05µl of distilled, deionized H_2O. 15µl of

each reaction were electrophoresed on a 1% ethidium bromide-stained agarose gel. To check that samples negative for *Wolbachia* were not because of (a) failed DNA extraction, (b) presence of PCR inhibitors, and (c) incorrect DNA concentration, control PCRs with the general eukaryotic 28S rDNA primers 28Sf and 28Sr were performed as described in Werren *et al.* (1995). Of the 218 individuals screened for *Wolbachia* infection, 12 did not amplify for 28S rDNA and were discarded.

4.3.3 PCR amplification and sequencing

All sequencing reactions were done at least twice (either with the forward and reverse primers for *wsp*, cytochrome *b* or twice with the forward primers for the ITS fragments) to minimise PCR artefacts, ambiguities and base-calling errors. Sequencing was carried out using the Perkin-Elmer BigDye Terminator chemistry and an ABI 377 sequencer.

Wolbachia diversity in infected populations was assessed by PCR and sequencing of a fragment of the *wsp* gene (nine specimens were analysed: 2 from Hungary, 1 from southern Spain, 2 from France - southern and central, 1 from Switzerland, 1 from the U.K., 1 from Ireland and 1 from Italy). This is the most polymorphic gene so far isolated from *Wolbachia* (Zhou, *et al.*, 1998) and hence the most likely to distinguish between two closely related *Wolbachia* strains. *Wolbachia*-infected individuals were sequenced for *wsp* using the 81F and 691R primers following methods described in Zhou et *al.* (1998). The total volume of 3 PCR reactions for each individual wasp was electrophoresed on a 1% agarose gel. The expected bands were cut from the gel and cleaned with the QIAQuick Gel Extraction Kit (QIAGEN cat. 28704). The clean DNA fragment was quantified and sequenced.

Twenty-nine individuals were sequenced for a 433 base pair fragment of cytochrome *b* (10 from central Spain, 5 from southern Spain, 3 from the U.K., 3 from southern France, 4 from Hungary, 1 from northern Spain, 1 from Switzerland, 1 from Ireland and 1 from Italy) showing 13 distinct haplotypes (Table 4.2). The cytochrome *b* fragment was amplified using the primers CB1 and CB2 as described in Stone and Cook (1998), and purified and sequenced as described above for *wsp*.

Internal transcriber regions were amplified using the universal primers ITS4 and ITS5 (White, *et al.*, 1990). The amplified fragment consists of the internal transcriber 1 (ITS1), the 5.8S rDNA and the internal transcriber 2 (ITS2) regions of the rDNA array. 54 clones of a 474 base pair ITS2 fragment were sequenced for 13 individuals (3 from central Spain, 1 from southern Spain, 1 from northern Spain, 2 from France, 2 from Italy, 1 from Hungary, 1 from Switzerland, 1 from Ireland and 1 from the U.K.) and 28 clones of a 635 base pair fragment of ITS1 for 6 (of the 13) individuals (see below). The PCR cycle consisted of an initial denaturation step of 94 °C for 2 min, followed by 35 cycles of 94 °C for 15 sec, 55 °C for 60 sec and 72 °C for 2 min, and a final extension step of 15 min at 72 °C. Reaction conditions were as for *Wolbachia* screening, except that 1.5µl of $MgCl_2$ and 18.55µl of deionized, distilled H_2O were used. The rDNA array is present in multiple copies in the typical eukaryote genome and undergoes concerted evolution (Hillis and Dixon, 1991). However, for fast evolving regions of the array concerted evolution is not always perfect, resulting in within-individual variation. To check for intra-individual variation, PCR products were cloned using the TOPO TA Cloning kit (Invitrogen, cat. 4500-01) and 2-6 clones from each specimen were subsequently sequenced. Plasmids containing the fragment of interest were isolated using the QIAprep Spin Miniprep Kit (QIAGEN, cat. 27104). Plasmid DNA was subsequently quantified and sequenced.

4.3.4 Allozyme screening

270 individuals were screened at five variable allozyme loci using cellulose acetate gel electrophoresis (zip zone equipment from Helena laboratories) as described in Stone and Sunnucks (1993) and Stone *et al.* (2001). Of an original set of 13 loci (see Stone, *et al.*, 2001) screened in *B. pallida,* 5 systems were found to be polymorphic: Peptidase b (PEP b) (EC 3.4.11), Aspartate aminotransferase (GOTm) (EC 2.6.1.1), 6-phosphogluconate dehydrogenase (6 PGD) (EC 1.1.1.44), Malate dehydrogenase (MDHs) (EC 1.1.1.37) and Glucose-phosphate isomerase (GPI) (EC 5.3.1.9). PEPb, GOTm, 6PGD and MDH-s were run on a sodium phosphate buffer (pH=6.3), and GPI was run using a Tris-EDTA-Maleate-$MgCl_2$ buffer (pH=7.6).

4.3.5 Analysis of Sequence Data

Sequences were aligned using CLUSTALW (Thompson, *et al.*, 1994) using the default settings. ITS1 and ITS2 alignments were manually checked to verify that there were no ambiguities. Departures from neutrality were tested using Tajima's *D* (Tajima, 1989) and Fu and Li's *D** and *F** (Fu and Li, 1993) statistics incorporated in the programme DNASP version 3.0 (Rozas and Rozas, 1999). For ITS1 and ITS2, sites with alignment gaps were excluded in neutrality tests. All generated sequences were utilised (including the multiple clones from each individual).

Phylogenies for cytochrome *b* were generated using parsimony (MP) and maximum likelihood (ML) in PAUP* version 4.0b3 (Swofford, 2000). MP was performed with 1000 bootstraps with all sites equally weighted, using the Tree-Bisection-Rooting (TBR) option in a heuristic search. All ML analyses were heuristic searches using the TBR option. To reduce computational time, only 100 bootstrap replications were performed for each ML analysis. A hierarchical series of increasingly complex models of sequence evolution was employed to identify the model that made the data most likely using Likelihood Ratio Test (LRT) statistics (reviewed by Huelsenbeck and Rannala, 1997). We tested, singly and together, the effects of unequal base frequencies, different rates between transitions and transversions (ti/tv) and rate variation over nucleotide sites (Hasegawa, *et al.*, 1985). The assumption of among-site rate heterogeneity (Yang, 1993) and the enforcement of a molecular clock were also tested for the best-fitting model. The shape parameter α of the gamma distribution and the ti/tv ratio were calculated from a 50% majority rule consensus tree using unweighted parsimony. Whenever unequal base frequencies were employed, we used the empirical frequencies of the nucleotides (Yang, 1994).

For ITS, phylogenies were constructed using combined data for ITS1 and ITS2. Each gap was coded as missing and as a fifth nucleotide character. Trees were constructed using MP (as above) and ML using PUZZLE (Strimmer and von Haeseler, 1996). MP bootstraps were carried out as above for 100 replicates. ML was performed using the quartet puzzling algorithm with 10 000 puzzling steps, the

HKY85 model of DNA substitution (Hasegawa, *et al.*, 1985) and a gamma shaped distribution to account for rate heterogeneity. All parameters were estimated from the data-set. A parsimony analysis was also performed using the gap insertions as unique indel (insertion/deletion) events. They were treated separately from the sequence data-set since no data exist concerning their frequency of substitution.

4.4 Results

4.4.1 Wolbachia *screening and diversity*

85% of the 206 *B. pallida* successfully screened were infected with *Wolbachia*. Over the entire sampled range only populations in central Spain were not infected (Fig. 4.1, Table 4.1). All sequences were 564 base pair long (Genbank accession number AF339629), with only two nucleotide positions polymorphic. Position 218 was polymorphic for T/C and position 230 was polymorphic for A/G in all the individual sequences. Since the sequencing was done from a PCR fragment, it could not be determined whether these sites are genuinely polymorphic (not fixed yet, four alleles) or represent two different strains (two alleles). The low *wsp* diversity suggests that all infected European populations carry the same strain of *Wolbachia*.

4.4.2 Mitochondrial phylogeography

The 29 individuals sequenced for the cytochrome *b* fragment were polymorphic for 25 positions (18 of them are parsimony-informative - 4.18%). This variation yielded a total of 13 discrete haplotypes (haplotypes and their polymorphic sites are shown in Table 4.2). Likelihood ratio tests showed that the best-fitting model for the *B. pallida* data used empirical base frequencies, a ti/tv ratio equal to 10.44 and variation in rate among sites (using a gamma distribution with α=0.0057 and four rate categories). The estimated value α=0.0057 denotes a very strong rate variation (see Yang *et al.*, 1994). The assumption of a molecular clock could not be rejected at the 0.05 level

(unconstrained model *vs.* model with molecular clock enforced, -ΔL=44.42, d.f.=27, $0.01<p<0.05$).

Table 4.2. List of haplotypes for the 433 base pair fragment of cytochrome *b* from *Biorhiza pallida*. Parsimony informative sites are indicated by an asterisk (*) and singletons by a full stop (.).

	1	112	222	233	333	334	4
	111	566	990	170	124	902	224	771	1
	034	912	020	535	499	321	250	394	8
	**.	**.	.*.	***	*..	***	***	.**	.
Haplotype 1	GTA	GTA	GAT	GTT	TAC	ATT	TGC	ATC	T
Haplotype 2	...	..G	...	...	.G.	...	...	...	.
Haplotype 3	...	...	...	...	...	...	...	...	C
Haplotype 4	...	...	...	...	...	...	...	G..	.
Haplotype 5	...	...	...	...	..T	...	.A.	...	.
Haplotype 6	...	...	...	...	...	...	...	..T	.
Haplotype 7	...	...	...	A..	...	...	..T	...	.
Haplotype 8	...	A..	...	A.C	C..	GAC	.A.	.AT	.
Haplotype 9	...	A..	A..	A.C	C..	GAC	.A.	.AT	.
Haplotype 10	...	A..	.G.	A.C	C..	GAC	.A.	.AT	.
Haplotype 11	A..	.A.	.G.	A.C	C..	.AC	CA.	..T	.
Haplotype 12	A.G	.A.	.G.	A.C	C..	.AC	CA.	..T	.
Haplotype 13	.C.	...	.GC	ACC	...	GAC	C.T	..T	.

Haplotype locations: Haplotype 1: St Jean pied de Porte - France, Clermont-l'Hérault - France, Szeghalom - Hungary, Broughton - U.K., Luin - Switzerland, Szentendre - Hungary, Oxford - U.K., Cambridge - U.K.; Haplotype 2: Puerto de Velate – Northern Spain; Haplotype 3: Mátrafüred - Hungary; Haplotype 4: St Jean pied de Porte - France; Haplotype 5: Bajna - Hungary; Haplotype 6: Dublin - Ireland; Haplotype 7: Casina - Italy; Haplotype 8: Avila-El Escorial – Central Spain; Haplotype 9: Guadalix de la Sierra – Central Spain; Haplotype 10: Guadalix de la Sierra – Central Spain, Zarzalejo – Central Spain, Villaviciosa - Central Spain, Los Molinos - Central Spain, Soto del Real - Central Spain; Haplotype 11:

Cercedilla - Central Spain, Zarzalejo - Central Spain; Haplotype 12: Cercedilla - Central Spain; Haplotype 13: Prado del Rey – Southern Spain (5 individuals).

All MP and ML analyses reveal a deep and well-supported split (based on 10 changes) between central and southern Spain versus populations in northern Spain and the rest of Europe (Fig. 4.2). Sequence diversity was higher in *Wolbachia*-free populations in central and southern Spain (π=0.0082) than in the *Wolbachia*-infected populations throughout northern Spain and the rest of Europe (π=0.0029). The 14 sequences from northern Spanish and rest of Europe populations showed significant deviation from predictions under neutrality (Tajima's D=-2.09, P<0.05, Fu and Li's D^*=-2.73, P<0.05 and F^*=-2.93, P<0.05). In contrast, central and southern Spanish sequences, whether the 10 uninfected central Spanish individuals (Tajima's D=0.52, P>0.10, Fu and Li's D^*=0.62, P>0.10 and F^*=0.67, P>0.10), or central and southern Spanish individuals combined (15 sequences) (Tajima's D=1.20, P>0.10, Fu and Li's D^*=0.89, P>0.10 and F^*=1.13, P>0.10) showed no significant deviation from neutrality.

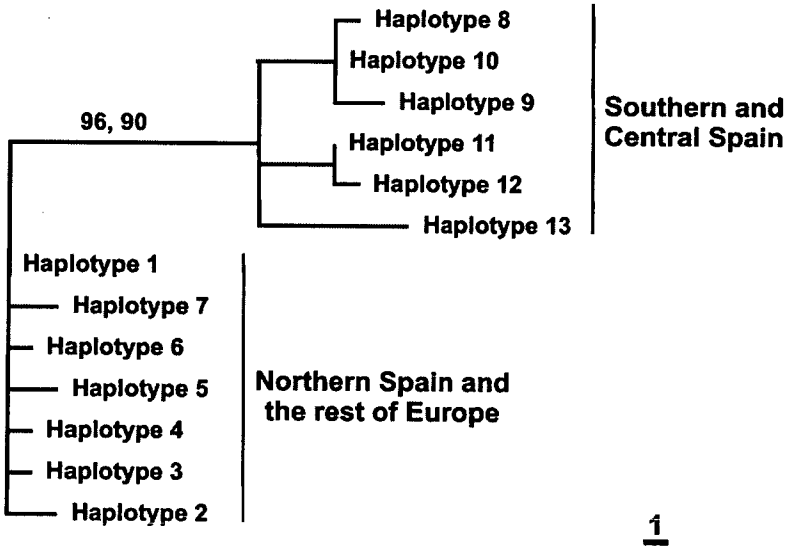


Figure 4.2. 50% majority rule consensus phylogram for cytochrome *b* haplotypes using a ML model which accounts for unequal base frequencies, a different ti/tv ratio and rate variation among sites. See Table 4.2 for haplotype information. Values above the branches denote bootstrap support under ML and MP respectively. The tree is 35 steps long with a consistency index (CI) of 0.714 and a retention index (RI) of 0.931. Genbank accession numbers: AF339616-AF339628.

	ITS1	ITS2
	2223444455555	11122222222333333344444444
	894775578800111	13436601111444144558900000012
GROUP	545675223456345	58891260123678689017956789 62
1	??????????????	T--A--TTA-----AGT--CTA-----T
2	AAA--TAA--TAATT	.--.--.....GT..----A.
3	AAA--TAA--TAATT	.--.--.....--.....----A.
4	...AAC.T-.....	C-C.--G.-----.....----A.
5	...AA..T-.....	C-C.--G.-----.....----A.
6	....--..T--..---	.-CG--G.-----.....----A.
7	....--.....--.....	.--G--.....-----.....----A.
8	??????????????	.--.--...TACAT...GT.A.----A.
9	...--.GT--..---	C--.GT.----CAT...GT.A.----A.
10	...AA..T-.....	C--.--...TACAT...GT.A.A---A.
11	....--.....--.....	C--.GT.----CAT...GT.A.A---A.
12	...--CGT--..---	C--.GT.----CAT...GT.A-----A.
13	...--.GT--..---	C--.GT.----CAT...GT.A-----A.
14	..---..TTA-----	CT-.GT.-------.----.A.A---A.
15	..---..TTA-----	CT-.GT.-------.----.A.A--AA.
16	..---..TTA-----	CT-.GT.-------.----.A.----A.
17	??????????????	CA-.GT.-------.----TAGAAAATC
18	??????????????	CA-.GT.-------.----.AGAAAAT.
19	.T---..TTA-----	CA-.GT.-------.----.A---AAA.
20	.T---..TTA-----	CA-.GT.-------.----.AA----A.
21	C..--..T-------	CT-.GT.-------G----.A.A---A.
22	C..--..TTA-----	CT-.GT.-------G----.A.----A.
23	C..--..TTA-----	CT-.GT.-------G----.A-----A.
24	..---..TTA-----	CT-.GT.-------G----.A-----A.
25	??????????????	C--.GT.----CAT...--.....----A.

Table 4.3. Parsimony-informative sites for ITS1 and ITS2. A dash (-) denotes a gap and a full stop (.) denotes an identical site.

Individuals with the same parsimony-informative sites were grouped together (see below). Numbers indicate number of individual and number of clone respectively. All numbers refer to the individuals in Fig. 4.3. The asterisk (*) indicates individuals that have been sequenced both for ITS1 and ITS2. Individuals without asterisk have been sequenced only for ITS2. Groups: **1:** 123.5, **2:** 140.10*, **3:** 117.3*, 123.11, 123.13, 48.4, 48.11, 53.4, 212.11, 35.3, 48.12, 53.11, 96.20*, 53.2, 48.7, 35.1, 48.3, **4:** 212.3, 212.9, 212.7, 212.8, 96.13*, **5:** 96.10*, 96.18*, 96.16*, **6:** 117.15*, **7:** 117.2*, **8:** 123.1, **9:** 141.6*, 140.5*, 141.10*, 141.3*, 48.5, **10:** 96.19*, **11:** 140.13*, **12:** 141.5*, **13:** 141.12*, **14:** 147.5*, **15:** 147.14*, **16:** 147.15*, **17:** 157.6, 157.16, **18:** 157.13, 157.7, 172.5, **19:** 147.19*, **20:** 147.13*, **21:** 32.5*, **22:** 32.7*, 32.3*, 32.8*, **23:** 32.10*, **24:** 147.3*, **25:** 123.3.

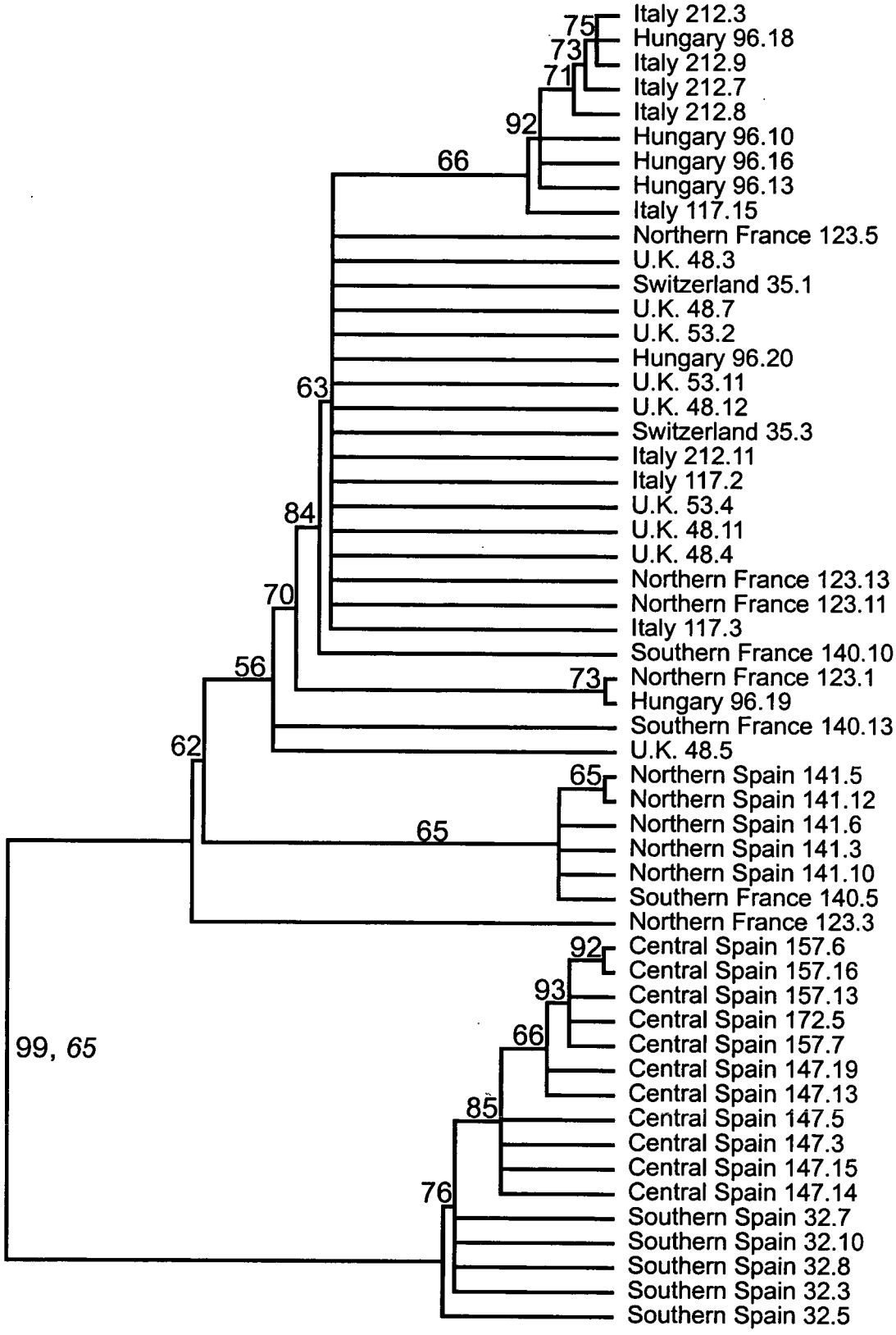


Figure 4.3. 50% majority rule consensus tree for ITS using MP with gaps coded as the fifth base. Numbers in branches denote bootstrap support under MP (alignment gaps as fifth base). For the southern and central Spain versus northern Spain and rest of Europe split the bootstrap support given by MP analysis of the indel data set is shown in italics. Numbers next to countries' names indicate number of individual and number of clone respectively. The tree is 141 steps long with a consistency index (CI) of 0.794 and a retention index (RI) of 0.925. Genbank accession numbers: ITS1: AF340069-AF340096 / ITS2: AF340097-AF340150.

4.4.3 ITS phylogeography

The 13 individuals sequenced for ITS2 and the 6 individuals sequenced for ITS1 and ITS2 revealed 43 parsimony-informative positions (Table 4.3). The length of the fragments varied between 456-467 base pairs for ITS2 and between 626-632 base pairs for ITS1. Length differences from clones from the same individual were usually much smaller or absent. The ML tree for ITS constructed with the quartet puzzling algorithm, and the MP tree coding alignment gaps as missing information, are both poorly resolved. However, MP analysis with gaps coded as a fifth base strongly supports the central and southern Spanish versus northern Spanish and rest of Europe split revealed by the mitochondrial cytochrome *b* data (Fig. 4.3). MP analysis of the ITS indel data-set (with 10 of 12 indels parsimony-informative) revealed the same topology, though with lower bootstrap support.

Deviation from neutrality was tested separately for ITS1 and ITS2. Given that these are nuclear loci, within which recombination might be occurring, the estimates for deviation from neutrality are conservative. Significantly, regional variation in departures from neutral expectations parallel those seen in cytochrome *b*. For both loci, Fu and Li's D^* and F^* and Tajima's D showed significant departure from neutrality for populations outside Spain (ITS1: Tajima's D=-1.92, P<0.05, Fu and Li's D^*=-2.77, P<0.05 and F^*=-2.92, P<0.05; ITS2: Tajima's D=-1.98, P<0.05, Fu and Li's D^*=-4.15, P<0.05 and F^*=-4.05, P<0.05) but were non-significant for central Spanish specimens (ITS1: Tajima's D=-1.01, P>0.10, Fu and Li's D^*=-0.99, P>0.10 and F^*=-1.07, P>0.10, Tajima's D=-1.01, P>0.10; ITS2: Tajima's D=-0.77, P>0.10, Fu and Li's D^*=-0.86, P>0.10 and F^*=-0.95, P>0.10).

4.4.4 Allozyme variability

The sample sizes obtained from each location (maximum 12 galls, and so 12 females screened) were too small to allow meaningful analysis of allele frequencies. We thus limit our interpretation of the data to the presence/absence of specific alleles, as summarised in Table 4.4. Four alleles are found only in Spain. Of these, two are present only in central Spain; allele 1 at 6-PGD is present only in one population,

whereas allele 2 at 6-PGD is present in four populations. The third allele is present in the only southern Spanish population (allele 4 at GPI) and one is present in all ten Spanish populations (allele 1 at MDHs). Populations in the rest of Europe possess two alleles absent from Spain. One is from a single individual from Stoughton, U.K. (allele 2 at GPI) and the other is very common in individuals from Phoenix Park, Ireland (allele 4 at MDHs). These regional differences in locally-restricted alleles support the substantial genetic divergence between central and southern Spain versus northern Spain the rest of Europe implied by cytochrome *b* and ITS sequence analyses.

Table 4.4. *Biorhiza pallida* allozyme alleles from 5 allozyme loci, from 4 areas of Europe. Locally-restricted alleles are written in bold. Numbers in parentheses denote the number of individuals and number of populations screened, respectively.

Area \ Allozyme	**6 PGD**	**GOTm**	**PEP b**	**MDHs**	**GPI**
Central and Southern Spain (43 - 10)	**1,2**,3,4	2,4	1,2,3,4	**1**,2,3	3,**4**
Southern France (20 - 3)	4	4	1,2,3,4	2	3
Italy, Hungary (51 - 11)	4	2,4	1,2,3,4	2	3
U.K., Eire, Northern France (156 - 30)	3,4	4	1,2,3,4	2,3,**4**	**2**,3

4.5 Discussion

4.5.1 Geographic variation in Wolbachia *infection*

Extensive screening revealed that, with the exception of central Spain, all European populations of *B. pallida* sampled are infected with *Wolbachia*. The apparent single strain infecting *B. pallida* appears to be relatively cosmopolitan; it is shared with three other European gallwasp species (Chapter 8), and database searches suggest that it also infects tsetse flies. There are three possible explanations for existence of infected populations of *B. pallida* both northern and southern of a *Wolbachia*-free region in central Spain: (a) there was a single infection event and central Spanish populations have lost their infection, (b) there was a single infection in a common

ancestor of southern Spanish, northern Spanish and the rest of Europe *B. pallida* not also shared with populations in central Spain, or (c) infection has occurred independently in southern Spain and in Europe from Pyrenees northwards. Too little is currently known about mechanisms of *Wolbachia* transfer between lineages for these possibilities to be distinguished, although evidence for horizontal transfer of *Wolbachia* between and within species is accumulating (Vavre, *et al.*, 1999; Huigens, *et al.*, 2000). We also know very little about the phenotypic consequences of *Wolbachia* infection in *B. pallida*. *Wolbachia* has been shown to induce parthenogenesis in some non-oak cynipids (e.g. Plantard, *et al.*, 1999), but the presence of males in all our samples suggests that this does not occur in *B. pallida*. Additionally, the slow lifecycle of this host, and obligate development of the larva within oak tissues make the mating and curing experiments needed to examine *Wolbachia*-associated effects (such as for example, cytoplasmic incompatibility) extremely difficult (see also Plantard, *et al.*, 1998a; Plantard, *et al.*, 1999).

4.5.2 Patterns of genetic diversity in B. pallida

Variability at the mitochondrial locus was higher in populations free of *Wolbachia* than in northern Spanish and the rest of Europe populations infected with the symbiont. Similarly, infected northern Spanish and rest of Europe populations showed significant departures from neutrality, while central and southern Spanish populations showed no such signature. Such covariation between infection and sequence diversity is compatible with a *Wolbachia*-induced selective sweep.

Large scale demographic processes such as range expansion can have a similar impact on mtDNA variation to a selective sweep in a population (Donnelly and Tavare, 1995). Studies on other European oak gallwasps predict a significant impact of range expansion on spatial patterns in *B. pallida,* and should be regarded as a more parsimonious underlying cause if compatible with the data. As discussed above, demographic processes tend to generate qualitatively similar patterns at both mitochondrial and nuclear markers. This is exactly what we see in *B. pallida*, indicating that lower genetic diversity in central and northern Europe is far more likely to result from either historically low genetic diversity in nearby refuges, or loss

of genetic diversity associated with range expansion, rather than a *Wolbachia*-associated selective sweep. *Wolbachia* infection may also generate similar patterns of genetic diversity in nuclear and mitochondrial markers, provided that the effect of *Wolbachia* on its host is to induce parthenogenesis. However, the presence of males in every population of *B. pallida* that has been collected argues against a scenario of *Wolbachia*-induced parthenogenesis.

The dominant feature of genetic variation in *B. pallida* is the deep split between central and southern Spanish versus northern Spanish and rest of Europe populations, supported by all *B. pallida* data sets. The division of these two groups is robustly supported by cytochrome *b* (Fig. 4.2), and less robustly by ITS (Fig. 4.3). The lower support observed in ITS is due to the fact that the informative sites are in the indel regions. When gaps are encoded as the fifth base or when they are treated as unique indel insertions, the split between central and southern Spain versus northern Spain and the rest of Europe is more strongly supported. Genetic discontinuities between areas northern and southern of the Pyrenees are known for many plants and animals, and many show hybrid zones at the Pyrenees (see reviews by Taberlet, *et al.*, 1998; Hewitt, 1999). For many species, including gallwasps (Stone and Sunnucks, 1993; Stone, *et al.*, 2001), expansion following retreat of the ice was principally from Italy and the Balkans, while Spanish populations failed to expand far into France. A similar scenario could explain the patterns seen in *B. pallida*. Region-specific haplotypes and nuclear alleles imply the existence of discrete refuge populations in Spain and in regions to the east. Some oak gallwasps show further differentiation between distinct refuges in Italy and the Balkans (Atkinson, 2000), but there is inadequate resolution in the *B. pallida* data to see if the same is true for this species. Range expansion from central or eastern Europe and associated rapid population growth would generate both the similarity among sites and the departure from expectations of neutrality seen outside Spain. Inability of Spanish populations to expand and spread beyond the Pyrenees would lead to absence of Spanish haplotypes or nuclear alleles from more northerly populations, and the absence of any deviation from neutrality. Inability to escape from the Spanish refuge has been clearly demonstrated in one other oak gallwasp, resulting from the evolution of oak-specific ecotypes whose hosts are essentially limited to Spain (Stone, *et al.*, 2001). It

is interesting to note that patterns of post-Pleistocene range expansion for gallwasps and for their host oaks do not match. Molecular evidence for oaks suggests that individuals from all three refuges (Spain, Italy and the Balkans) contributed to the colonisation of central Europe (Ferris, *et al.*, 1993; Dumolin-Lapegue, *et al.*, 1997). In contrast, gallwasps expanded principally from Italy and the Balkans (see above, Atkinson, 2000; Stone, *et al.*, 2001). This discordance may be explained by adaptation of gallwasp populations in the Spanish refuge to local endemic oak species that failed to expand further north (for more discussion see Atkinson, 2000; Stone, *et al.*, 2001, Chapter 5).

This phylogeographic hypothesis is supported by dating estimates based upon a *B. pallida* mitochondrial molecular clock, assuming a 2.3% divergence of mtDNA sequences per million years (Brower, 1994). There are two issues for which timing estimates are meaningful. The first concerns the timing of the event leading to the current levels of genetic diversity among northern Spanish and rest of Europe populations. Assuming that the 0-0.9% divergence that is observed among these populations is the result of substitutions that have occurred just after that event, we get an estimation between 0 and 390 000 years before present. This estimation roughly encompasses the range expansion by the oak hosts of *B. pallida* following the end of the Pleistocene (Ferris, *et al.*, 1993; Dumolin-Lapegue, *et al.*, 1997). The second issue is the split between central and southern Spain versus northern Spain and the rest of Europe. Levels of divergence between the groups of around 2.5% suggest that this split is ancient, pointing to a separation long before the end of the Pleistocene. Such a division into long-standing eastern and western refuges is supported by similar data for other oak gallwasp species, such as *Andricus kollari* and *Andricus quercustozae* (Atkinson, 2000; Stone, *et al.*, 2001, Chapter 5). Our study shows that while patterns of variation in mitochondrial sequence diversity in this system do not allow a *Wolbachia*-induced selective sweep to be discounted, consideration of nuclear marker diversity points to a demographic cause.

4.6 Acknowledgements

A version of this chapter has already been published: Rokas, A., R. J. Atkinson, G. S.

Brown, S. A. West and G. N. Stone, 2001. Understanding patterns of genetic diversity in the oak gallwasp *Biorhiza pallida*: demographic history or a *Wolbachia* selective sweep? *Heredity* **87:** 294-304.

Gordon Brown and Rachel Atkinson generated and analysed the allozyme data. Rachel Atkinson, Stuart West, Graham Stone and many anonymous referees provided extensive comments on various versions of the manuscript. David Guiliano provided *Wolbachia* strain C DNA to test the WOLG-R primer and John Werren the *Nasonia* control strains. Members of the British Plant Gall Society, Gyuri Csóka, José-Luis Nieves-Aldrey, Felix Fontal-Cazalla, Gil McVean, Olivier Plantard and Jean-Yves Rasplus provided valuable help with gall collecting.

CHAPTER 5

The phylogeographic structure of the European oak gallwasp *Andricus quercustozae* reveals changes in genetic diversity with longitude and the importance of Anatolian populations

5.1 Abstract

The amount, distribution and pattern of intra-specific genetic diversity of many European taxa has been shaped by the Quaternary ice ages. More specifically, during the ice ages populations of most species were restricted to southern refuges in Iberia, Italy and the Balkans. For species with distributions that extend to the east, refuge populations may also have existed in Asia Minor, a centre of genetic diversity for many taxa. We have collected and analysed specimens of the oak gallwasp *Andricus quercustozae* from 7 circum-Mediterranean and central European countries. The species is found from Morocco around the north coast of the Mediterranean to Iran. Here we present sequence data from a fragment of the mitochondrial gene cytochrome *b* and allele frequency data from 12 polymorphic allozyme loci in an attempt to understand (a) which are the Pleistocene refuges for *A. quercustozae*, (b) the importance of Anatolia as a refuge and potential centre of origin for the species, and (c) its postglacial range expansion pattern and how it has been influenced by the distribution of its hosts, the oaks. Our results suggest that *A. quercustozae* was present in perhaps five distinct refuges (Iberia, Italy, the Balkans, south-western Anatolia and north-eastern Anatolia) and there has been recent genetic exchange between Italian and Hungarian refuge individuals. The refuge(s) in Anatolia harbour the highest levels of genetic diversity, providing support for the hypothesis that the European radiation of oak gallwasps may have been centred in Anatolia. More generally, the potential contribution of eastern-most European and Anatolian populations in studies on European taxa may have been underestimated. Finally, we provide evidence suggesting that the current distribution of *A. quercustozae* and its

range expansion pattern during this current interglacial has been limited by the preference of the parthenogenetic generation wasps for specific host oak species.

5.2 Introduction

Over the last 2.4 Myr of earth history (the Quaternary period) Earth's climate has been dominated by a series of major glacial periods with a periodicity of about 100 kyr. During the glacials large ice sheets covered Europe to about 52°N and tundra or cold steppe reached the southern European mountain ranges. These glacial periods have been separated by shorter, milder interglacials 5 – 70 kyr in length, in which the climate in the interglacials was more similar to that of the present day (Hewitt, 1996, 1999).

The climatic oscillations caused alternating cycles of range expansion and contraction in many plant and animal taxa (Hewitt, 1999). Genetic, fossil and palynological studies across a wide range of taxa have shown that during glacial periods, species occupied three main refuges in Europe; Iberia, Italy and the Balkans (Huntley and Birks, 1983; Hewitt, 1996, 1999). Recent studies have also indicated the importance of Anatolia and Caucasus as centres of genetic diversity for European taxa with contiguous distributions (Cooper, *et al.*, 1995; King and Ferris, 1998; Hewitt, 1999). Although many taxa expanded out of the refuge areas during the interglacials, refuge-specific genetic diversity shows these populations to have remained distinct and isolated from each other for the whole of the Pleistocene (Hewitt, 1999). This can be explained either by divergence of lineages prior to the Quaternary (and thus not linked to any cold period, Taberlet, *et al.*, 1998), or by extinction of populations in the expanded range – if climatic deterioration at the end of each interglacial occurred too rapidly for successful movement back into the refugial areas, only those individuals that never moved from the refuges will have survived through the glacials.

Oak gallwasps are a tribe of wasps (Hymenoptera: Cynipidae: Cynipini) that parasitise oaks (genus *Quercus*), inducing gall formation (Askew, 1984; Stone, *et al.*, 2002). The galls formed by oak gallwasps are structurally complex and are the product of the host tree through induction by the gallwasp larva (Rohfritsch, 1992).

Their life-cycle is cyclical parthenogenesis and consists of a short-lived spring sexual generation and a longer-lived autumn parthenogenetic generation (Askew, 1984). Host use by many gallwasp species in the genus *Andricus* alternates between white oaks (section *Quercus sensu stricto*, examples: *Q. infectoria*, *Q. faginea*, *Q. petraea*, *Q. pubescens*, *Q. pyrenaica* and *Q. robur*) for the parthenogenetic generation and black oaks (section *Cerris*, examples: *Q. cerris*, *Q. suber*) for the sexual generation (Cook, *et al.*, 2001, Chapter 7; Stone, *et al.*, 2002).

Detailed phylogeographic data (mitochondrial haplotype sequences and allozyme allele frequencies) have already been published for two European oak gallwasp species, the oak apple gallwasp *Biorhiza pallida* and the marble gallwasp *Andricus kollari* (Rokas, *et al.*, 2001a; Stone, *et al.*, 2001, Chapter 4, Appendix I). Both species are distributed from Morocco in the west to the Caucasus in the east. In Europe, both species are thought to have survived the glacials in three refuges (Iberia, Italy and the Balkans), with the most significant genetic differentiation found between Iberia and the rest of Europe (Rokas, *et al.*, 2001a; Stone, *et al.*, 2001, Chapter 4, Appendix I). Additional sampling of *A. kollari* in Turkey showed this refuge to be clearly distinct from all European refuges (Stone, *et al.*, 2001, Appendix I), a pattern that may also hold for *B. pallida*.

Studies of several species (e.g. *A. kollari*, *A. quercuscalicis*) show gallwasps to be capable of rapid range expansion (Stone and Sunnucks, 1993; Csóka, *et al.*, 1998; Stone, *et al.*, 2001, Appendix I). These species have a host-alternating lifecycle, and were restricted to Italy and the Balkans until human introduction and subsequent naturalisation of their sexual generation host in these regions, the Turkey oak (*Q. cerris*), into northern Europe 400-500 years ago. These species are now common and widespread as far north as Scotland and southern Scandinavia (Schönrogge, *et al.*, 1998).

These rapid range expansions illustrate the ultimate dependence of gallwasp distributions on those of their host oaks. While oak host specificity in gallwasps has usually been assumed to be at the level of oak section rather than species, there is increasing evidence for very high levels of host oak specificity in some species (Stone and Sunnucks, 1993; Abrahamson, *et al.*, 1998; Stone, *et al.*, 2001, Appendix I). The phylogeographic analysis of several European oak species, particularly *Q.*

robur and *Q. petraea* (Huntley and Birks, 1983; Dumolin-Lapegue, *et al.*, 1997; Ferris, *et al.*, 1998; Toumi and Lumaret, 1998), indicates that while distinction between the species is unclear, the differences between refuge populations are significant, suggesting refuge isolation since at least the beginning of the Pleistocene (Ferris, *et al.*, 1998; Dumolin-Lapegue, *et al.*, 1999). Thus if the interaction between oaks and gallwasps is a tight one, refuge-specific gallwasp populations may become locally adapted to their hosts. Although evidence showing inter-refuge variation (different host use, different host-organ use, gall morphology differences) in gallwasp populations supports this argument (Stone, *et al.*, 2001, Appendix I), it is at present unclear whether these inter-refugial differences are due to changes in the gallwasp genotypes, the host oak genotypes or in an interaction between the two.

The ranges of some oak species found in Europe stretch eastwards through Anatolia and the mountains of Iran and Iraq to Afghanistan, northwards into the Caucasus, and south into Syria and Lebanon (Sternlicht, 1968; Chodjai, 1980; Yaltirik, 1982). In the western Palaearctic, the species diversity of sections *Quercus sensu stricto* and *Cerris* is highest in Anatolia (numbering 23 species) and molecular phylogenetic analysis indicates that this might have been the origin of the European radiation of both oak sections (Manos, *et al.*, 1999). Whether each individual oak species persisted in these areas throughout the Pleistocene is unclear, due partly to the notorious difficulties associated with studies of oak phylogenetics. However, high levels of species endemism and identification of numerous subspecies and races suggest that regions of both Anatolia and the Caucasus may have provided discrete refuges for oaks during this period (Turrill, 1929; Konarov, 1936; Camus, 1936-1954; Zohary, 1966; Townsend and Guest, 1980; Yaltirik, 1982; Castroviejo, 1986). Anatolia may also be a centre of diversity for oak parasites, such as gallwasps. However, although oak gallwasp species richness is high in Anatolia (Stone and Csóka, in preparation), the hypothesis that Anatolia may have been the centre of origin of many European oak gallwasp species has not, as yet, been tested.

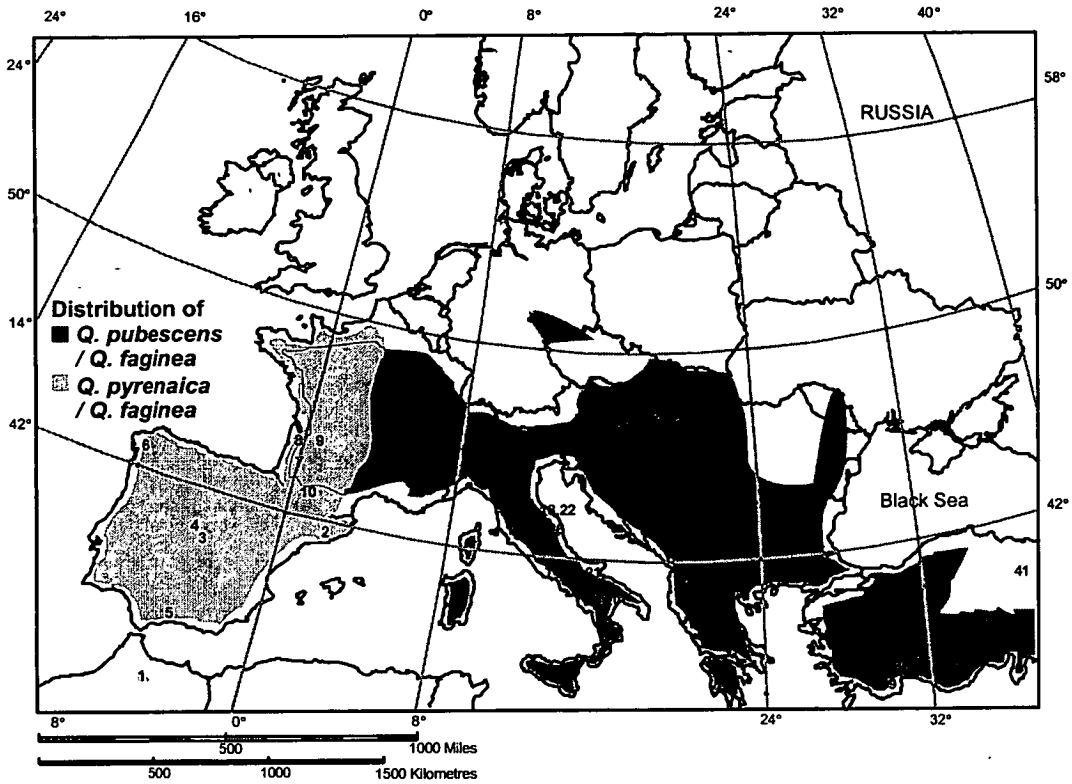
Most published phylogeographic studies of the western Palaearctic have concentrated on Europe even though the taxa may be much more widely distributed (e.g. Hewitt, 1996; Taberlet, *et al.*, 1998; Hewitt, 1999; Rokas, *et al.*, 2001a, Chapter 4, Appendix I), a state of affairs arising more from ease of sampling rather than any

intellectual boundary to the question. Additionally, the majority of studies have focused on latitudinal diversity gradients established by postglacial northwards migration, with few focusing on longitudinal diversity gradients resulting from more ancient patterns of range expansion (Hewitt, 1996; Taberlet, *et al.*, 1998; Hewitt, 1999, 2000). In this study, we have specifically sampled *A. quercustozae* to address these two questions: (a) the phylogeographic importance of Anatolian regions, and (b) longitudinal diversity gradients for *A. quercustozae*.

The choice of molecular marker for a particular question depends, practical considerations aside, upon the extent of genetic variation required to best answer the question (Parker, *et al.*, 1998). For phylogeographic studies, two of the most widely used markers have been mitochondrial DNA haplotype sequences and allozyme polymorphism data. Several differences in the nature and statistical analysis of the two molecular markers (e.g. different modes of inheritance, different mutation rates, single locus versus multi-locus, different sample sizes) have the potential, through their comparison, to provide a more accurate description of the historical processes that have acted in the populations examined than either can alone. For example, a combined use of both markers may provide both times of divergence through application of molecular clocks in the mitochondrial haplotype data and a genome-wide estimate of diversity through analysis of the allozyme data.

In this paper we study the refuge differentiation of the oak gallwasp *Andricus quercustozae* (Bosc, 1792). This gallwasp has a circum-Mediterranean distribution found from Morocco in the west through oak forests along the north of the Mediterranean into Anatolia and the Levant in the east. Some authors have considered the Anatolian and Iranian populations as a separate species, *A. insana* (Dalla Torre and Kieffer, 1910). Differences between the two species are principally in the gall structure, which is larger, more red and stickier in the *A. insana* gall. Whether these differences in gall structure reflect underlying differences in the gallwasps' genotypes or oak-host mediated differences remains to be answered. Evidence from other gallwasps provides support for both hypotheses (Stone and Cook, 1998; Stone, *et al.*, 2002).

A.



B.

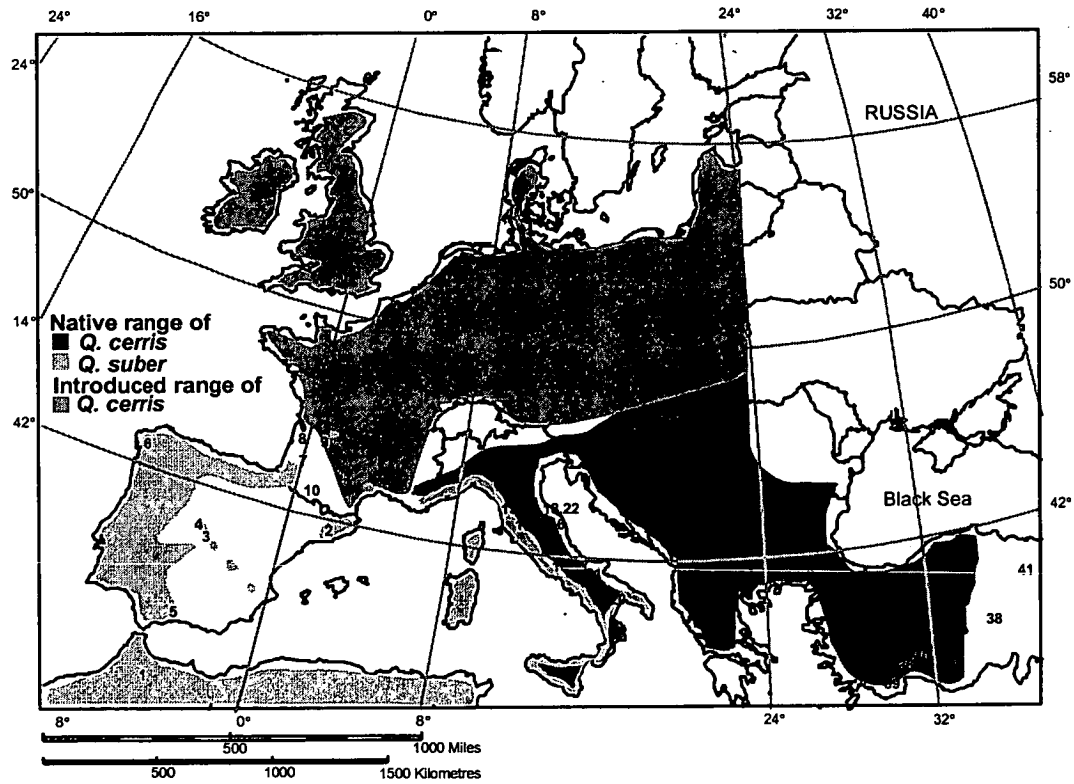


Figure 5.1. Distribution of (A) parthenogenetic generation oak hosts and (B) putative sexual generation oak hosts throughout the western Palaearctic for *Andricus quercustozae*, according to Jalas and Suominen (1987) for European and Yaltirik (1982) for Anatolian distributions. Overlapping distributions are indicated by thin lines. Collection locations are indicated by numbers, which correspond to the populations in Table 5.1. Site 42 (Yeniyol) is outside the range covered on this map (in the eastern-most point of the Black Sea coast).

A. quercustozae, unlike the species already mentioned, is known only from its parthenogenetic generation, which occurs on oaks within the section *Quercus sensu stricto* (Melika, *et al.*, 2000), specifically *Q. pubescens*, *Q. pyrenaica*, *Q. faginea* and *Q. infectoria* and other species in the east (Fig. 5.1A). Population genetic analyses in the refuge distribution of *A. quercustozae* imply the presence of a cryptic sexual generation (Atkinson, 2000; Atkinson, *et al.*, 2001) and although the host of this putative sexual generation is unknown, phylogenetic evidence strongly suggests that it is from section *Cerris* (*Q. cerris* or *Q. suber*) (Cook, *et al.*, 2001, Chapter 7). Given that gallwasp distributions are ultimately dependent on their oak hosts' distributions, the current distribution and pattern of postglacial range expansion of *A. quercustozae* populations must have been shaped by the distributions of the host with the more restricted distribution. If the sexual generation host has a more limited distribution, we might expect to see a pattern similar to that for the marble gallwasp *A. kollari* (which also galls section *Cerris* oaks for its sexual generation, Stone, *et al.*, 2001, Appendix I), and would predict a range expansion of *A. quercustozae* populations from Italy and the Balkans following human planting of its sexual generation host (*Q. cerris*) in these regions (Fig. 5.1B). Furthermore, as a consequence of the natural range limit of its sexual generation host in Iberia, *Q. suber*, we would predict limited expansion from Iberia (Fig. 5.1B). In contrast, if the limiting factor of *A. quercustozae*' s distribution is the parthenogenetic generation oak host (*Q. pubescens*, *Q. pyrenaica*, etc.), the prediction would be that *A. quercustozae* has not expanded into northern Europe, despite the presence there of its postulated sexual generation host, *Q. cerris*, due to the poor postglacial range expansion of these oak species (Fig. 5.1A) (Jalas and Suominen, 1987).

In this study, we will address the following questions: (a) how many and which are the Pleistocene refuges for *A. quercustozae*?, (b) what is the importance of Anatolia as a refuge and putative centre of origin for *A. quercustozae* and how does genetic diversity change in relation to longitude?, (c) what is the distribution of *A. quercustozae* outside the refuge areas?, and (d) has *A. quercustozae*' s range expansion outside Pleistocene refuges been constrained by the oak hosts of either generation?

To answer these questions we sampled extensively *A. quercustozae* populations from Morocco in the west to the north-east of Anatolia and subjected the resulting samples to molecular analyses. More specifically, we obtained DNA sequences from a fragment of the mitochondrial locus cytochrome *b* from 47 individuals and allele frequencies on 12 polymorphic allozyme loci from 609 individuals.

5.3 Materials and Methods

5.3.1 Collection of specimens

Surveys for *Andricus quercustozae* parthenogenetic generation galls were performed, as part of a larger project on the population biology of oak gallwasps, in all autumns between 1990 and 2000 within (Morocco, Spain, Italy, Hungary, Slovenia, Greece and Turkey) and outside refuge areas (Germany, France, Belgium, the Netherlands, Austria, the Czech Republic, Switzerland, Sweden, the United Kingdom, Ireland and Denmark) (Stone and Sunnucks, 1993; Schönrogge, *et al.*, 1995; Stone, *et al.*, 1995; Sunnucks and Stone, 1996; Cook, *et al.*, 1998; Schönrogge, *et al.*, 1998; Stone and Cook, 1998; Atkinson, 2000; Rokas, *et al.*, 2001a; Stone, *et al.*, 2001, Chapter 4, Appendix I).

Table 5.1. Sampling locations and number of individuals sequenced for cytochrome *b* or electrophoresed for allozymes from each population. Degrees latitude and longitude are decimalized. Abbreviations: Seq.: number of individuals sequenced; Alloz.: number of individuals electrophoresed for allozyme variation; *: sequences from these individuals were cloned.

No. – Population	Country	Location latitude, longitude	Number of individuals	
			Seq.	Alloz.
1. Azrou	Morocco	33.45, -5.23	2	39
2. Barcelona	Spain	41.42, 2.17	1	0
3. Madrid	Spain	40.42, -3.72	1	9

4. Navacerrada	Spain	40.72, -4.02	1	0
5. Prado del Rey	Spain	36.8, -5.55	1	17
6. Santiago de Compostela	Spain	42.87, -8.55	1	0
7. Aire de Querane	France	45.37 , -0.97	1	0
8. Bordeaux	France	44.83, -0.57	1	12
9. Perigeux	France	45.12, 0.73	1	0
10. Tarbes	France	43.23, 0.08	1	9
11. Bombiana	Italy	44.2, 10.95	1	16
12. Casina	Italy	44.52, 10.5	1	0
13. Chiusi	Italy	43.03, 11.95	1	0
14. Felitto	Italy	40.37, 15.25	1	0
15. Gildone	Italy	41.5, 14.67	1	0
16. Greve in Chianti	Italy	43.58, 11.32	1	39
17. Jelsi	Italy	41.53, 14.8	1	28
18. Lame	Italy	44.63, 9.7	1	0
19. Massa Marritima	Italy	43.05, 10.88	1	27
20. Moio	Italy	40.15, 15.17	0	28
21. Monte Sant' Angelo	Italy	41.72, 15.97	0	40
22. Poppi	Italy	43.72, 11.77	1	39
23. Radicofani	Italy	42.9, 11.77	1	0
24. Rieti	Italy	42.4, 12.85	1	22
25. Salsomaggiore	Italy	44.19, 9.62	1	0
26. San Venanzo	Italy	42.87, 12.27	1	27
27. Lake Balaton	Hungary	47.1, 17.9	1	0
28. Eger	Hungary	47.88, 20.47	1	0
29. Gödöllö	Hungary	47.6, 19.33	2	0
30. Sopron	Hungary	47.67, 16.58	1	40
31. Szeghalom	Hungary	47.23, 16.70	1	39
32. Tatabanya	Hungary	47.52, 18.42	1	10
33. Veszprem	Hungary	47.1, 17.9	1	40

34. Arnissa	Greece	40.46, 21.56	3	8
35. Pisoderi	Greece	40.46, 21.13	2	0
36. Ağlasun	South-western Anatolia	37.65, 30.53	0	15
37. Gezende	South-western Anatolia	36.53, 33.15	1*	40
38. Küllüce	Central Anatolia	38.20, 34.60	2	7
39. Lysandra	South-western Anatolia	36.48, 30.05	1*	6
40. Madenli	South-western Anatolia	38.13, 31.02	0	40
41. Refahiye	North-eastern Anatolia	39.90, 38.75	3	7
42. Yeniyol	North-eastern Anatolia	41.4, 41.63	2	5

5.3.2 DNA sequence data

5.3.2.1 DNA extraction, amplification and sequencing

DNA was extracted by using the DNeasy Tissue kit (QIAGEN cat. 69504) and following the protocol of the manufacturer for insect DNA extraction. Sequencing was carried out using the Perkin-Elmer BigDye Terminator chemistry and an ABI 377 sequencer. PCR products for each species were sequenced in both directions to minimise PCR artefacts, ambiguities and base-calling errors.

A 433 base pair (bp) fragment of the mitochondrial cytochrome *b* gene was amplified and sequenced for 47 individuals (this includes the sequence published in Stone and Cook, 1998; Genbank accession no. AJ228467). Detailed molecular protocols used have been described elsewhere (e.g. Stone and Cook, 1998; Rokas, *et al.*, 2001a; Stone, *et al.*, 2001, Chapter 4, Appendix I). All sequences will be available from Genbank on publication.

Following an initial demonstration that multiple sequences were produced by PCR amplification for certain gallwasp specimens from Turkey (Gezende and Lysandra), a strategy of cloning was adopted. PCR products were cloned in a blunt-end vector following the manufacturer's instructions (Zero Blunt® TOPO® Cloning kit, Invitrogen, cat. K4500-01). Ten colonies from each individual were selected, cultured and purified (QIAprep spin miniprep kit, QIAGEN cat. 27104) for subsequent cloning.

5.3.2.2 Phylogenetic analysis

Sequences were aligned manually. Datasets with and without the haplotypes from the two Turkish wasps that gave multiple products and with and without outgroups were analysed. A total of six progressively more distant gallwasp species were used as outgroups, based on the phylogeny produced by Stone and Cook (1998). All sequence phylogenies were estimated by maximum likelihood (ML), using the PAUP* package, version 4.0b8 (Swofford, 2000). The best-fit ML model for each dataset was identified using MODELTEST, version 3.0 (Posada and Crandall, 1998). Modeltest utilises likelihood ratio tests (Huelsenbeck and Rannala, 1997; Lewis, 1998b) to identify the ML model of sequence evolution. Parameters allowed to vary in model-fitting were base composition, substitution rates and rate heterogeneity among sites. The models with the best-fit for all the datasets were rather similar and showed rather minor differences in parameter values (see legend of Fig. 5.2). The parameter values suggested by MODELTEST for each dataset were used to perform heuristic searches (with tree bisection and reconnection) on 100 bootstrap replicates.

The alignments for all datasets and topologies reported in this study will be available electronically from TreeBASE (http://www.herbaria.harvard.edu/treebase/) on publication.

5.3.3 Allozyme data

5.3.3.1 Allozyme electrophoresis

609 individuals were screened at 12 variable allozyme loci using cellulose acetate gel electrophoresis (zip zone equipment from Helena laboratories). The methods are

described in detail in Stone and Sunnucks (1993) and Stone *et al.* (2001, Appendix I). The following loci were used: αGPD1 & 2 (EC 1.1.1.8), GOT-m & -s (EC 2.6.1.1) HK (EC 2.7.1.1), PGM (EC 2.7.5.1), MDHm (EC 1.1.1.37), ME (EC 1.1.1.40), 6PGD (EC 1.1.1.44), AK (EC 2.7.4.3), GPI (EC 5.3.1.9), PEPb (EC 3.4.11). They were run on the buffers described in Richardson *et al.* (1986). HK and PGM were run on Tris Glycine pH 8.6; 6PGD, GPI, MDHm, ME and GOT were run on Tris-EDTA-Maleate-$MgCl_2$ pH 8.3; and PEPb, αGPD and AK were run on a phosphate buffer pH 6.1-6.3.

5.3.3.2 Allozyme data analysis

Genotypic data were analysed for deviation from Hardy-Weinberg equilibrium (HWE) and evidence for linkage disequilibrium (LD) using the multisample test in GENEPOP (Raymond and Rousset, 1995). By assuming that loci and populations are independent, this test allows the results to be summarised at the level of the population (HWE) and locus (LD) and the significance of the result estimated using an exact test. Deviation in any population from HWE or evidence for LD can then be traced back to its source. To correct for the multiple comparisons that occur in the calculations, we used a Dunn-Sidak correction with an experiment-wise error rate of 0.01.

Genetic diversity measures. Two genetic diversity measures (mean expected heterozygosity and average number of alleles per locus) were calculated for each population using GENEPOP (Raymond and Rousset, 1995) and GENETIX (Belkhir, *et al.*, 2000).

Cluster analysis. The data were analysed using two different clustering methods.

(a) Pairwise-distance matrix method. Phylogenies were constructed using three different criteria (neighbour joining, least squares and maximum likelihood) on two different pairwise distance measures; Nei' s genetic distance (Nei, 1972) and Cavalli-Sforza' s chord measure (Cavalli-Sforza and Edwards, 1967). For neighbour-joining and least squares, and on allele frequency data for maximum likelihood, 100

bootstrap replicates were generated for each method using PHYLIP (Felsenstein, 1993).

(b) Model-based method. The data were analysed using the programme STRUCTURE to infer the number of discrete genotypic pools the sampled individuals were drawn from (Pritchard, *et al.*, 2000). STRUCTURE assumes a model in which there is a specified number of pools and characterises each one by a set of allele frequencies derived from multilocus genotype data. Individuals are then assigned probabilistically to the pools under the assumptions of Hardy-Weinberg and linkage equilibrium using Markov Chain Monte Carlo (MCMC) simulation. The simulation is rerun for models specifying different numbers of pools and the posterior probabilities for each simulation are compared to infer the pool number that is best supported by the underlying genotypic data. We ran the model for pool sizes of 1-5, using only genotypic data from the refugial areas both including and not including any population admixture. Admixture allows for individuals to be assigned jointly to more than one pool if their genotype indicates a mixed ancestry, as may have occurred through mixing between refugial areas during interglacial periods. Having determined the best estimate for the number of pools, we reran the simulation to include individuals from the two populations outside of the refugial areas (Tarbes and Bordeaux). This approach allows us to separate the origin of recently established populations from the more ancient separation of refuges.

5.4 Results

5.4.1 Distribution of A. quercustozae *outside the refuge areas*

Parthenogenetic generation galls of *A. quercustozae* were found and collected from 40 localities in 6 countries known to have served as Pleistocene refuges across the Mediterranean from Morocco in the west to Turkey in the east (Table 5.1, Fig. 5.1). Outside refuge areas, *A. quercustozae* galls were only found in four localities in south-western France (Table 5.1, Fig. 5.1). Therefore, *A. quercustozae* has not expanded its range in the current interglacial into central and northern Europe, a

pattern observed in other oak gallwasps (e.g. Stone and Sunnucks, 1993; Atkinson, 2000; Rokas, *et al.*, 2001a; Stone, *et al.*, 2001, Chapter 4, Appendix I).

5.4.1.1 DNA sequence data

Sequence variation. For most individual wasps sequenced (45 out of 47 individuals), PCR amplification produced a single sequence. These 45 sequences yielded a total of 25 haplotypes (Table 5.2). 35 nucleotide sites (8.08%) were polymorphic, 24 of which were parsimony-informative (5.54%). Direct sequencing from cleaned PCR fragments from the two south-western Anatolian specimens (locations Gezende and Lysandra) suggested the existence of multiple cytochrome *b*-like products. Sequencing of ten clones from each of these specimens yielded a total of six distinct sequences (one sequence was shared between the two individuals). All the sequences had a correct reading frame. The length and genetic divergence was high both within (uncorrected: 0.46% - 4.85%) and between these two individuals (uncorrected: 0.23% - 5.31%) and even higher between them and the rest of the Anatolian samples (uncorrected: 3.93% - 7.62%). Given that all the cytochrome-*b* like sequences recovered from these two Turkish specimens have a correct reading frame and length, there are two alternative explanations: (a) one of the sequences is the 'true' mitochondrial copy with the others being mitochondrial pseudogenes in the nucleus (numts) (Bensasson, *et al.*, 2001), or (b) all of the sequences are recently (given that they are not, as yet, mutationally degenerated) derived numts from the 'true' mitochondrial copy. Our data do not allow discrimination between these two hypotheses, and the six distinct sequences (haplotypes) will be referred to subsequently in the text as 'pseudohaplotypes'. Other potential explanations for the existence of mitochondrial-like sequences additional to the 'true' mitochondrial copy include heteroplasmy and intra-mitochondrial duplications (Mirol, *et al.*, 2000; Bensasson, *et al.*, 2001); reasons why both of these alternatives are rather unlikely explanations in studies like this one have been thoroughly discussed by Mirol *et al.* (2000).

Table 5.2. Haplotypes for the 433 bp fragment of cytochrome *b* from *Andricus quercustozae*. Each column corresponds to a variable nucleotide site. The number above each column indicates the position of the nucleotide site along the 433 bp cytochrome *b* fragment. The parsimony informative sites are indicated by an asterisk in the first row of the table.

		1111111122	2222333333	33333
	111457889	0014679902	2356222223	34567
	1039354251	7922460650	6208012595	70849
Parsimony informative?	*** * ***	**** *	* ******	** *
Haplotype_1	TGCATGTGGA	CTTGTAAATT	TTTTGTTAGG	ATATA
Haplotype_2	....	.....G....	.........A	.....
Haplotype_3	....C.....	....	....	...C.
Haplotype_4	....	....	........A.	.....
Haplotype_5	.....A....	....	....	.....
Haplotype_6	??......A.	....	........A.	.....
Haplotype_7	.AT..A.AAG	TA.A...G..	.C.....G..	T....
Haplotype_8	?AT..A.AAG	TA.A..GG..	.C.....G..	T....
Haplotype_9	.AT....AAG	TA.A...G..	.C.....GA.	T....
Haplotype_10	.AT..A.AAG	TA.A...G..	.C.....GA.	T....
Haplotype_11	.AT....AAG	TA.A...G..	.C.....G..	T....
Haplotype_12	.AT....AAG	TA.A...G.C	.C.....G..	T....
Haplotype_13	.AT..C.AAG	TA.A...G..	.C........	T....
Haplotype_14	.AT..C.AAG	TA.A...G..	.......G..	T....
Haplotype_15	.AT..C.AAG	TA.A...G..	.C.....G..	T....
Haplotype_16	CAT..T.AA.	TA.A...G..	.C.....G..	T....
Haplotype_17	CAT..C.AA.	TA.A...G..	.C.....G..	T....
Haplotype_18	.AT..T.AA.	TA.AC..G..	.C.....G..	T....
Haplotype_19	.AT..C.AA.	TA.A...G..	.C.....G..	T....
Haplotype_20	.AT..C.AA.	TA.A...G..	.C.....GA.	TC...
Haplotype_21	.AT...CAAG	TA.A...G..	.C.....GA.	T.G..
Haplotype_22	.AT....A.G	TA.A......	......CG..	TC...

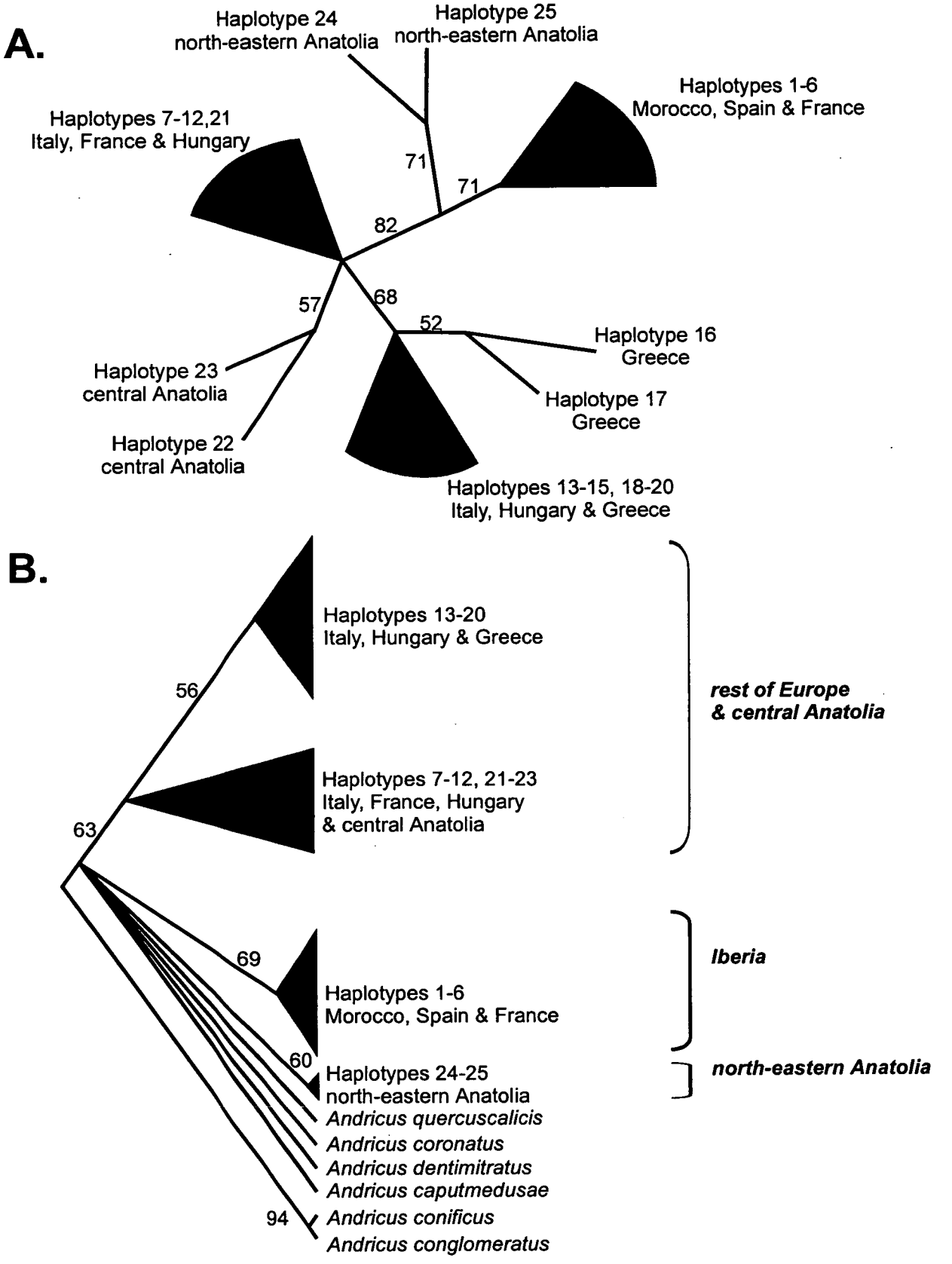
```
Haplotype_23      .AT....AAG TA.A...... ......CG.. TC...
Haplotype_24      .AT..A.AA. ..CA...... ...CAC..A. T...C
Haplotype_25      ..TG.A.AA. ..CA...GC. C.CCAC.GA. TC.C.
```

Haplotype locations: haplotypes 1-6 are from Iberia, Morocco and France, haplotypes 7-14 are mainly from Italy, haplotypes 15-17 from Greece, haplotypes 18-21 from the Balkans and haplotypes 22-25 from Anatolia. 1 – Barcelona, Prado del Rey, Santiago de Compostela (S), Tarbes (F); 2 – Madrid (S); 3 – Azrou (M); 4 – Azrou (M); 5 – Bordeaux (F); 6 – Perigeux (F); 7 – Lame (I); 8 – Casina (I); 9 – Chiusi, Salsomaggiore, San Venanzo (I); 10 – Jelsi (I); 11 – Aire de Querane (F), Felitto, Greve, Poppi, Radicofani, Rieti (I), Gödöllö, Sopron, Szeghalom (H); 12 – Bombiana (I); 13 – Greve in Chianti (I); 14 – Gildone, Massa Marritima (I); 15 – Arnissa (G); 16 – Pisoderi (G); 17 – Pisoderi (G); 18 – Lake Balaton (H); 19 – Eger, Gödöllö, Tatabanya (H), Arnissa (G); 20 – Veszprem (H); 21 – Sopron (H); 22 – Küllüce (T); 23 – Küllüce (T); 24 – Yeniyol (T); 25 – Refahiye (T).
Letters in parentheses indicate the country of origin (M: Morocco, S: Spain, F: France, I: Italy, H: Hungary, G: Greece, T: Turkey).

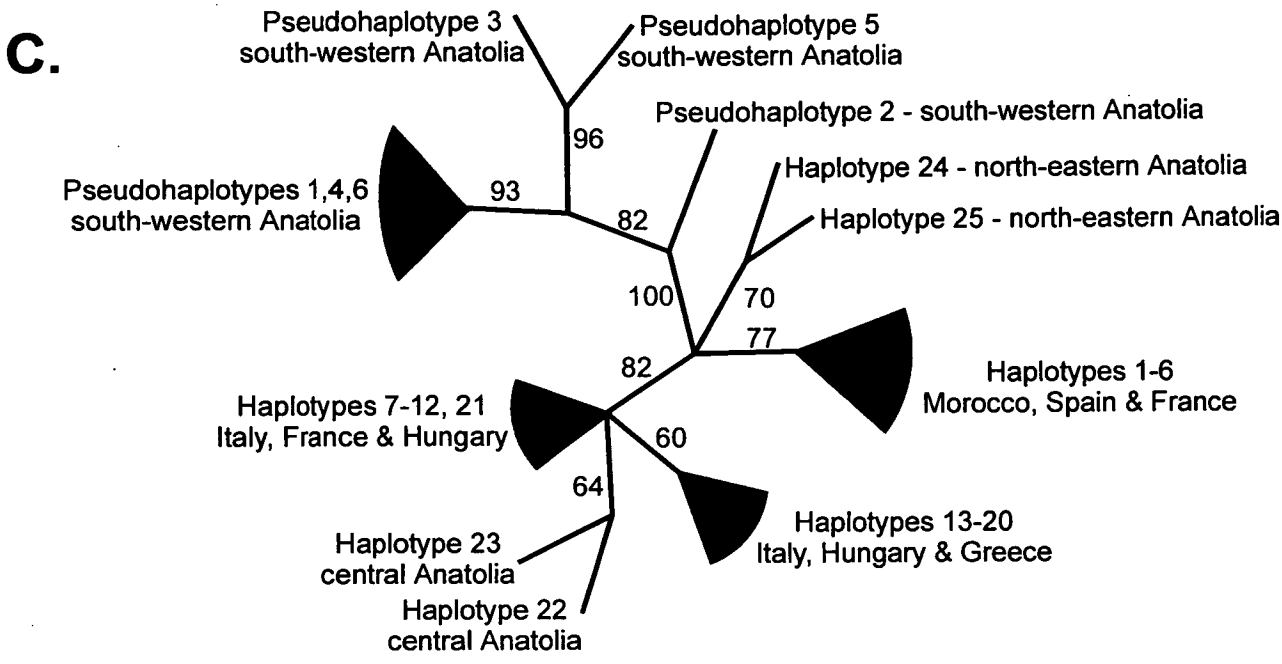
Phylogenetic patterns. Consensus trees, with and without outgroups and with and without the pseudohaplotypes, are shown in Figs. 5.2A-D. Although the details differ in these four reconstructions, four clades are strongly supported by most reconstructions (see Figs. 5.2B and 5.2D): (a) the Iberia clade (haplotypes 1-4), composed of individuals from Spain and Morocco; (b) the rest-of-Europe clade (haplotypes 7-23), composed of individuals from Italy, Hungary, Greece and central Anatolia; (c) the north-eastern-Anatolia clade (haplotypes 24-25), composed of two individuals from north-eastern Turkey; and (d) the south-west Anatolia pseudohaplotypes clade (pseudohaplotypes 1-6), composed of all the haplotypes recovered from the two south-western individuals that showed multiple haplotypes. Each of the first three clades may perhaps be considered as a distinct refuge. The fourth clade cannot be regarded as a distinct refuge on the basis of mitochondrial data, unless we assume that one of the pseudohaplotypes is the 'true' mitochondrial copy (if all the pseudohaplotypes are numts then the possibility that they are also distinct from the 'true' mitochondrial haplotypes of this region cannot be rejected). The south-western French populations, the only area where *A. quercustozae* is found

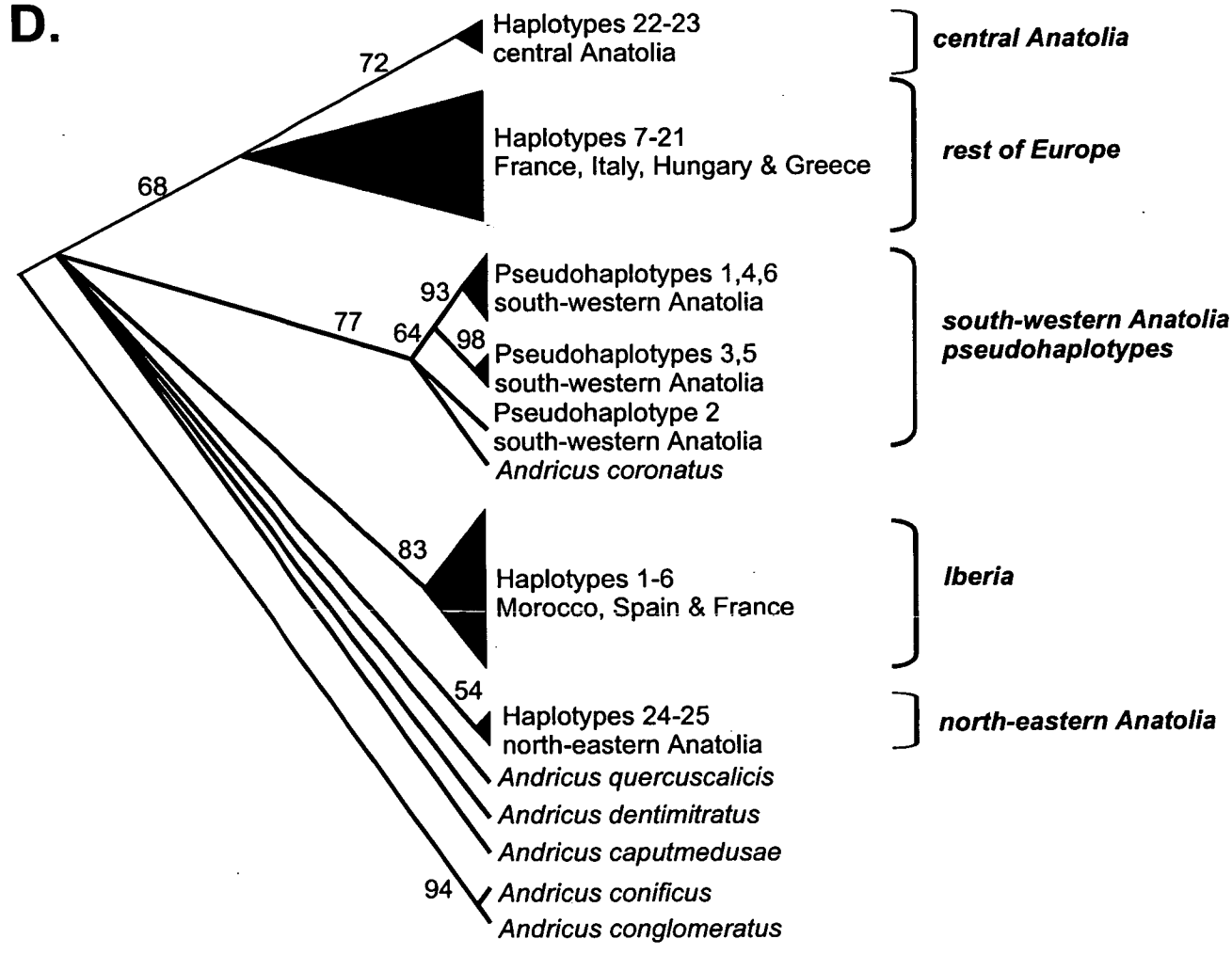
outside of the putative refugial areas, predominantly comprised haplotypes placed within the Iberia clade (haplotypes 1, 5, 6), however one French individual contained a haplotype belonging to the rest of Europe clade (haplotype 11). Six progressively more genetically distant *Andricus* species were used as outgroups either singly or together. In the two rooted trees (Figs. 5.2B, 5.2D), the four *A. quercustozae* clades did not form a monophyletic group in the exclusion of the outgroup sequences but were placed in a basal multifurcating (polytomous) node together with outgroup species of the *Andricus quercuscalicis* clade *sensu* Stone and Cook (1998, see also Chapter 6); these are *A. quercuscalicis*, *A. coronatus, A. dentimitratus* and *A. caputmedusae*. Occasionally, single outgroup species from this clade grouped together with the pseudohaplotypes clade or the north-eastern Anatolian clade (as for example in Fig. 5.2D and in data not shown), probably due to phylogenetic error associated with high sequence divergence (long-branch attraction). None of these groupings were persistent in the different combinations of outgroups performed (data not shown).

The two outgroup taxa outside the *A. quercuscalicis* clade (*A. conificus*, *A. conglomeratus*) were always placed as a monophyletic outgroup but it was not possible to identify which of the four *A. quercustozae* clades is basal. The bootstrap support for the various groupings was always below 50% and it was unaffected by the use of different combinations of outgroup species.



igure 5.2. (A, B) Unrooted and rooted trees of the cytochrome *b* haplotypes of *A. quercustozae* without the seudohaplotypes. (C, D - see next page) Unrooted and rooted trees of the cytochrome *b* haplotypes with the seudohaplotypes. Outgroup species are from Stone and Cook (1998). All trees are 50% majority-rule onsensus trees that have been constructed using a ML algorithm on 100 bootstrap replications with the arameter values suggested by MODELTEST (Posada and Crandall, 1998). The best-fit model in all four datasets ncluded unequal base frequencies (AT-bias ~ 77%), a transition / transversion ratio (ranging from 6.3 to 18.7) nd rate heterogeneity among sites (gamma distribution shape parameter α ranged from 0.82 to 1.25 and the roportion of invariable sites ranged from 0.72 to 0.84). Values above branches indicate bootstrap values.


C.
Pseudohaplotype 3
south-western Anatolia
Pseudohaplotype 5
south-western Anatolia
Pseudohaplotype 2 - south-western Anatolia
Haplotype 24 - north-eastern Anatolia
Haplotype 25 - north-eastern Anatolia
Pseudohaplotypes 1,4,6
south-western Anatolia
96
93
82
100
70
77
82
Haplotypes 1-6
Morocco, Spain & France
Haplotypes 7-12, 21
Italy, France & Hungary
60
64
Haplotypes 13-20
Italy, Hungary & Greece
Haplotype 23
central Anatolia
Haplotype 22
central Anatolia


D.
Haplotypes 22-23
central Anatolia
central Anatolia
72
Haplotypes 7-21
France, Italy, Hungary & Greece
rest of Europe
68
Pseudohaplotypes 1,4,6
south-western Anatolia
93
77
64
98
Pseudohaplotypes 3,5
south-western Anatolia
Pseudohaplotype 2
south-western Anatolia
Andricus coronatus
south-western Anatolia
pseudohaplotypes
83
Haplotypes 1-6
Morocco, Spain & France
Iberia
54
Haplotypes 24-25
north-eastern Anatolia
north-eastern Anatolia
Andricus quercuscalicis
Andricus dentimitratus
Andricus caputmedusae
94
Andricus conificus
Andricus conglomeratus

5.4.1.2 Allozyme data

There was no evidence for significant deviation from Hardy-Weinberg equilibrium in any population, or evidence for linkage disequilibrium between loci.

Genetic diversity. The average expected heterozygosity ($\hat{H}_{exp}$) shows a marked decrease with longitude westwards from Anatolia ($\hat{H}_{exp}$=0.24-0.30) to Iberia ($\hat{H}_{exp}$=0.07-0.08) (Fig. 5.3). The mean number of alleles also decreases (halving from a value of 2.83 in Ağlasun to a value of 1.42 in Azrou, see Fig. 5.3) although this measure has a wider variance due to its high sensitivity to sample size (the latter being correlated with population-specific rare alleles, Table 5.3).

Twelve population-specific (private) alleles were found over the sampled area, nine of which were restricted to Anatolia. Five are restricted to populations in the south-west (three uniquely to Madenli and two shared by populations in this area), the other four are only found in one population in the north-east; three in Küllüce and one in Refahiye. Of the remaining three alleles, two are private to Greve in Chianti (Italy) and one to Veszprem (Hungary). Of the twelve private alleles, only two may be refuge-specific (the two found in south-western Anatolian populations).

Despite the apparent lack of refuge-specific alleles, the most common alleles at several loci appear in frequencies distinct to each refuge. For example, there are two common alleles at locus MDHm. In Hungarian populations, only one occurs commonly, in Italian populations the two alleles are found at frequencies of 0.6 and 0.4, whereas in Iberian populations the latter ratio is reversed. Similar refuge-specific patterns are also evident at six other loci (GOTs, GPI, HK, PGM, HK and PEPb, Appendix 5.1).

Cluster analysis. (a) Pairwise distance-matrix methods. The topologies produced from each method of phylogenetic reconstruction were very similar (data not shown), and identify at least five distinct groups in the refugial areas; (a) Iberia & Morocco; (b) Italy; (c) Hungary; (d) south-western Anatolia and (e) north-eastern and central Anatolia. Populations outside the refugial areas (French populations) cluster most closely with the Iberian refuge. Interestingly, the Iberian refuge appears to cluster more closely to the Hungarian refuge than the Italian one, despite the fact that Italy and Iberia are geographically closer together.

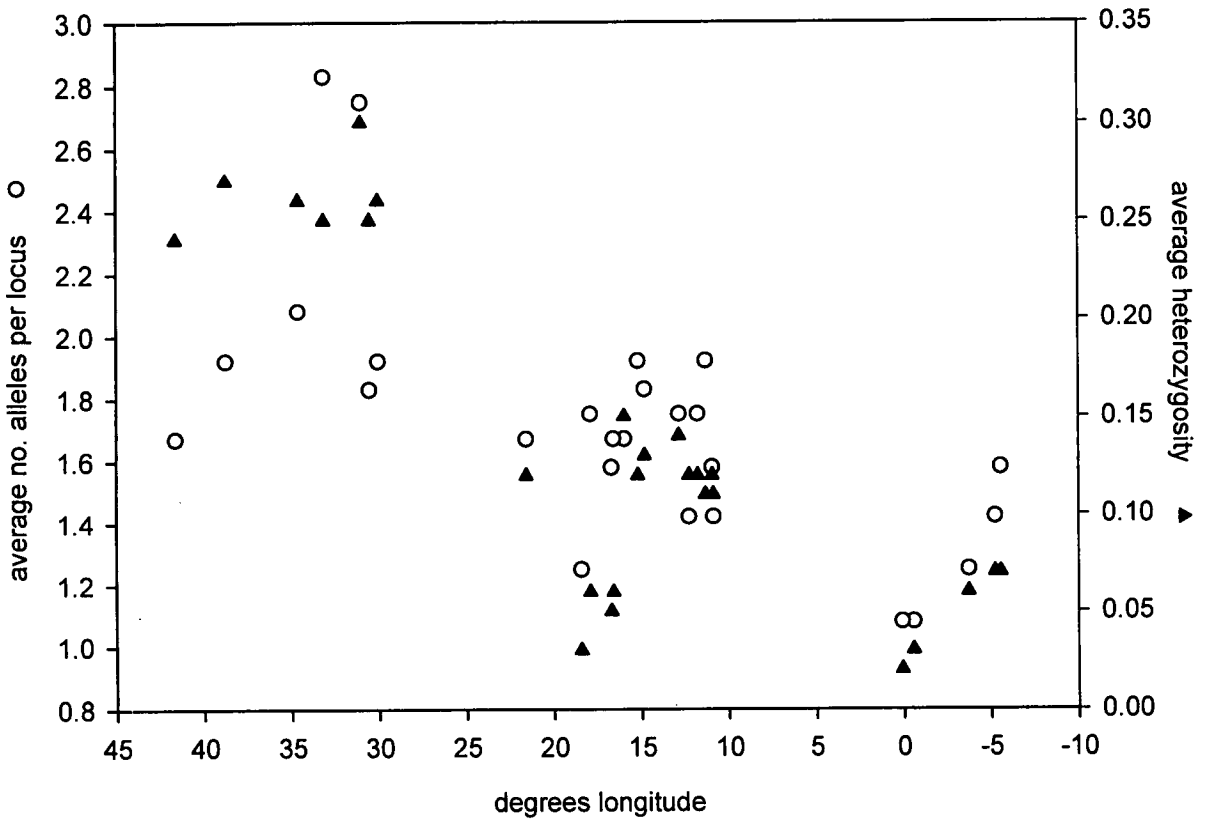


Figure 5.3. The correlation between longitude (decimalized) and genetic diversity indices (expected heterozygosity ○, mean number of alleles per locus ▲) in the allozyme allele frequency data of the *A. quercustozae* populations.

Table 5.3. Measure of the genetic diversity from allozyme allele frequency data. Populations are numbered according to Table 5.1. $\hat{H}_{exp}$: expected heterozygosity; Numbers in parentheses indicate standard errors.

No. – Population	Country	N	$\hat{H}_{exp}$	Alleles / locus
1. Azrou	Morocco	39	0.07 (0.02)	1.42
3. Madrid	Spain	9	0.06 (0.05)	1.25
5. Prado del Rey	Spain	17	0.07 (0.03)	1.58
8. Bordeaux	France	12	0.03 (0.03)	1.08
10. Tarbes	France	9	0.02 (0.03)	1.08
11. Bombiana	Italy	16	0.12 (0.05)	1.58
16. Greve in Chianti	Italy	39	0.11 (0.03)	1.92
17. Jelsi	Italy	28	0.13 (0.04)	1.83
19. Massa Marritima	Italy	27	0.11 (0.03)	1.42
20. Moio	Italy	28	0.12 (0.03)	1.92
21. Monte Sant' Angelo	Italy	40	0.15 (0.03)	1.67
22. Poppi	Italy	39	0.12 (0.03)	1.75
24. Rieti	Italy	22	0.14 (0.04)	1.75
26. San Venanzo	Italy	27	0.12 (0.03)	1.42
30. Sopron	Hungary	40	0.06 (0.02)	1.67
31. Szeghalom	Hungary	39	0.05 (0.01)	1.58
32. Tatabanya	Hungary	10	0.03 (0.02)	1.25
33. Veszprem	Hungary	40	0.06 (0.01)	1.75
34. Arnissa	Greece	8	0.12 (0.06)	1.67
36. Ağlasun	Anatolia	15	0.25 (0.07)	1.83
37. Gezende	Anatolia	40	0.25 (0.04)	2.83
38. Küllüce	Anatolia	7	0.26 (0.11)	2.08
39. Lysandra	Anatolia	6	0.26 (0.12)	1.92
40. Madenli	Anatolia	40	0.30 (0.04)	2.75
41. Refahiye	Anatolia	7	0.27 (0.11)	1.92
42. Yeniyol	Anatolia	5	0.24 (0.12)	1.67

(b) Model-based method. The best supported estimate for the number of genotypic pools (K) is four (Table 5.4). The addition of admixture made no difference to this result. The four pools can be characterised as (a) Italy, (b) Hungary & Iberia, (c) north-eastern and central Anatolia & Greece, and (d) south-western Anatolia. The addition of a fifth pool did not result in a group dominated by individuals from Iberia (as might be expected from our understanding of refugial populations), instead the pool comprised wasps from Italy.

Table 5.4. Conditional and posterior probabilities from STRUCTURE associated with each pool number (K) from a model with and without admixture.

K	With admixture		Without admixture	
	lnPr(X|K) [a]	Pr(K|X) [b]	lnPr(X|K)	Pr(K|X)
1	-5061	~0	-5061	~0
2	-4157	~0	-4138	~0
3	-4040	~0	-4009	~0
4	-3859	~1	-3795	~1
5	-4061	~0	-3954	~0

[a] The estimated conditional probability of observing the data under a model with the specified number of pools.

[b] The posterior probability (calculated from Bayes' rule with a uniform prior) to compare the relative likelihood of each pool number given the underlying genotypic information.

The probability of each individual belonging to each of the four pools is illustrated in Fig. 5.5. This shows clearly that the two Anatolian pools are the best resolved, while many Italian individuals show high probabilities of being associated with the Hungarian pool (Fig. 5.5). In addition, although the majority of individuals from Iberia show a higher probability of association with the Hungarian pool, a minority show a higher probability of association with the Italian pool (as illustrated by the scatter of Iberian individuals along the axis linking these two refuges).

In simulations including non-refugial populations, individuals from the two French populations Tarbes and Bordeaux show highest probability of membership to the pool characterised by Hungarian and Iberian individuals.

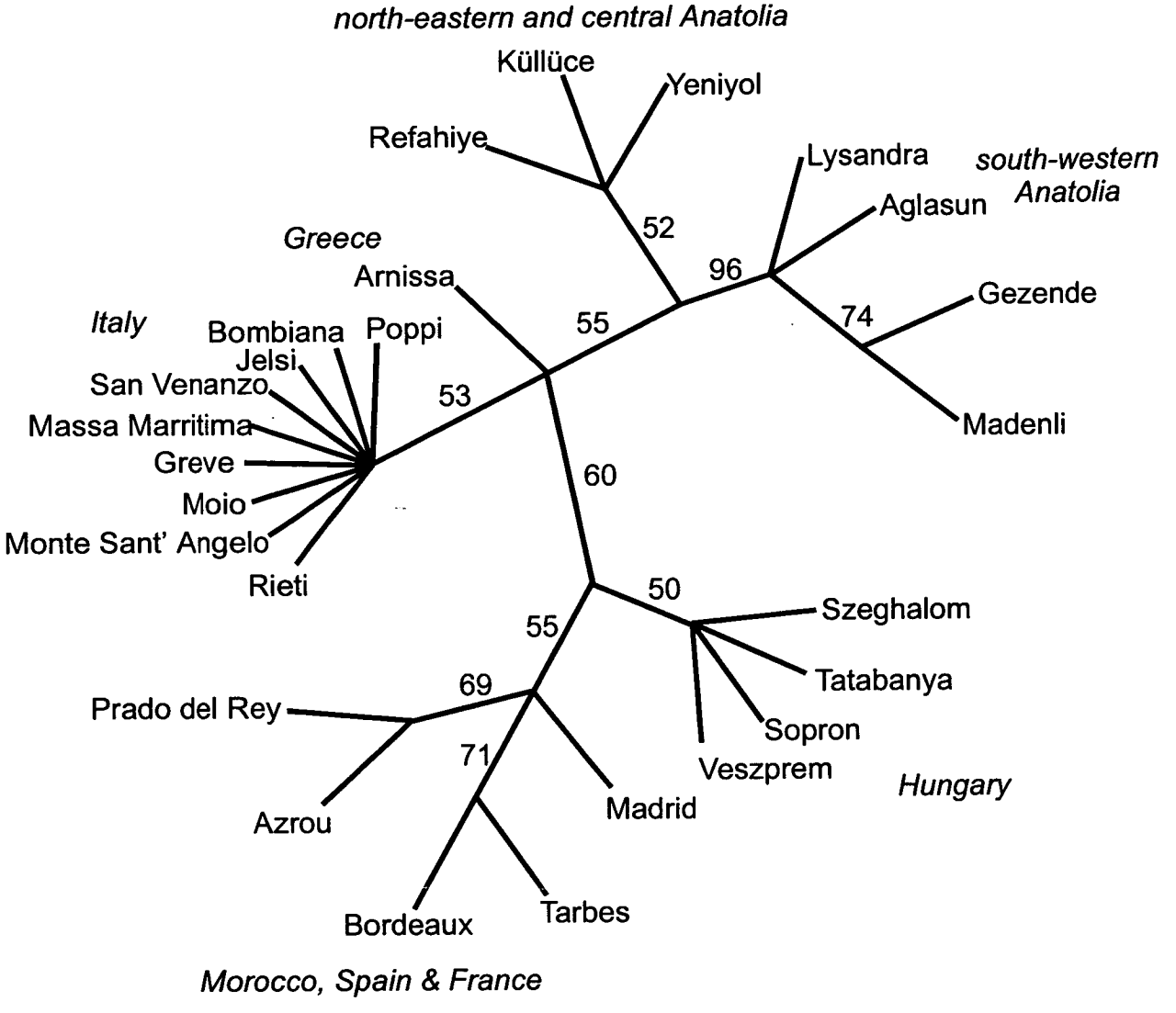


Figure 5.4. A 50% majority-rule consensus tree based on allozyme allele frequencies in *A. quercustozae* populations. The phylogeny was generated using the Least squares criterion on Cavalli-Sforza' s chord distance (Cavalli-Sforza and Edwards 1967).

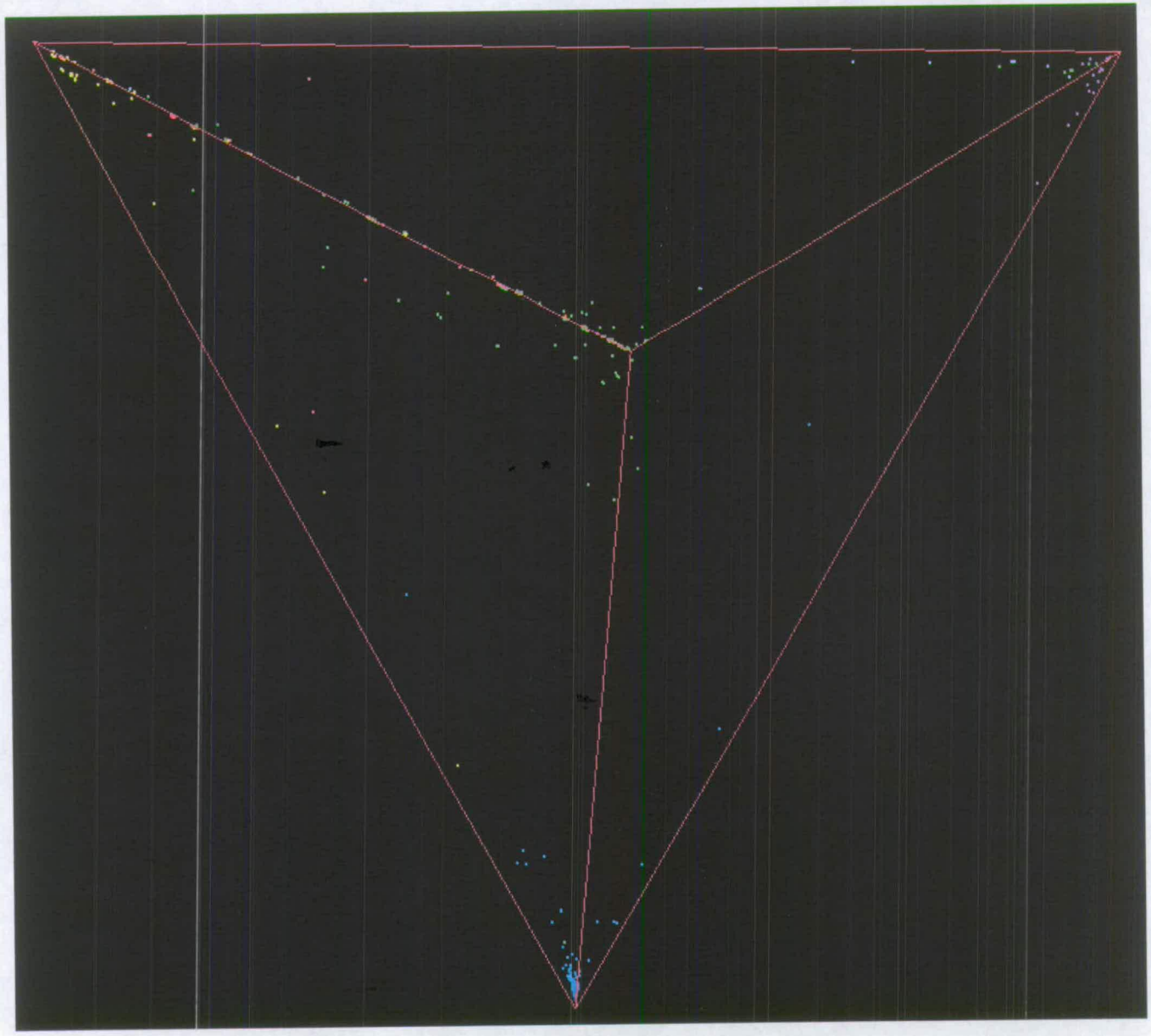


Figure 5.5. Clustering of data from the STRUCTURE analysis with four discrete pools. Each axis represents the probability of membership into either of the four groups (represented by the four corners of the tetrahedron). The points are colour-coded by their country of origin. Red = Morocco and Spain, Yellow = Hungary, Green = Italy, Pink = north-eastern and central Anatolia, Blue = south-western Anatolia, White = France.

5.5 Discussion

5.5.1 *Allozyme- and sequence-based phylogeographic patterns of* A. quercustozae

Both mitochondrial and allozyme datasets support the same basic pattern of refugial distinction of *A. quercustozae* populations (Iberia, Italy, the Balkans and at least two refuges in Anatolia). However, differences in the resolution offered by the two datasets also exist, perhaps indicating the difference in the nature of the two markers (different inheritance, sampling, analysis). For example, whereas the allozyme dataset suggests that Hungary and Italy represent two distinct refuges and that the Iberian refuge is not clearly separated from the Hungarian refuge, the sequence data do not differentiate between Hungarian and Italian haplotypes and show high levels of divergence for the Iberian haplotypes when compared to haplotypes in the rest of Europe.

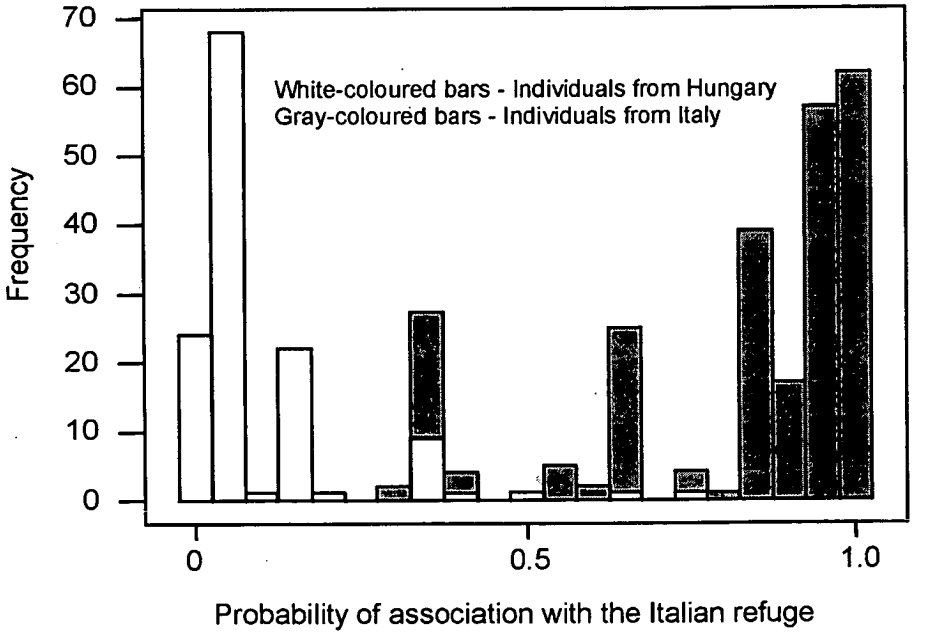


Figure 5.6. Association probabilities of Hungarian and Italian individuals to the Hungarian (x=0) and Italian (x=1) refuges.

The failure of phylogenetic analysis of cytochrome *b* haplotypes to identify Italian and Hungarian populations as distinct refuges (Fig. 5.2), although also observed in other oak gallwasp species (Rokas, *et al.*, 2001a; Stone, *et al.*, 2001, Chapter 4, Appendix I), is in sharp contrast with those studies on a range of taxa, including gallwasps' hosts – the oaks, that provide evidence that Italy and the Balkans have served as distinct refuges (e.g. Cooper, *et al.*, 1995; Dumolin-Lapegue, *et al.*, 1997; Ferris, *et al.*, 1998). We suggest that the allozyme data from individuals of these two countries may be able to offer an explanation for the mitochondrial patterns observed. The presence of refuge-specific signatures in terms of allele frequencies in individuals from Italy and Hungary is very clear in the allozyme data. Nevertheless, as shown in Fig. 5.6, a number of Italian individuals contain Hungary-specific signatures and vice-versa. But when the data are averaged over all sampled loci within an Italian individual or over all individuals sampled within an Italian population, the signature specific to Italy predominates. This mixing of refuge-specific signatures is suggestive of interbreeding (or admixture) between individuals from the two refuges. Furthermore, the mixing process must have started occurring relatively recently since the underlying refuge-specific signatures are still identifiable. In the light of this evidence, two hypotheses may explain the failure of mitochondrial haplotypes to distinguish between the two refuges. One possibility is that the haplotypes associated with one of the two refuges have been driven to fixation recently (through selection or genetic drift) in both refuges (through the interbreeding of the two refuge populations), leading to the extinction of the haplotypes associated with the other refuge. Loss of mitochondrial haplotypes is much easier than loss of nuclear haplotypes due to their different patterns of inheritance (maternal *versus* Mendelian) and population size (the mitochondrial effective population size is ¼ of the nuclear one). Discordant patterns between mitochondrial and nuclear markers have been also observed in other species. For example, in the honeybee *Apis mellifera* the African mitochondrial haplotypes have diffused into southern Europe but, crucially, this mitochondrial diffusion did not lead to the 'Africanization' of the nuclear markers (Franck, *et al.*, 2001). Another possibility is that the split between the two refuges happened towards the end of the Pleistocene, too recently for accumulation of sufficient mitochondrial variation,

although the presence of sufficient variation in the allozyme loci argues against this scenario.

Phylogeographic data for mitochondrial haplotypes and allozymes exist for two other oak gallwasp species, *B. pallida* and *A. kollari* (Rokas, *et al.*, 2001a; Stone, *et al.*, 2001, Chapter 4, Appendix I). However, in these two datasets both allozyme and mitochondrial markers argue against the distinction between the Italian and Hungarian refuges (Rokas, *et al.*, 2001a; Stone, *et al.*, 2001, Chapter 4, Appendix I). A potential explanation for this discordance between these two species and *A. quercustozae* may be that analysis of the allozyme allele frequency data in these species has been carried out only at the population level and not at the level of a single individual, as is the case for *A. quercustozae* (Fig. 5.6).

The distinction of Iberian (Morocco and Spain) and Hungarian refuges by the mitochondrial and allozyme data shows the reverse trend to that of refuges in Hungary and Italy; in this case, the mitochondrial haplotypes provide more resolution than the allozyme data. This may simply be an artefact of the low levels of genetic diversity seen in the allozyme data, perhaps due to repeated bottlenecking of Iberian populations (see genetic diversity estimates in Table 5.3). Although there is some evidence for refuge-specific signatures from allele frequencies, it is likely that bottlenecking events in Iberian populations may have eroded any clearer pattern resulting in lack of differentiation when compared with other refuges, as for example seen in Iberian populations of *A. kollari* (Stone, *et al.*, 2001, Appendix I).

5.5.1.1 The south-western Anatolian pseudohaplotypes are derived from truly distinct regional haplotypes

Cloning of the PCR product from the two individuals which failed direct sequencing for the cytochrome *b* fragment, revealed the presence of multiple pseudohaplotypes within each individual. Most or all of these pseudohaplotypes are probably nuclear mitochondrial pseudogenes (numts) (Bensasson, *et al.*, 2001). These pseudohaplotypes do not contain indels and/or stop codons and were not highly divergent in relation to the other haplotypes, thus the possibility that one of them represents the 'true' mitochondrial copy cannot be rejected. Furthermore, the recovery of these multiple 'functional' sequences makes the identification of the

'true' mitochondrial copy difficult. This is in contrast with the majority of previously published cases (e.g. Bensasson, *et al.*, 2000; Mirol, *et al.*, 2000; Stone, *et al.*, 2001, Appendix I), where it is generally straightforward to identify the 'true' mitochondrial copy due to the mutational degeneration of the numts. However, all the south-western Anatolian pseudohaplotypes form a distinct lineage (see Figs. 5.2C-D) and the allozyme data also support the existence of a distinct south-western Anatolia clade. Therefore, although most, or all, of the pseudohaplotypes from this region may be numts, the allozyme data strongly suggest that these pseudohaplotypes are derived from truly distinct regional haplotypes (or alternatively, one of the sequences is the 'true' mitochondrial haplotype).

5.5.2 The importance of Anatolia

The importance of Anatolia for the origin of European species is increasingly apparent in the published literature, as suggested by the patterns observed in a number of taxa, such as the black alder *Alnus glutinosa* (King and Ferris, 1998), the grasshopper *Chorthippus parallelus* (Cooper, *et al.*, 1995), the honeybee *Apis mellifera* (Franck, *et al.*, 2001) and the domestic mouse *Mus musculus* (Boursot, *et al.*, 1993). Despite the conclusions drawn from these studies, there are still relatively few data available for the eastern-most distributions of many well-studied species. The lack of sampling from these areas is perhaps indicative of the bias in interest into how genetic diversity varies with latitude rather than longitude. However, such a bias may be underestimating the contribution of putative easternmost refuges localised in Europe and / or Asia (Taberlet, *et al.*, 1998). While the haplotype data for *A. quercustozae* cannot resolve which refuge is basal to the others, the allozyme allele frequency data are suggestive of a decrease in genetic diversity along a latitudinal gradient from Anatolia in the east to Iberia and Morocco in the west. This pattern is seen in a decrease in the number of population-specific alleles, the average expected heterozygosity and the mean number of alleles per locus (Table 5.3, Fig. 5.3). It is a pattern often associated with range expansions (e.g. Stone and Sunnucks, 1993; Hewitt, 1999; Atkinson, 2000; Avise, 2000) brought about by the steady erosion of allelic diversity through a succession of bottlenecking events (Luikart, *et*

al., 1998). This is likely to have been the process causing the patterns seen in *A. quercustozae*, although the clear distinction between refuges indicated by refuge-specific patterns in allele frequencies and population-specific alleles in addition to the phylogenetic and clustering analyses provides evidence that the refugial populations were founded a long time ago (perhaps before the Pleistocene, Stone, *et al.*, 2001, Appendix I, Rokas, *et al.*, unpublished data) and have remained isolated until recently (see also Taberlet, *et al.*, 1998).

The distinctness of the Anatolian *A. quercustozae* populations, as indicated by the presence of many unique allozyme alleles and mitochondrial haplotypes as well as by the morphological differences between the galls from the two regions, are suggestive of a prolonged isolation from the European populations studied. Populations from Anatolia exhibited the highest level of genotypic diversity of any sampled populations across the species range (see Table 5.3), a pattern also found in other oak gallwasp species (Atkinson, 2000). The high species richness of gallwasps and their oak hosts in Anatolia coupled with the high genetic diversity of among Anatolian populations, perhaps argue for an eastern origin for the European radiation of *Andricus* oak gallwasps. However intriguing and biologically plausible such an 'out-of-Anatolia' radiation scenario might be for *A. quercustozae* (as well as for European oak gallwasps and other taxa in general, Boursot, *et al.*, 1993; Cooper, *et al.*, 1995; King and Ferris, 1998; Franck, *et al.*, 2001), two caveats merit further discussion. Firstly, the higher genetic diversity estimates in Anatolia (Fig. 5.3) are not necessarily proof of an 'out-of-Anatolia' radiation. These populations may have higher diversity because they have been less influenced by bottlenecks associated with Quaternary ice episodes than the European refuges. Secondly, attempts for rooting the generated *A. quercustozae* haplotype phylogenies, were not conclusive as to which one of the refuges is the most basal (Figs. 5.2B, 5.2D), perhaps indicating that divergence and subsequent radiation may have occurred in a short period of time, or that none of the sampled populations was the epicentre of the radiation.

The isolation of Anatolia appears to have been significant enough to affect the phenotype of *A. quercustozae* galls. Galls found in most of Europe are only sticky during growth but not when mature (G. Stone, personal observations), while galls found in Albania, Anatolia, Palestine and Asia Minor remain sticky throughout their

ontogeny (Dalla Torre and Kieffer, 1910). The latter phenotype was previously thought to represent a different species, *Andricus insana* (Dalla Torre and Kieffer, 1910), but detailed studies of the adult wasp morphology (Melika, *et al.*, 2000) and molecular phylogenetics (this study, Rokas, *et al.*, unpublished data) suggest otherwise. Furthermore, it is unclear whether the different phenotypes are caused by host-parasite interactions or by differences in the gallwasp genotypes *per se*, since *A. quercustozae* is using a different oak host in the eastern part of the Mediterranean (*Q. infectoria*, a species absent from central and western Europe).

The long-standing genetic isolation of Anatolian populations from the rest of Europe may be explained by the separation of the two areas by substantial mountain ranges and the Black and Mediterranean seas (Hewitt, 1999). Additionally, the isolation is compounded by the presence of populations in Greece, Bulgaria and Romania, so that range expansion into Europe during an interglacial could potentially be thwarted by the meeting of a region at carrying capacity for the species involved (Hewitt, 1999). This might also be the reason for the relatively close affinity between certain Greek and north-eastern Anatolian individuals/populations (Figs. 5.2, 5.4, 5.5), although there is also some evidence for trade of this and other species between the northern Anatolia and the Mediterranean.

Anatolia represents the meeting place of three phytogeographical regions; the Irano-Turanian, the Euro-Siberian and the Mediterranean region, each with a distinct climate and flora. The Irano-Turanian region which covers the majority of the interior of Anatolia, stretching east from the beginnings of the Central Anatolian Plateau into Iran is further bisected by the Anatolian diagonal, a line of mountains running laterally from the Anti-Taurus north-east to the Black Sea coast near Trabzon. This represents an important floristic divide separating a predominantly Asiatic flora from one with characteristics of Europe, a divide that is also visible in the distribution of oak species. The populations of oak gallwasps from Anatolia fall into two clusters in both mitochondrial and allozyme analyses, perhaps representing two separate refuges on either side of this divide. Alternatively, the observed clustering within Anatolia may be an artefact of our limited within-Anatolia geographic sampling, created by sampling only two (out of many) localised areas that have been generated by the mountainous nature of the Anatolian landscape.

5.5.3 The origin of populations outside of the refuge areas

Range expansion patterns following the retreat of the ice sheet at the end of the last glacial period vary among gallwasp species, perhaps reflecting the constraints imposed on them by the distribution of their oak hosts. For example, *B. pallida*, which has both generations on white oaks of the *Q. petraea* / *Q. robur* type, probably expanded its range into central and northern Europe as soon as the ice melted ten thousand years ago, tracking its oak hosts (Dumolin-Lapegue, *et al.*, 1997; Rokas, *et al.*, 2001a, Chapter 4). In contrast, *A. quercuscalicis* and *A. kollari* expanded their ranges only in the last few centuries following the planting of their sexual generation host, the Turkey oak (*Q. cerris*), through central and northern Europe (Fig. 5.1B) (Stone and Sunnucks, 1993; Stone, *et al.*, 2001, Appendix I). Other species, including *A. quercustozae*, have either not expanded their range from putative refugial areas, or are expanding northwards very slowly (Atkinson, 2000).

Individuals from the French populations of *A. quercustozae* show both allozyme allele frequency and mitochondrial haplotype similarities with Iberian individuals (although not all of the specimens), suggesting that the most likely point of origin for the colonisation of south-western France is Iberia. The distribution of *A. quercustozae* outside its native range is reminiscent of the Iberian race of *A. kollari*, their northernmost limit being the south-west of France. Stone *et al.* (2001, Appendix I) have suggested that Iberian *A. kollari* has been restricted in its range expansion by the distribution of its sexual generation host in this region, *Q. suber*. This species is endemic to Iberia and Italy and is only found north of the Pyrenees along the west coast of France as far north as Bordeaux (Fig. 5.1B). Although the sexual generation host of *A. quercustozae* remains unknown, the high likelihood of it being a section *Cerris* oak and the similarity in distribution of Iberian *A. quercustozae* and *A. kollari* suggests that this species may also be utilising *Q. suber*. *A. kollari* utilises *Q. cerris* as its sexual generation host across the rest of its range and, as mentioned above it is found extensively over much of the anthropogenic range of *Q. cerris* in northern Europe. Given that *A. quercustozae* probably has a similar lifecycle outside Iberia, it is unclear why this species has not also escaped the

Italian and Balkan refuges. The ability of other *Q. cerris*-dependent *Andricus* species to expand their ranges (Schönrogge, *et al.*, 1998), suggests that the distribution of *A. quercustozae* in eastern and central Europe is not likely to be restricted by the distribution of its sexual generation host, *Q. cerris*.

Another possible barrier to range expansion may be the wasp's specificity to its parthenogenetic generation hosts. *A. quercustozae* is rarely found on *Q. robur* or *Q. petraea* (the two most common species in northern Europe) (Dumolin-Lapegue, *et al.*, 1997), preferring instead species related to *Q. pubescens*. Only two species from this complex have expanded their ranges to any extent - *Q. pubescens* and *Q. pyrenaica* (Fig. 5.1A). The former is patchily distributed and uncommon outside Italy and the Balkans, being restricted to south-facing slopes on very alkaline soils, while *Q. pyrenaica* (an Iberian endemic), has followed a pattern of range expansion similar to that of *Q. suber*, its natural distribution extending only as far north as south-western France. The distribution of these two species coincides closely with the current distribution of *A. quercustozae*, suggesting that, in contrast to *A. kollari* and *A. quercuscalicis* (Stone and Sunnucks, 1993; Stone, *et al.*, 2001, Appendix I), northwards range expansion in *A. quercustozae* has been limited by the post-glacial history of its parthenogenetic generation host.

Host specificity, apart from providing an explanation the lack of range expansion of *A. quercustozae*, might provide an additional mechanism for isolation of refugial populations. If the pattern of oak tree range expansion in the present interglacial is typical of other interglacials throughout the Pleistocene, the high levels of differentiation between refuge populations may have been maintained simply because the species never left these areas during the milder interglacial periods. If so, a testable expectation would be that gallwasp species galling *Q. pubescens* / *Q. pyrenaica* oaks should show a greater refugial distinction when compared to those able to gall *Q. petraea* / *Q. robur* oaks (assuming that post-glacial patterns in the current interglacial accurately reflect taxonomic differences in past interglacials too). It should be noted that the recent movement of individuals between Hungary and Italy does not contradict the host specificity hypothesis, but rather points to a recent recolonisation of the foothills of the dolomites by suitable oak trees, thus opening up a corridor for gene exchange.

In conclusion, the evidence presented for *A. quercustozae* points strongly to a lack of sexual generation host limitation and much more to limitation by the parthenogenetic generation host(s). The inability to switch host in parthenogenetic generation from a *Q. pubescens / Q. pyrenaica* group host to a *Q. robur / Q. petraea* oak group, is consistent with the general rarity of host shifts in oak cynipids and may explain the inability of other host alternating species to follow their sexual generation host oaks (section *Cerris*) in northern Europe (e.g. *A. gemmeus*, *A. coriarius*, *A. caputmedusae*).

5.6 Conclusions

The Quaternary ice ages had a profound impact on the genetic structure of the European fauna and flora. Accumulating evidence from studies of European oak gallwasps (Stone and Sunnucks, 1993; Atkinson, 2000, this study; Rokas, *et al.*, 2001a; Stone, *et al.*, 2001, Appendix I) as well as from their obligate oak hosts (Ferris, *et al.*, 1993; Dumolin-Lapegue, *et al.*, 1997; Ferris, *et al.*, 1998; Toumi and Lumaret, 1998) highlights the genetic isolation and distinctness of circum-Mediterranean southern refuge populations, including the eastern-most ones, their importance for the survival of the taxa in question, as well as the source populations for the subsequent invasion of central and northern Europe after the retreat of the ice sheet. To this end, the combination of a suite of available molecular markers is necessary for better description of the processes in action. Perhaps more importantly, knowledge of species-specific ecological constraints, such as availability of hosts for different stages of the lifecycle, may be used in generating testable null hypotheses (Stone, *et al.*, 2001, Appendix I, this study) and in understanding how different species react to common environmental changes.

5.7 Acknowledgements

A version of this chapter is being prepared for publication. Rokas, A., R. J. Atkinson, L. M. I. Webster and G. N. Stone. The phylogeographic structure of the European

oak gallwasp *Andricus quercustozae* reveals changes in genetic diversity with longitude and the importance of Anatolian populations.

Part of the mitochondrial haplotype data were collected by Lucy Webster as part of her Honours project in the University of Edinburgh. The allozyme data were collected and analysed by Rachel Atkinson. Rachel Atkinson and Graham Stone provided extensive comments on previous versions of this chapter. José-Luis Nieves-Aldrey, Alison Atkinson, Gyuri Csóka, Sara Lourie, Gil McVean, John Pannell, Alex Rowe and Yan Wong helped in sample collection and Rob Smallshire helped in Figure 5.5.

Appendix 5.1. Sampling effort (N) and allozyme allele frequencies at the 12 polymorphic loci for the 26 (13 in the first table and 13 in the second) populations of *Andricus quercustozae*.

First thirteen Populations.

Locus and allele number	**Azrou (Mor)**	**Madrid (Sp)**	**Prado del Rey (Sp)**	**Bordeaux (Fr)**	**Tarbes (Fr)**	**Bombiana (It)**	**Greve in Chianti (It)**	**Jelsi (It)**	**Massa Marittima (It)**	**Moio (It)**	**M. San' Angelo (It)**	**Poppi (It)**	**Rieti (It)**
N	39	9	17	12	9	16	39	28	27	28	40	39	22
GPD1													
1	0	0	0	0	0	0	0.039	0.071	0.074	0.036	0.113	0	0.046
2	1	1	1	1	1	1	0.961	0.911	0.926	0.964	0.862	1	0.931
3	0	0	0	0	0	0	0	0.018	0	0	0.025	0	0.023
GPD2													
1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0.013	0	0	0	0	0	0
3	1	1	1	1	1	1	0.987	1	1	0.982	1	1	1
4	0	0	0	0	0	0	0	0	0	0.018	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0
GOTs													
1	0	0	0	0	0	0	0	0	0	0.054	0	0	0
2	1	0.944	1	1	1	1	1	1	1	0.928	0.962	0.974	0.886

3	0	0.056	0	0	0	0	0	0	0	0.018	0.038	0.026	0.114
4	0	0	0	0	0	0	0	0	0	0	0	0	0
GOTm													
1	0	0	0	0	0	0	0	0	0	0	0.175	0	0
2	1	1	1	1	1	1	1	1	1	1	0.825	1	1
3	0	0	0	0	0	0	0	0	0	0	0	0	0
GPI													
1	0	0	0	0	0	0.125	0.166	0.125	0.130	0.196	0.175	0.167	0.205
2	0.923	1	0.971	1	1	0.844	0.821	0.875	0.870	0.786	0.812	0.833	0.795
3	0.077	0	0.029	0	0	0.031	0.013	0	0	0.018	0.013	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0
HK													
1	0	0	0	0	0	0.031	0	0	0	0	0	0.013	0
2	1	1	1	1	1	0.969	0.987	1	1	1	1	0.987	0.977
3	0	0	0	0	0	0	0.013	0	0	0	0	0	0.023
MDHm													
1	0.026	0	0	0	0	0	0	0.054	0	0	0	0	0
2	0.577	0.556	0.647	0.792	0.833	0.375	0.192	0.357	0.296	0.268	0.288	0.346	0.318
3	0.397	0.444	0.353	0.208	0.167	0.625	0.808	0.589	0.704	0.732	0.712	0.654	0.682

4	0	0	0	0	0	0	0	0	0	0	0	0	0
ME													
1	1	1	1	1	1	1	0.974	1	1	1	1	1	1
2	0	0	0	0	0	0	0.026	0	0	0	0	0	0
PEPb													
1	0	0	0	0	0	0	0	0.036	0	0	0	0	0
2	0	0	0	0	0	0.344	0.397	0.303	0.389	0.178	0.325	0.346	0.318
3	1	1	0.971	1	1	0.656	0.590	0.661	0.611	0.804	0.675	0.641	0.682
4	0	0	0.029	0	0	0	0.013	0	0	0.018	0	0.013	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0
PGM													
1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0.077	0.056	0.059	0	0	0.031	0	0.036	0	0	0	0.013	0
3	0.923	0.944	0.912	1	1	0.906	1	0.946	1	1	1	0.961	1
4	0	0	0.029	0	0	0.063	0	0.018	0	0	0	0.026	0
AK													

1	0	0	0	0	0	0	0	0	0.019	0	0	0	0.023
2	0.026	0	0.029	0	0	0	0.026	0	0	0	0	0.013	0
3	0.974	1	0.942	1	1	1	0.974	1	0.981	1	1	0.987	0.977
4	0	0	0.029	0	0	0	0	0	0	0	0	0	0
6PGD													
1	0	0	0	0	0	0	0.013	0.036	0	0.018	0	0	0.046
2	1	1	1	1	1	1	0.987	0.964	1	0.928	1	1	0.954
3	0	0	0	0	0	0	0	0	0	0.054	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0

Last thirteen populations.

Locus and allele number	**San Venanzo (It)**	**Sopron (Hu)**	**Szeghalom (Hu)**	**Tatabanya (Hu)**	**Veszprem (Hu)**	**Arnissa (Gr)**	**Ağlasun (Tur)**	**Gezende (Tur)**	**Küllüce (Tur)**	**Lysandra (Tur)**	**Madenli (Tur)**	**Refahiye (Tur)**	**Yeniyol (Tur)**
N	27	40	39	10	40	8	15	40	7	6	40	7	5
GPD1													
1	0.093	0	0	0	0.013	0	0	0.013	0	0	0	0	0
2	0.907	1	1	1	0.987	1	1	0.974	1	1	1	1	1
3	0	0	0	0	0	0	0	0.013	0	0	0	0	0
GPD2													

1	0	0	0	0	0	0	0	0	0	0	0.013	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	0.975	0.987	1	0.987	1	1	1	1	1	0.987	1	1
4	0	0.025	0.013	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0.013	0	0	0	0	0	0	0	0
GOTs													
1	0	0.025	0.077	0	0.075	0.063	0.2	0.038	0	0.167	0.225	0	0.2
2	1	0.975	0.923	1	0.925	0.937	0.4	0.512	1	0.5	0.388	0.857	0.8
3	0	0	0	0	0	0	0.333	0.425	0	0.333	0.387	0.143	0
4	0	0	0	0	0	0	0.067	0.025	0	0	0	0	0
GOTm													
1	0	0	0	0	0	0.875	0	0.013	0.929	0	0.05	1	1
2	1	1	1	1	1	0.125	1	0.987	0	1	0.95	0	0
3	0	0	0	0	0	0	0	0	0.071	0	0	0	0
GPI													
1	0.167	0	0.013	0	0.012	0.063	0	0.037	0.071	0	0	0.071	0
2	0.833	1	0.936	0.95	0.975	0.812	1	0.9	0.929	1	0.963	0.929	1
3	0	0	0.051	0.05	0.013	0.062	0	0.05	0	0	0.037	0	0
4	0	0	0	0	0	0.063	0	0.013	0	0	0	0	0

HK													
1	0	0	0.013	0	0	0	0	0.013	0	0	0.013	0	0
2	1	0.975	0.987	1	0.975	1	0.967	0.937	1	0.917	0.912	1	1
3	0	0.025	0	0	0.025	0	0.033	0.05	0	0.083	0.075	0	0
MDHm													
1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0.315	0.987	0.974	1	0.912	0.875	0.433	0.5	0.429	0.833	0.325	0.429	0.6
3	0.685	0.013	0.026	0	0.088	0.125	0.567	0.5	0.571	0.167	0.662	0.571	0.4
4	0	0	0	0	0	0	0	0	0	0	0.013	0	0
ME													
1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	0	0	0	0	0	0	0	0	0	0	0	0	0
PEPb													
1	0	0	0	0	0	0	0	0.1	0.071	0.083	0.05	0	0.1
2	0.333	0.1	0.128	0.1	0.113	0.312	0.233	0.187	0.214	0.167	0.188	0.357	0.6
3	0.667	0.825	0.872	0.9	0.887	0.625	0.6	0.35	0.5	0.5	0.412	0.357	0.3
4	0	0.075	0	0	0	0.063	0.167	0.275	0.071	0.25	0.287	0	0
5	0	0	0	0	0	0	0	0.088	0.071	0	0.05	0.072	0
6	0	0	0	0	0	0	0	0	0	0	0.013	0	0

7	0	0	0	0	0	0	0	0	0	0	0	0.214	0
8	0	0	0	0	0	0	0	0	0.071	0	0	0	0
PGM													
1	0	0	0	0	0	0	0	0.013	0	0.25	0.05	0	0
2	0	0	0	0	0.013	0	0.167	0.125	0.143	0.167	0.125	0.357	0.5
3	0.963	1	1	1	0.987	1	0.833	0.837	0.643	0.583	0.775	0.643	0.5
4	0.037	0	0	0	0	0	0	0.025	0.214	0	0.05	0	0
AK													
1	0	0.013	0	0	0	0	0.533	0.875	0.286	0.583	0.688	0.286	0.3
2	0	0.125	0	0	0	0	0.467	0.125	0	0.417	0.288	0.429	0.1
3	1	0.862	1	1	1	1	0	0	0	0	0.025	0.143	0.6
4	0	0	0	0	0	0	0	0	0.714	0	0	0.143	0
6PGD													
1	0	0	0	0.05	0	0	0	0.025	0.214	0	0.088	0	0.1
2	1	1	1	0.95	1	1	0.8	0.812	0.572	0.917	0.625	0.786	0.9
3	0	0	0	0	0	0	0.2	0.163	0	0.083	0.287	0.214	0
4	0	0	0	0	0	0	0	0	0.214	0	0	0	0

SECTION III: APPLICATIONS OF 'TREE-THINKING' AT THE INTER-SPECIFIC LEVEL

CHAPTER 6

A mitochondrial molecular phylogeny of the European oak gallwasps (Hymenoptera: Cynipidae: Cynipini)

6.1 Abstract

Oak gallwasps are distributed throughout the Holarctic and have the ability to induce a wide diversity of highly complex species- and generation-specific galls on oaks and other Fagaceae. We have sequenced a fragment of the mitochondrial cytochrome *b* locus from 57 European and 5 non-European oak gallwasp species from 10 genera and analysed the resulting data in a statistical framework (maximum likelihood and Bayesian phylogenetic inference) in an attempt to reconstruct the phylogenetic associations within European oak gallwasps. Additionally, for 16 of the 57 European species, we have included 23 additional cytochrome *b* haplotype sequences from distinct phylogeographic European refuges, in order to understand the relationship between intra- and inter-specific divergence in oak gallwasps. The reconstructed phylogenies show good intra-generic resolution and identify several conserved clades, but fail to reconstruct either very recent or very ancient divergences. Addition of the intra-specific haplotypes significantly improves the resolution of topologies generated by both methods and reveals cases of lineage sorting of ancestral polymorphisms as well as introgressive hybridisation, especially in species of the genus *Andricus*. Crucially, the effective population size of mitochondrial loci is much smaller than that of nuclear autosomal loci (due to their different mode of inheritance) and as a consequence, mitochondrial intra-specific polymorphisms have much smaller coalescence times than nuclear markers. This suggests that the lineage sorting observed in European oak gallwasps for the cytochrome *b* locus will be more

pronounced in nuclear loci. More generally, phylogenies of widely-distributed insect groups, like the oak gallwasps, may be sensitive to all factors that produce incongruence between a gene phylogeny and the species phylogeny. These include (a) variation in the inheritance pattern of the molecular marker(s) used, and (b) impact of the geographical origin of the specimens from each species.

6.2 Introduction

Gallwasps (Hymenoptera: Cynipidae) are members of the Cynipoidea, a major lineage of predominantly parasitoid wasps within the Hymenoptera (Ronquist, 1995, 1999). All gallwasps are obligate parasites of plants, and either induce their own galls in plant tissues or develop as inquilines within the galls induced by other gallwasps. Each gallwasp species induces a highly characteristic and often highly complex gall, leading to an incredible diversity of gall structures within the whole group. The gallwasp family is composed of about 1360 described species, currently divided, on the basis of morphological data, into six tribes (Fig. 6.1) (Liljeblad and Ronquist, 1998). Biogeographic and fossil evidence suggest that the family of gallwasps originated at least as long ago as the mid Cretaceous (83 mya) (Ronquist, 1999). Fossil data also suggest that the inquiline Synergini diverged at least 45 mya (Ronquist, 1999). This dating also suggests that the woody rosid gallers (Fig. 6.1), as obligate hosts to the inquilines, must also have diverged at least 45 mya.

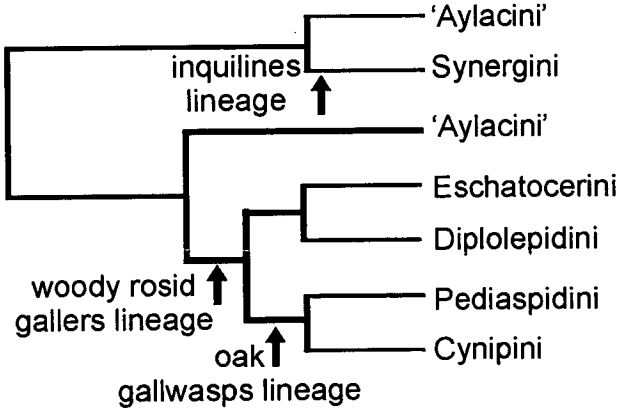


Figure 6.1. Morphology-based phylogeny of the six tribes in the family of gallwasps. The topology was generated using parsimony (Liljeblad and Ronquist, 1998). Arrows indicate lineages discussed in the text.

Oak gallwasps (tribe Cynipini) are the most species-rich tribe of gallwasps, with around 1000 known species world-wide, predominantly in the Northern Hemisphere (Ronquist, 1999) and have a lifecycle consisting of an alternation between a spring sexual generation and a summer/autumn parthenogenetic generation (cyclical parthenogenesis) (for a review of oak gallwasps see Stone, *et al.*, 2002). Geographic variation in cynipid diversity generally tracks patterns in oak species richness (Cornell and Washburn, 1979). The greatest richness of oak gallwasps is found in the Nearctic (and particularly in Mexico), where there are an estimated 700 species in 29 genera (Weld, 1957, 1959, 1960). Oak gallwasps are thought to have undergone major radiations in central America (Kinsey, 1936) alongside their oak hosts (Manos, *et al.*, 1999), suggesting that oak gallwasps may have originated in the Nearctic. The Palaearctic fauna is less species-rich (*ca.* 140 species in 11 genera) (Askew, 1984; Nieves-Aldrey, 1987; Stone, *et al.*, 2002). The majority of European oak gallwasps (94 species) are contained within the genus *Andricus*, with the genera *Neuroterus* (12) and *Cynips* (9) following in species-richness (Table 6.1) (Askew, 1984; Stone, *et al.*, 2002).

Table 6.1. World-wide and European species richness for the genera analyzed in this study (from Askew, 1984; Stone, *et al.*, 2002).

Genus	World-wide	European	Analyzed in this study
Andricus	1000	94	42 (+ 4 non-European)
Aphelonyx	1	1	1
Biorhiza	2	2	1
Callirhytis	150	7	1
Chilaspis	1	1	1
Cynips	?	9	6
Neuroterus	100	12	3
Plagiotrochus	5	5	1
Trigonaspis	4	4	1
Trichagalma	2	—	1 (non-European)

Although an increasing body of work has been addressing phylogenetic questions on gallwasps (Ronquist, 1994; Liljeblad and Ronquist, 1998; Stone and Cook, 1998; Cook, *et al.*, 2001, Chapter 3; Rokas, *et al.*, 2001b), still relatively little is known about cynipid phylogenetic relationships, especially below the tribe level, although the taxonomic status of most European species has been well-studied and repeatedly revised in recent years (Ambrus, 1974; Melika, *et al.*, 1999; Melika, *et al.*, 2000). This in contrast to the taxonomy of Nearctic species, which is much more problematic (Drown and Brown, 1998), despite significant recent efforts towards a more coherent classification (Lyon, 1984, 1993, 1996; Melika and Abrahamson, 1997b, a, 2000).

Mitochondrial DNA loci have been extensively used in phylogenetic studies of taxa in low taxonomic levels, because of their rapid evolution and easiness of amplification (e.g. Hillis, *et al.*, 1996). In studies where multiple gene trees are employed to infer the species tree, the utility of multiple mtDNA loci has been questioned on the grounds that mitochondrion is essentially a single locus; mtDNA loci are inherited as a single linkage group and thus do not provide independent estimates of the species tree (Pamilo and Nei, 1988; Wu, 1991). Therefore in multi-locus studies, the species tree will be best inferred from multiple nuclear gene trees, since they are independently inherited. However, the use of a mtDNA locus (than a nuclear one) may be favored in single-locus studies aimed at inferring the correct species tree. Specifically, the probability that a gene tree will be congruent with a species tree depends on the effective population size of the ancestral species between the two successive speciation events; the smaller the effective population size, the higher the probability that a given gene tree will track the species tree (Moore, 1995). Although the inter-node distance will be the same for both mitochondrial and nuclear loci, the effective population size for mtDNA loci is one-fourth of a nuclear autosomal gene (Moore, 1995; see Hoelzer, 1997, for exceptions; Palumbi, *et al.*, 2001). This is due to the different transmission genetics of the two types of loci; mitochondrial loci are both haploid and transmitted only through females, whereas nuclear autosomal genes are carried in two copies by each of the two sexes. Therefore, a mitochondrial gene tree has a better chance of tracking the species tree than any single nuclear gene tree (Moore, 1995; Palumbi, *et al.*, 2001). This

advantage of mitochondrial loci becomes even more important for taxonomic groups that have experienced high speciation rates, a situation found in many insect groups (e.g. Kambysellis, *et al.*, 1995; Farrell, 1998).

In an earlier paper (Rokas, *et al.*, 2001b, Chapter 3) we assessed the utility of eight popular DNA sequence markers in reconstructing phylogenetic relationships at various levels of divergence within the family of gallwasps. Mitochondrial loci were shown to be fast-evolving and potentially useful for lower-level phylogenetics of gallwasp taxa. In this study, we have sequenced a fragment of the mitochondrial gene cytochrome *b* for 57 species from 9 genera of European oak gallwasps as well as from 5 non-European species belonging in 2 genera in an attempt to (a) reconstruct the phylogenetic relationships within the tribe Cynipini. This sampling includes more than a third of the total number of species known to exist in Europe (Table 6.1). Additionally, we have included 23 cytochrome *b* haplotypes from distinct Pleistocene glacial refuges for 16 of the 57 European species, resulting in an extended dataset of 85 taxa, in an attempt to (b) understand the effect of intra-specific variation on inter-specific patterns. Additionally, (c) we comment on the utility of the cytochrome *b* marker for insect lower-level phylogenetics.

6.3 Materials and Methods

6.3.1 Study species

Two datasets were analysed in this study. Dataset I is composed of 62 oak gallwasp species (Appendix 6.1). Five species in this dataset are non-European (*Andricus mukaigawae*, *Andricus kashiwaphilus*, *Andricus symbioticus* and *Trichagalma serratae* from Japan, eastern Palaearctic and *Andricus spectabilis* from the USA, Nearctic). Dataset II is an extension of dataset I; 23 individuals from 16 species (10 *Andricus* species, 4 *Cynips* species, 1 *Neuroterus* species and 1 *Biorhiza* species – all of them already included in dataset I) have been added. These additional 23 taxa are from distinct Pleistocene ice refuges in Europe and Anatolia (Appendix 6.1) and were collected as part of a larger project on the phylogeography of European gallwasps (Atkinson, 2000; Rokas, *et al.*, 2001a; Stone, *et al.*, 2001, Chapter 4,

Appendix I, Rokas, *et al.*, unpublished data). Species nomenclature follows the revisions of Melika *et al.* (2000). No outgroup was used since the other gallwasp tribes are, according to the molecular marker employed in this study, distantly related to Cynipini (Rokas, *et al.*, 2001b, Chapter 3).

6.3.2 Molecular methods

DNA was extracted by using the DNeasy Tissue kit (QIAGEN cat. 69504) and following the protocol of the manufacturer for insect DNA extraction. A 433 bp fragment of the mitochondrial cytochrome *b* gene was amplified by PCR using the CB1/CB2 primer combination of Jermiin and Crozier (1994; alias CB-J-10933 and CB-N-11367 of Simon, *et al.*, 1994) and previously described conditions (Stone and Cook, 1998; Rokas, *et al.*, 2001a, Chapter 4, Appendix I; Stone, *et al.*, 2001). Both strands were sequenced to minimise PCR artefacts, ambiguities and base-calling errors. Sequencing was carried out using Perkin-Elmer BigDye Terminator chemistry on an ABI 377 sequencer. All sequences will be available from Genbank on publication.

Following an initial demonstration that multiple sequences were produced by PCR amplification of cytochrome *b* from four gallwasp species (*Andricus aries*, *Andricus theophrastii*, *Andricus tomentosus* and *Neuroterus lanuginosus*), a strategy of cloning was adopted. PCR products were cloned in a blunt-end vector following the manufacturer's instructions (Zero Blunt® TOPO® Cloning kit, Invitrogen, cat. K4500-01). For each species, up to ten colonies were selected, cultured and purified (QIAprep spin miniprep kit, QIAGEN cat. 27104) for subsequent cloning. All sequences will be available from Genbank on publication.

6.3.3 Phylogenetic analysis

6.3.3.1 Maximum likelihood - ML

Phylogenies were estimated by maximum likelihood (ML) analysis of the sequence data, using the PAUP* package, version 4.0b8 (Swofford, 2000). The best-fit ML model for each locus was identified using MODELTEST 3.0 (Posada and Crandall,

1998). Parameters allowed to vary in model-fitting were base composition, substitution rates and rate heterogeneity among sites. The ML model of sequence evolution was identified using likelihood ratio tests (Huelsenbeck and Rannala, 1997). The parameter values suggested by MODELTEST were used to perform heuristic searches, using the nearest-neighbour-interchange (NNI) algorithm on 500 bootstrap replicates.

6.3.3.2 Bayesian inference - $(MC)^3$

Phylogenies were estimated using a variant of Markov chain Monte Carlo (MCMC) (e.g. Larget and Simon, 1999) using the program MRBAYES, version 2.0 (Huelsenbeck and Ronquist, 2001). MRBAYES implements a variant of MCMC called Metropolis-coupled Markov chain Monte Carlo – $(MC)^3$ for short. The $(MC)^3$ algorithm allows running of n chains, n-1 of which may be 'heated'. A 'heated' chain effectively explores more easily tree space, by taking leaps across deep valleys in the landscape of trees. Importantly, swaps between the 'heated' and the 'non-heated' (cold) chains are allowed, leading to a much better exploration of tree space. For each of the two datasets (I and II), four independent runs, each with four chains (three 'cold' and one 'heated'), were generated. Two of the runs were performed for 2,000,000 generations and the other two for 1,000,000 generations. Each analysis was run under a model allowing GTR+G+I model of sequence evolution with all parameters (base frequencies, substitution matrix values, rate heterogeneity by gamma distribution and invariable sites) estimated.

The alignments for all datasets and topologies reported in this study will be available electronically from TreeBASE on publication.

6.4 RESULTS

6.4.1 Cytochrome b *characteristics*

Genetic distances among basal species as well as among oak gallwasp genera are very high (Table 6.2, upper diagonal). Between the different oak gallwasp genera the proportion of transitions against the overall number of substitutions is around 40%

(Table 6.2, lower diagonal). Comparisons between less divergent taxa typically show a higher proportion of transitions; for example, within the genus *Cynips* values range between 52% – 92% with a mean of 69%. This pattern of the proportion of transitions comprising approximately 40% of the observed substitutions in sequences that are very divergent seems to be common in insect phylogenetic studies using mitochondrial markers (Liu and Beckenback, 1992; Page, *et al.*, 1998), and is probably indicative of substitutional saturation.

The cytochrome *b* region is very AT-rich (observed frequencies, A: 34.8%, C: 12.3%, G: 10.8%, T: 42.1%), following the general rule of high AT-bias in insect mitochondrial genomes, especially in the third positions of codons (Table 6.3). Such extreme AT-bias results in many sites essentially being transformed from four-state (A, C, G or T) to two-state characters (A or T). As a consequence, such sites become saturated faster and show higher levels of homoplasy. Demonstration of this effect is shown in Figure 6.2 where the number of nucleotide changes per nucleotide site on the ML-estimated consensus tree has been plotted for 1st and 2nd, and 3rd codon positions only. In the highly AT-biased 3rd positions, some nucleotide sites are estimated to have undergone 15 or more changes of character states, in contrast to 1st and 2nd position sites, which have undergone significantly fewer changes.

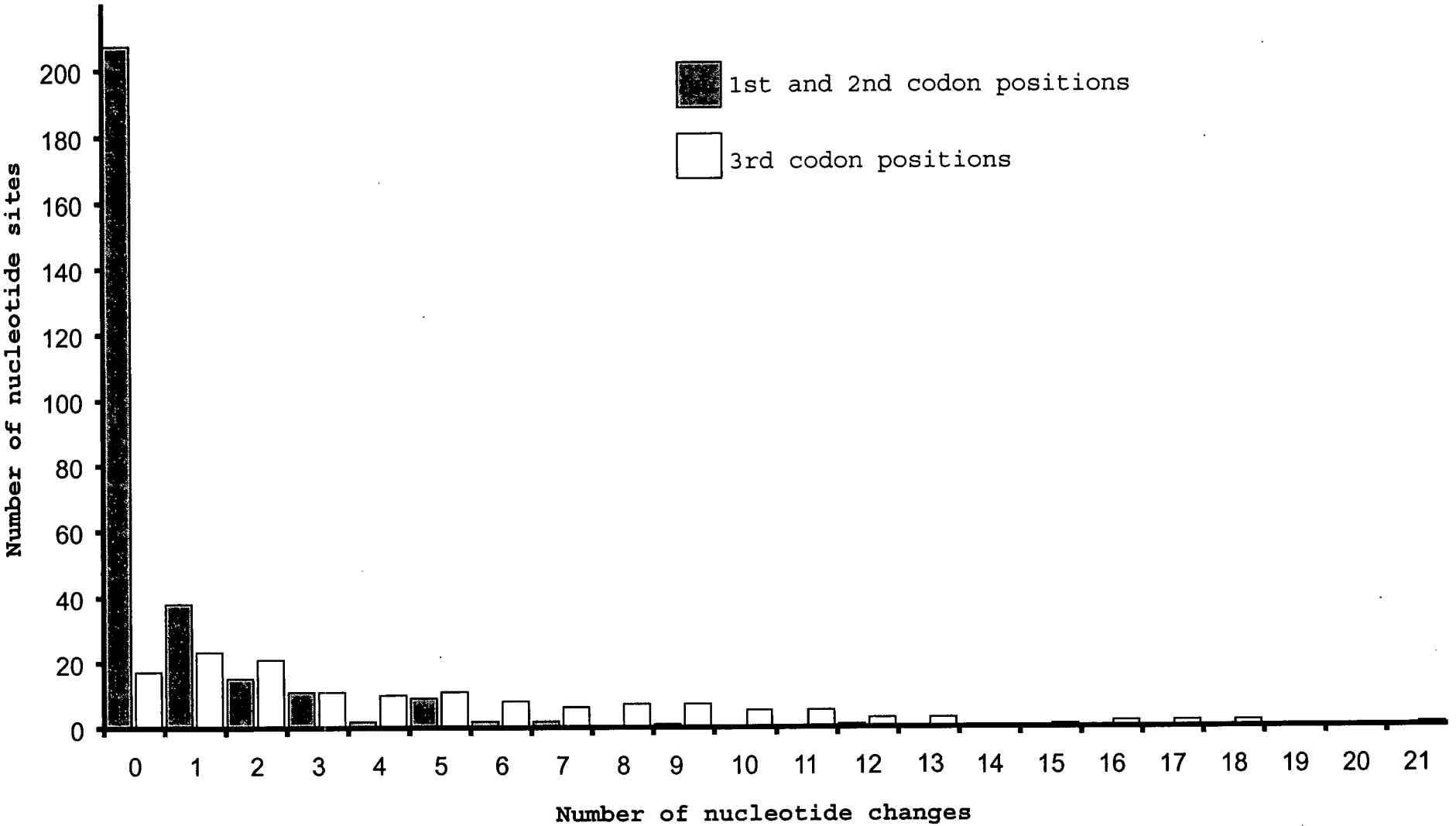


Figure 6.2. Distribution of the number of changes per nucleotide site for the cytochrome *b* fragment on the ML-estimated consensus tree.

Genus	*Andricus*	*Aphelonyx*	*Biorhiza*	*Callirhytis*	*Chilaspis*	*Cynips*	*Neuroterus*	*Plagiotrochus*	*Trichagalma*	*Trigonaspis*
Andricus[1]	—	17.3 / 54.9	11.1 / 25.3	11.5 / 27.2	12.7 / 33.2	12.2 / 29.9	12.9 / 34.0	13.6 / 33.6	15.4 / 45.2	14.1 / 35.2
Aphelonyx	38.7	—	15.7 / 47.3	15.9 / 50.4	13.6 / 36.4	16.4 / 48.0	12.7 / 30.0	15.2 / 44.3	15.7 / 43.1	17.3 / 56.2
Biorhiza	35.4	36.8	—	8.5 / 16.2	12.7 / 35.9	9.2 / 16.7	12.0 / 29.4	14.3 / 40.6	13.6 / 35.6	10.4 / 21.6
Callirhytis	32.0	33.3	32.4	—	9.9 / 21.8	8.8 / 14.8	9.5 / 20.0	11.1 / 24.5	13.9 / 37.9	12.7 / 29.2
Chilaspis	30.1	32.2	25.5	25.6	—	11.3 / 25.4	9.0 / 17.3	11.1 / 23.5	12.7 / 29.2	14.1 / 39.4
Cynips[2]	35.8	40.8	42.5	47.4	34.7	—	11.1 / 22.7	12.9 / 29.5	14.2 / 37.6	9.9 / 18.3
Neuroterus[3]	32.1	38.2	32.7	26.8	33.3	41.7	—	11.3 / 24.4	11.8 / 24.6	13.2 / 29.5
Plagiotrochus	39.0	33.3	33.9	33.3	37.5	42.9	38.8	—	12.7 / 28.3	13.9 / 31.6
Trichagalma	37.3	41.2	37.3	33.3	41.8	38.7	45.1	45.5	—	14.1 / 38.2
Trigonaspis	44.3	38.7	35.6	41.8	34.4	48.8	47.4	50.0	37.7	—

Table 6.2. Upper diagonal: percent sequence divergence (uncorrected *p* distance / ML-estimated distance) for cytochrome *b* between representative species of the European oak gallwasp genera (with the exception of *Trichagalma*). Lower diagonal: Proportion of transitional substitutions against all substitution types in the 433 bp cytochrome *b* fragment for the 10 oak gallwasp genera. Representative species for each genus: [1]: *Andricus hungaricus*; [2]: *Cynips quercus*; [3]: *Neuroterus saliens*.

Insect genus (Order: Family)	Max. divergence (%)				AT-content (%)				Study
	all	1st	2nd	3rd	all	1st	2nd	3rd	
Andricus gallwasps (Hymenoptera: Cynipidae)	16.9	13.2	7.6	33.3	76.8	69.4	67.1	93.8	This study
Cynips gallwasps (Hymenoptera: Cynipidae)	6.0	3.5	0	16.7	76.7	70.2	67.3	92.7	This study
Euura sawflies (Hymenoptera: Tenthredinidae)	10.8	9.7	5.8	20.3	74.8	69.4	69.4	84.5	(Nyman, *et al.*, 2000)
Phyllocolpa sawflies (Hymenoptera: Tenthredinidae)	12.5	10.6	4.0	27.1	76.0	70.1	69.9	90.5	(Nyman, *et al.*, 2000)
Pontania sawflies (Hymenoptera: Tenthredinidae)	15.2	11.5	4.9	30.2	75.6	70.2	69.4	86.5	(Nyman, *et al.*, 2000)
Dennyus lice (Phthiraptera: Menoponidae)	37.4	31.5	17.5	65.1	68.9	62.2	61.7	82.8	(Page, *et al.*, 1998)
Ischnura damselflies (Odonata: Coenagrionidae)	16.2	7.6	3.6	38.0	71.0	62.2	64.7	86.2	(Chippindale, et al., 1999)
Hepialus butterflies (Lepidoptera: Hepialidae)	9.8	5.2	2.8	23.1	75.4	69.7	66.6	90.1	(Chen - Genbank)
Nebria ground beetles (Coleoptera: Carabidae)	10.6	7.6	1.4	22.9	73.5	62.4	64.6	93.4	(Clarke, 1998 - Genbank)
Phlebotomus sandflies (Diptera: Psychodidae)	24.9	18.8	5.0	51.9	72.9	62.1	66.8	90.0	(Esseghir, *et al.*, - Genbank)
Triatoma bugs (Hemiptera: Reduviidae)	24.8	17.2	5.3	55.3	64.1	55.4	66.5	71.0	(Lyman, *et al.*, 1999)

Table 6.3. Compilation of maximum sequence divergence (uncorrected *p* distance) and AT-content for all cytochrome *b* codon positions from a variety of insect genera.

6.4.1.1 Multiple cytochrome b-like fragments in four gallwasp species

Sequencing of multiple clones for cytochrome *b* from the four gallwasp species (*A. aries*, *A. theophrastii*, *A. tomentosus* and *N. lanuginosus*) that failed direct sequencing, revealed multiple cytochrome *b*-like fragments. Out of the seven *A. aries* clones, one had the correct length (433 bp) and reading frame (no stop codons / indels), whereas the other six clones had a 1 bp indel (434 bp) and differed by 3.7% (uncorrected *p* distance) to the first clone. In *A. theophrastii*, five clones had the correct length and reading frame, whereas the other five although of correct length had a stop codon. The two clone groups differed by 10.2%. In *A. tomentosus*, five distinct sequences were found. Only one clone had the correct length and reading frame. Five clones had the correct length but translation revealed the presence of a stop codon. One clone was 2 bp longer (435 bp), whereas the last two clones were shorter (375 and 376 bp respectively). The lowest divergence observed between these five sequences was 2.6%. In *N. lanuginosus*, all four clones recovered had the correct length and reading frame. For the phylogenetic analyses presented in this study, only the haplotype with the correct length and reading frame from each species was used.

6.4.2 Phylogenetic patterns of European oak gallwasps

Phylogenetic reconstructions of dataset I from the two different methods of statistical phylogenetic inference, maximum likelihood and Bayesian inference, and their respective bootstrap / posterior probability branch support values, are shown in Figures 6.3 and 6.4. Several conclusions may be drawn regarding the largest genus of European Cynipini, *Andricus*. The majority of *Andricus* species form a monophyletic clade, although with low bootstrap / posterior probability support (< 50%). Three species (*A. inflator*, *A. spectabilis*, *A. symbioticus*) are placed outside the monophyletic *Andricus* clade, but their groupings are also not robustly supported. Additionally, all the non-European *Andricus* species (*A. symbioticus*, *A. mukaigawae* and *A. symbioticus* from Japan – eastern Palaearctic; *A. spectabilis* from the U.S. – Nearctic) are located basally on the tree. Within *Andricus*, five main clades may be identified, although the topologies within each clade differ slightly between the two

methods; these are *A. mayri*, *A. kollari*, *A. hartigi*, *A. quercuscalicis* and *A. fecundator* (Figs. 6.3, 6.4). With the exception of the *A. fecundator* clade, the monophyly of all clades is well supported. However, within each clade, branch support values (bootstrap / posterior probability) are low. For example, only 1/8 and none out of 6 branches within the *A. quercuscalicis* and *A. fecundator* clades respectively, show bootstrap values higher than 50% in the ML analysis (for the $(MC)^3$ analysis, posterior probabilities above 50% are found in 2/8 and in 1/6 branches within the two clades –Figs. 6.3, 6.4).

Distances within *Andricus* are much higher (uncorrected *p* distance: 0.2 – 15.0% / ML-estimated distance: 0.2 – 39.8%) than distances within the genus *Cynips* (2.8 – 6% / 2.9 – 8.3%). Distances within *Cynips* have a much better match with distances within the various clades of *Andricus* (*A. mayri* clade: 0.7 – 5.3% / 0.7 – 6.4%; *A. kollari* clade: 0.2 – 6.5% / 0.2 – 8.2%; *A. quercuscalicis* clade: 0.7 – 5.8% / 0.7 – 7.8%; *A. hartigi* clade: 2.5 – 5.1% / 2.7 – 6.4%).

Outside *Andricus*, the phylogenetic patterns observed are not very robust. The genus *Cynips* is monophyletic with relatively good bootstrap support in the ML reconstructions (Figs. 6.3, 6.5), but weakly supported in the $(MC)^3$ analyses (Figs. 6.4, 6.6). Monophyly is not supported for the genus *Neuroterus* either. Two of the three representative species (*N. macropterus* and *N. saliens*) are grouped together, but the third species (*N. lanuginosus*) groups with *Trichagalma serratae*, although the whole *Neuroterus*-containing clade is weakly supported.

A maximum likelihood phylogeny of a reduced set taxa containing only a few representatives from each oak gallwasp genus was also generated. The resulting phylogeny (data not shown) shows the same patterns as the phylogenies generated from the two larger datasets – low resolution of the affinities between the various genera.

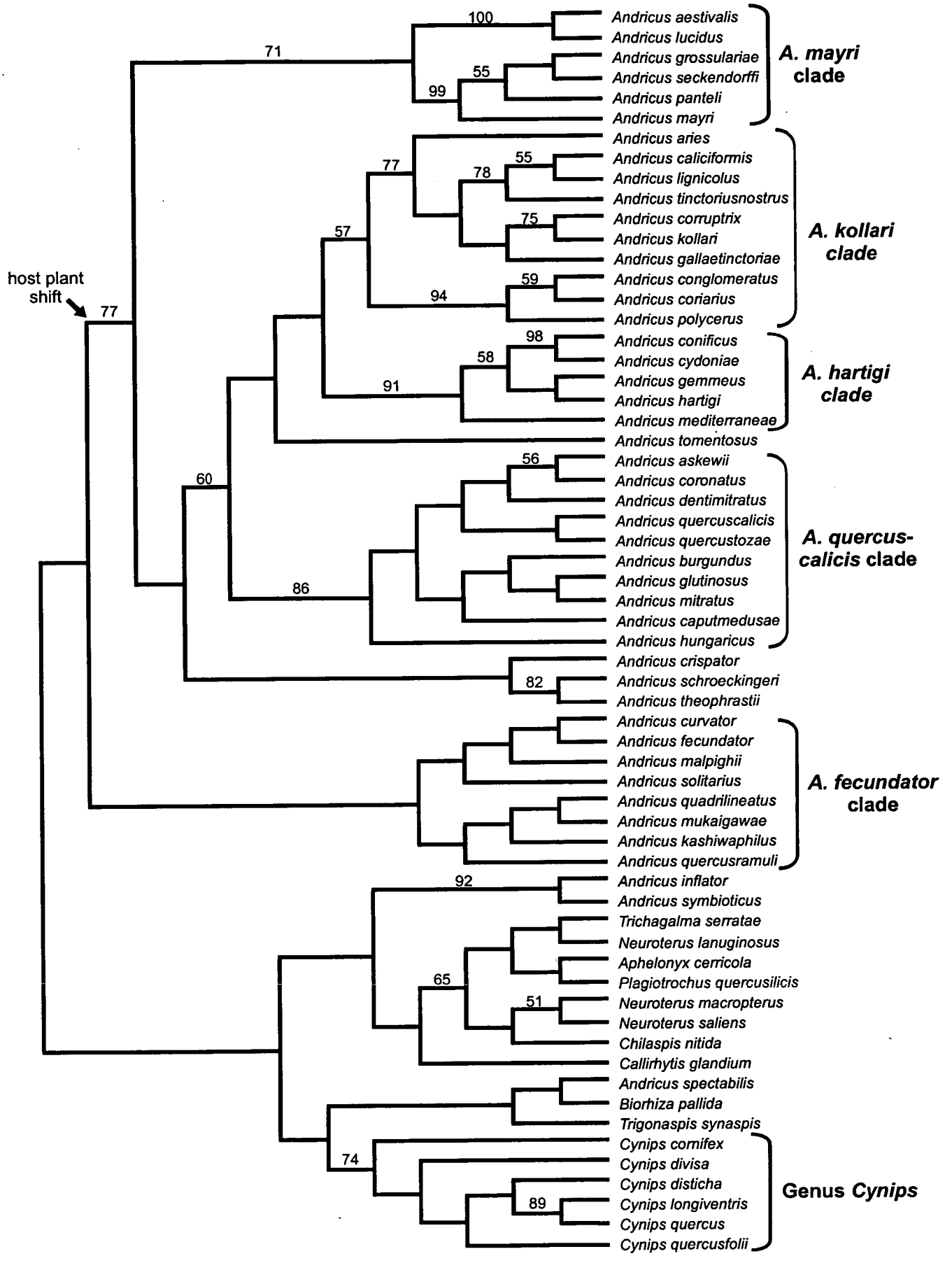


Figure 6.3. 50% majority-rule consensus cladogram from ML analysis of cytochrome *b* from 62 species of oak gallwasps. The root of the tree is arbitrary and has been chosen for best visualisation. Branches without bootstrap support values are supported by less than 50% of replicates.

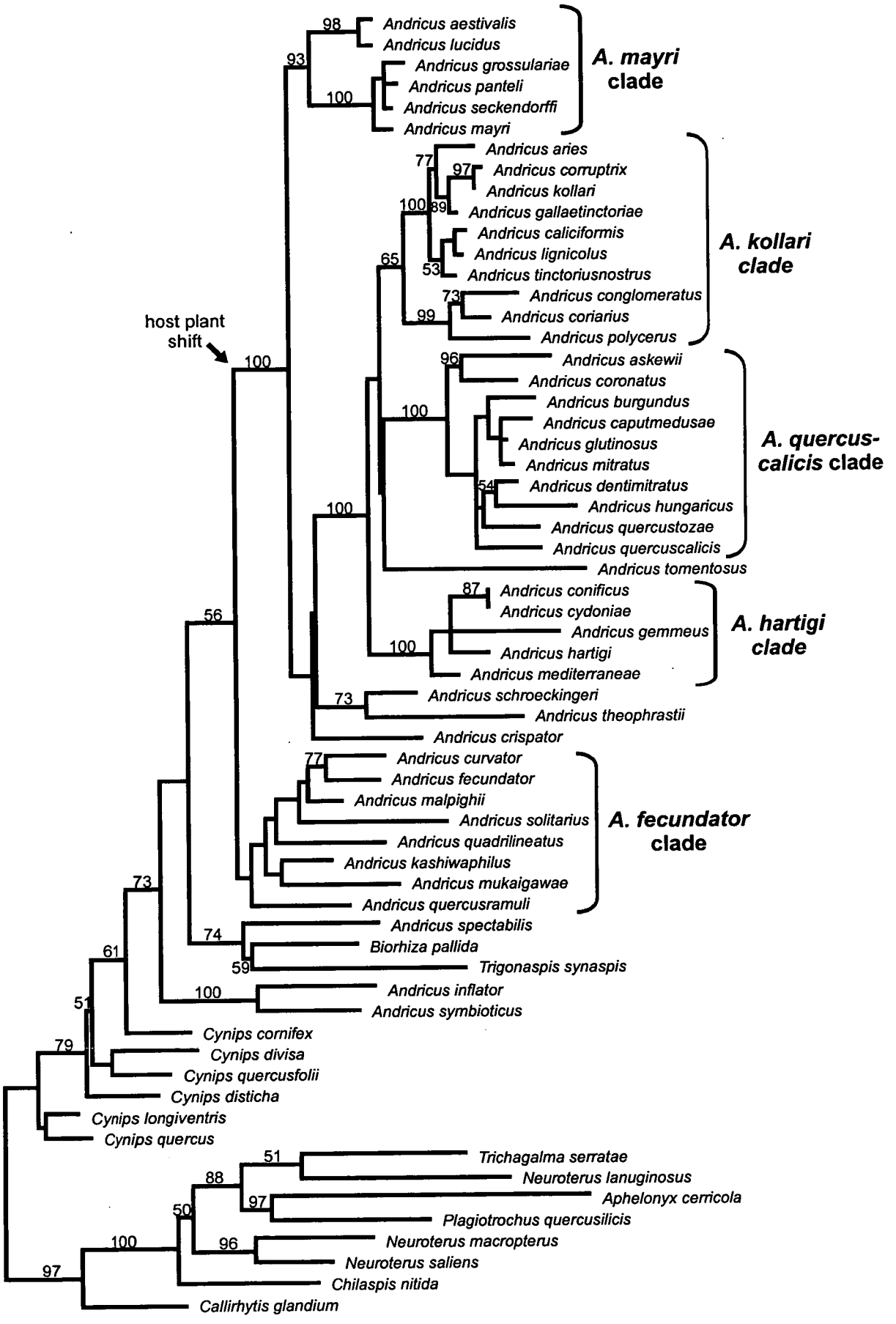


Figure 6.4. 50% majority-rule consensus phylogram from $(MC)^3$ analysis of cytochrome *b* from 62 species of oak gallwasps. The tree is mid-point rooted. Branches without posterior probability values are supported by less than 50% of the sampled trees.

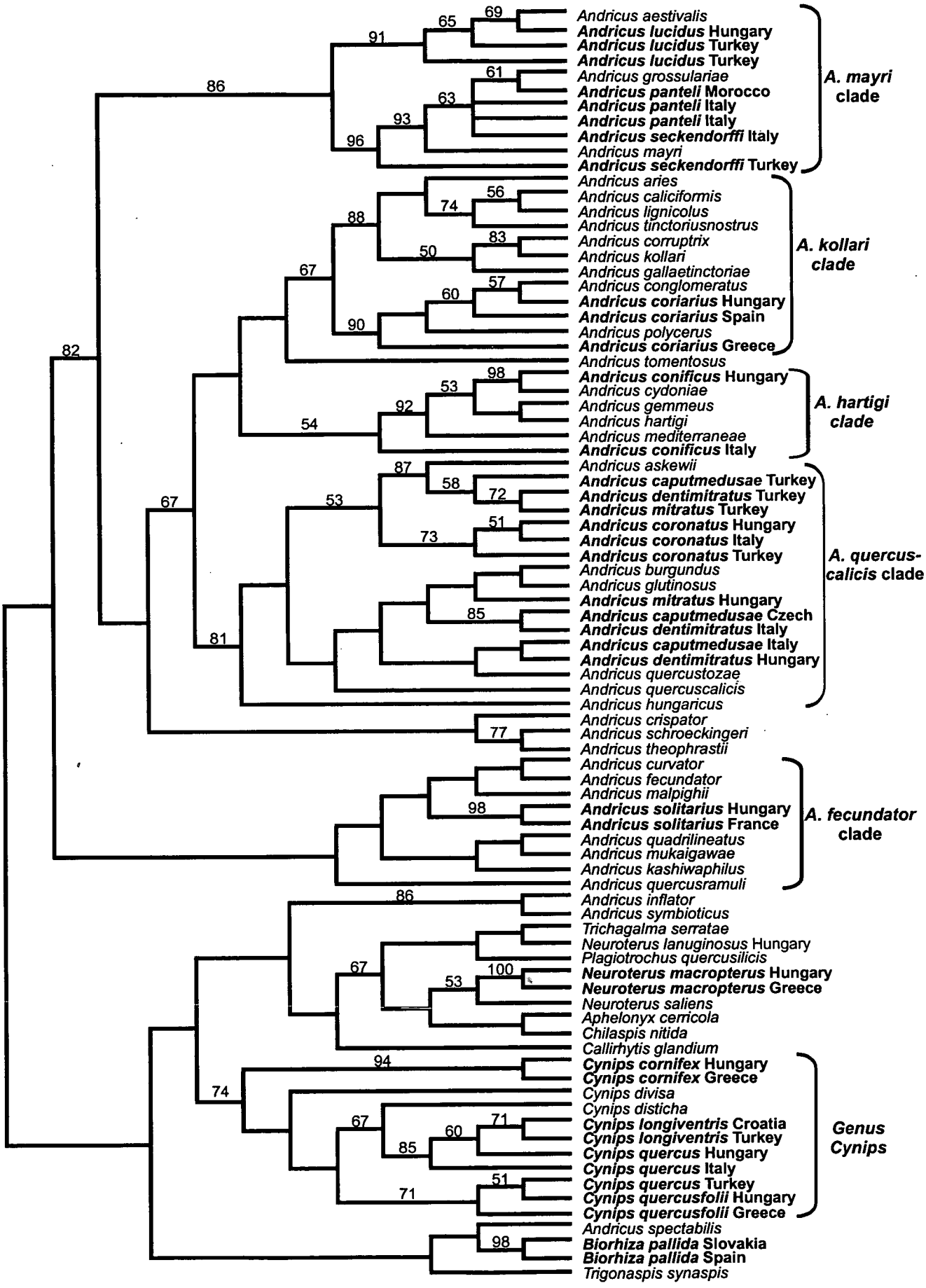
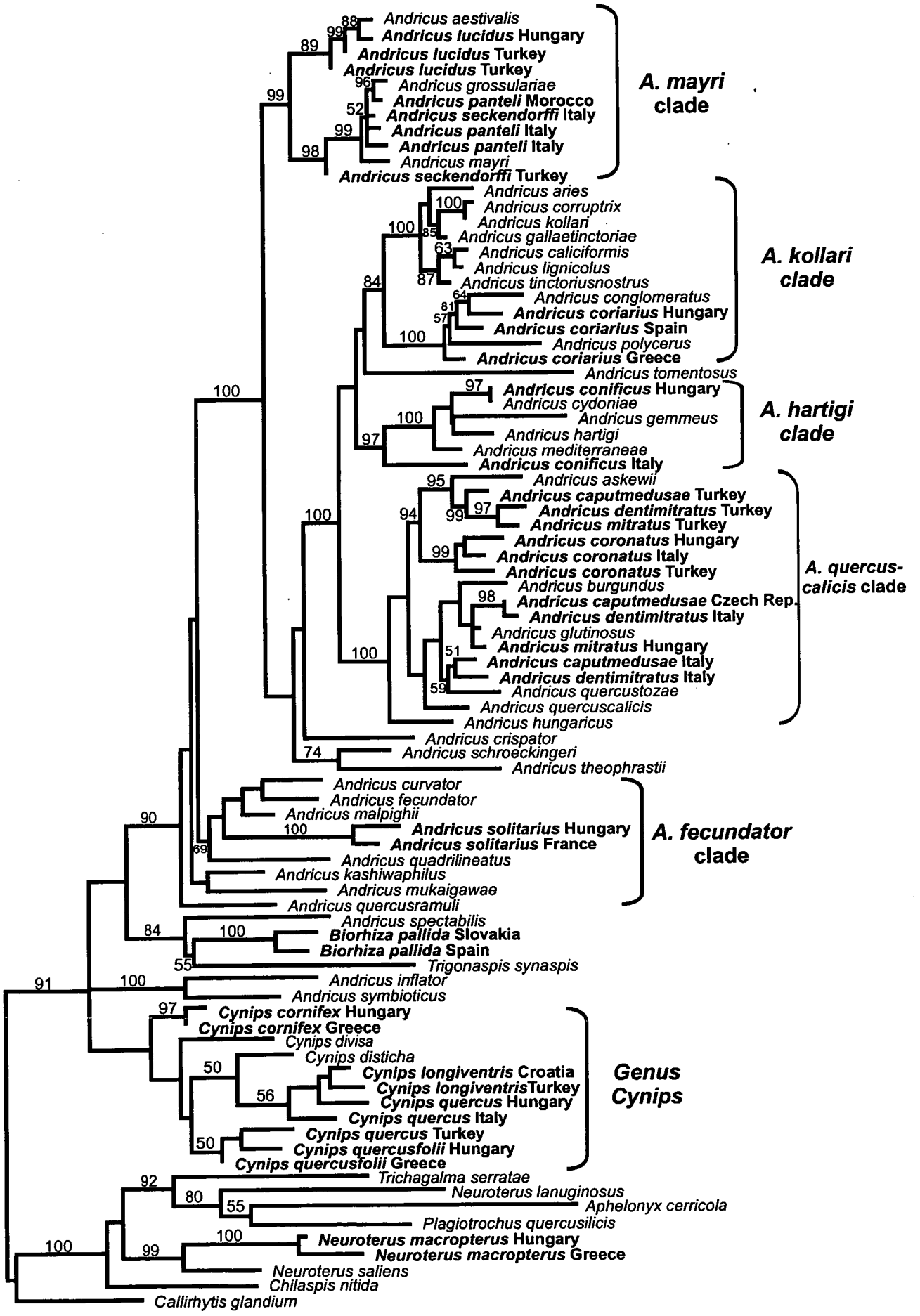


Figure 6.5. 50% majority-rule consensus cladogram from ML analysis of cytochrome *b* for the 85-taxa dataset. The root of the tree is arbitrary and has been chosen for best visualisation. Branches without bootstrap support values are supported by less than 50% of replicates.



gure 6.6. 50% majority-rule consensus phylogram from $(MC)^3$ analysis of cytochrome *b* for the 85-taxa taset. The tree is mid-point rooted. Branches without posterior probability values are supported by less than % of the sampled trees.

6.4.3 The effect of intra-specific variation on phylogenetic patterns

Adding 23 more haplotypes for 16 species improved the phylogenetic resolution for both methods (Figs. 6.5, 6.6). The most striking result of this addition of taxa is the non-monophyly of many species' haplotypes, especially within *Andricus*, where only two of the 10 species, *A. coronatus* and *A. solitarius*, show monophyletic haplotypes (Figures 6.5, 6.6 – Table 6.4). Haplotypes from each non-monophyletic species are always grouped within the clade where each species belongs; for example, all the *Andricus caputmedusae* specimens are placed within the *Andricus quercuscalicis* clade. Non-monophyly is less pronounced outside *Andricus*, where 4 out of 6 species are monophyletic. Only two species, *C. quercus* and *C. quercusfolii* are not monophyletic, with only the *C. quercus* non-monophyly robustly supported. Monophyly is not correlated with levels of intra-specific divergence (Table 6.4).

6.4.4 Comparison of maximum likelihood and Bayesian inference

The two methods of statistical phylogenetic inference differed in branch support and topology (for example, compare Figs. 6.5, 6.6 for the *Cynips* genus). Each of the four $(MC)^3$ replicates for both datasets gave slightly different results although all of them had quickly converged on the same likelihood values. Interestingly, $(MC)^3$ usually gives higher support than ML (for example, compare branch support for the major *Andricus* clades in Figs. 6.5, 6.6), perhaps indicating that $(MC)^3$ is more strongly dependent on the underlying data, although many deviations from this pattern exist in our phylogenetic reconstructions. The evaluation methods for clade support provided by the two methods (bootstrap support in maximum likelihood and posterior probability in Bayesian inference) are not the same. However, no general pattern emerges from our analyses regarding correspondence between bootstrap support and posterior probabilities.

Table 6.4. Maximum sequence divergence of mitochondrial haplotypes for cytochrome *b* reported from the 16 species of this study for which more than one haplotype was included in the analysis (dataset II). The values estimated are the percentage of observed substitutions per nucleotide (*p* distance).

Species	*p*-value (%)	Grouping status
Andricus caputmedusae	4.39	Non-monophyletic
Andricus conificus	6.03	Non-monophyletic
Andricus coriarius	2.77	Non-monophyletic
Andricus coronatus	3.00	Monophyletic
Andricus dentimitratus	4.85	Non-monophyletic
Andricus lucidus	1.40	Non-monophyletic
Andricus mitratus	4.62	Non-monophyletic
Andricus panteli	1.16	Non-monophyletic
Andricus seckendorffi	1.19	Non-monophyletic
Andricus solitarius	2.54	Monophyletic
Biorhiza pallida	2.54	Monophyletic
Cynips cornifex	0.69	Monophyletic
Cynips longiventris	1.85	Monophyletic
Cynips quercus	6.47	Non-monophyletic
Cynips quercusfolii	2.08	Non-monophyletic*
Neuroterus macropterus	2.54	Monophyletic

* : weakly supported.

6.5 Discussion

The phylogenies reconstructed from the three gallwasp datasets show good resolution of lineages at intermediate distances and identify several robustly supported clades, but fail to reconstruct very recent or very old divergences. Multiple substitutions at high divergences and the high ratio of taxa to characters are the probable causes (Table 6.2, Fig. 6.2 – see below) (e.g. Cruickshank, *et al.*, 2001). When more than one haplotype for a given species is added the topologies obtained

are improved; this is probably due to the fact that the taxa included break long branches and help reduce the average branch length throughout the tree (Hillis, 1996; Swofford, *et al.*, 1996; but see Rosenberg and Kumar, 2001). The reconstruction of different topologies by the two methods of phylogenetic inference employed in this study, Bayesian inference and maximum likelihood, is also probably due to the high levels of divergence between the various genera. This is supported by the fact that incongruences in the topologies obtained by the two methods are more concentrated on the basal branches of the trees.

6.5.1 *Utility of cytochrome* b *as a molecular marker*

The phylogenetic analyses presented here show low resolution of basal lineages and evidence for substitution saturation of the cytochrome *b* marker (Fig. 6.2, Table 6.2). Limited resolution of deep branches using mitochondrial loci has been shown for other arthropod groups, but until recently this had been a property of phylogenetic studies addressing questions at higher taxonomic levels (e.g. among insect orders Liu and Beckenback, 1992; among beetle families Howland and Hewitt, 1995; among mite families Soller, *et al.*, 2001). However, an increasing number of studies are revealing poor resolution of mitochondrial phylogenies at lower taxonomic levels (e.g. Belshaw and Quicke, 1997; Gleeson, *et al.*, 1998, this study). One of the main problems with cytochrome *b*, and more generally with mitochondrial loci, for insect phylogenetics is that synonymous sites quickly become saturated with substitutions, whereas non-synonymous sites are 'not free to vary', due to selective constraints on protein function (Belshaw and Quicke, 1997; Brower and DeSalle, 1998; Lee, 1999). This saturation problem is aggravated by the high AT-bias in insect mitochondrial genomes (Table 6.3, especially in the third codon position) (e.g. Crozier, *et al.*, 1989; Dowton and Austin, 1995).

Differences in the rate of mtDNA evolution among taxonomic groups and the arbitrariness of taxonomic categories (see below) can only aggravate the problems related to choice of molecular markers for phylogenetics. A range of mechanisms has been proposed to account for differences in rates of mitochondrial DNA evolution among taxa, including generation time, metabolic rate, base composition and

population size (Rand, 1994), and rate acceleration has been demonstrated for honeybees and lice (Crozier, *et al.*, 1989; Crozier and Crozier, 1993; Hafner, *et al.*, 1994; Jermiin and Crozier, 1994; Page, *et al.*, 1998). Although there is no evidence that oak gallwasp mitochondrial have an elevated rate of molecular evolution relative to other hymenopterans, a more slowly evolving marker such as the long-wavelength opsin, 28S rDNA or elongation factor-1α (Cook, *et al.*, 2001, Chapters 3, 7, Appendix I; Rokas, *et al.*, 2001b) will be more useful for the resolution of the basal lineages in the oak gallwasp tribe.

6.5.2 *Cytochrome* b *pseudogenes (numts)*

Cloning of cytochrome *b* for the four oak gallwasp species that failed direct sequencing, revealed multiple cytochrome *b*-like sequences. Only one sequence from each species was found to be of the correct length and reading frame, suggesting that the other sequences found are probably nuclear mitochondrial pseudogenes (numts), as indicated by their mutational degeneration (presence of stop codons and / or indels) (Bensasson, *et al.*, 2001). Alternative explanations for the existence of mitochondrial-like sequences additional to the 'true' mitochondrial copy include heteroplasmy and intra-mitochondrial duplications (Mirol, *et al.*, 2000; Bensasson, *et al.*, 2001); reasons why both of these alternatives are rather unlikely explanations in studies like this one have been discussed in detail by Mirol *et al.* (2000). However, sequences with correct length and reading frame are not necessarily amplifications of the 'true' mitochondrial copy; evidence from other gallwasp species suggests that, occasionally, more than one haplotypes of correct length and reading frame exist (Chapter 5). This might be the case for *N. lanuginosus* and the reason for the non-monophyly of the genus *Neuroterus*, although the relationships within the clade including the *Neuroterus* species are not robustly supported. This is a serious problem for phylogenetic reconstruction, since the derived numts are essentially paralogues to the mitochondrial loci. However, further work on the nature of these mtDNA-like sequences and their phylogenetic implications requires establishment of their nuclear location (Bensasson, *et al.*, 2001).

6.5.3 Possible explanations for the non-monophyly of oak gallwasp intra-specific haplotypes

The two striking features from the analysis of dataset II are the rather high intra-specific divergences (Table 6.4) and the lack of monophyly for most species (Table 6.4, Figs. 6.5, 6.6). Coalescence theory suggests that lineage sorting of ancestral polymorphisms is statistically more likely to be shorter in mtDNA loci when compared with nuclear autosomal loci, due to the smaller effective population size of the former (Moore, 1995). The high percentage of non-monophyletic haplotypes for a given species of oak gallwasp in the mitochondrial phylogenies presented in this study suggest that the species phylogeny of oak gallwasps (at least at the within-genera level) may never be correctly reconstructed. Nuclear loci are statistically likely to be more prone to lineage sorting of ancestral polymorphisms than mitochondrial loci; hence, the proportion of nuclear autosomal lineage sorting events is expected to be much higher than the proportion observed for the mitochondrial data.

However, a number of alternative explanations may be considered for the apparent non-monophyly of many gallwasp species. Several published studies on insects exist where the mitochondrial genome from one species has introgressed into the nuclear genome background of another species (e.g. Powell, 1983; Mason, *et al.*, 1995; Deschamps-Cottin, *et al.*, 2000; Thelwell, *et al.*, 2000), and evidence from oak gallwasps suggests that matings across species borders do occur and produce viable offspring (Folliot, 1964), although the fitness of the hybrids is unknown. A case where mtDNA introgression may have occurred is found on the *A. quercuscalicis* clade (Figs. 6.5, 6.6). Specifically, the Turkish haplotypes from *A. mitratus*, *A. caputmedusae* and *A. dentimitratus*, together with *A. askewii* (also from Turkey – Appendix 6.1), form a separate monophyletic group distinct from the Italian and Balkan haplotypes of the same species which form their own monophyletic group. This is in contrast with the topologies obtained for other clades, e.g. *Cynips*. Assuming local hybridisation between these species might explain equally well this clustering of refuge-specific haplotypes shown in Figs. 6.5 and 6.6, but more work is needed to test this interesting hypothesis.

Another complication is associated with oak gallwasp lifecycles. The two generations (sexual and parthenogenetic) are morphologically very distinct, and rearing experiments are difficult. Parthenogenetic generation wasps can be up to five times as large as the sexual generation wasps, the latter being much more similar morphologically. The same is true for galls; parthenogenetic generation galls are large, long-lived and structurally diverse, while the sexual generation galls are usually small, ephemeral and much less variable in morphology (Ambrus, 1974; Stone, *et al.*, 1995; Stone and Cook, 1998). Furthermore, the gallwasp lifecycle cannot be completed in the laboratory and until recently, the pairing of the parthenogenetic and sexual generation of a species has required caging experiments in the field (Folliot, 1964; Pujade-Villar, 1992; Lund, *et al.*, 1998; Wehrmaker, 1998). Such experiments often fail due to our limited knowledge about the way female wasps select oviposition sites, a problem further intensified by the low number of generations each year. As a result, the two generations of many cyclical parthenogens were originally classified as separate species; the pairing of these generations into a cyclically parthenogenetic lifecycle has been slow and is far from complete (Folliot, 1964; Lund, *et al.*, 1998; Wehrmaker, 1998), and therefore clades containing sexual-only and parthenogenetic-only species may actually be the two alternate generations of a cyclical parthenogenetic species. The phylogenies presented here suggest two potential pairings. The first case is found on the *A. mayri* clade; if *A. aestivalis* (currently known only from its sexual generation) is the sexual generation *of A. lucidus* (currently known only from its parthenogenetic generation), then the apparent non-monophyly of *A. lucidus* haplotypes disappears (the *A. aestivalis* sequence is from Hungary and, as expected, groups with a haplotype from the same refuge – Appendix 6.1). A second case is found in the *A. hartigi* clade, where *A. conificus* (currently known only from its parthenogenetic generation) may be the parthenogenetic generation of *A. cydoniae* (currently known only from its sexual generation), although in this case the non-monophyly of *A. conificus* does not disappear.

The above-mentioned alternative explanations for the non-monophyly of the 10 taxa (introgressive hybridisation and pairing of alternate generations) suggest that our conclusions about the consequences of lineage sorting of ancestral

polymorphisms and the potential value of nuclear markers for oak gallwasp phylogenetics may be exaggerated. Furthermore, behavioural factors (e.g. polygyny, sex-biased dispersal) may be very important in determining the similarity between nuclear- and mitochondrial-derived trees (Hoelzer, 1997). More generally, although the transmission genetics of mitochondrial genomes reduces its effective population size, the natural history of the taxon in question may reduce the effective population size for autosomal nuclear loci, leading in both cases to the same effect; reduction of time to coalescence and a better chance that the gene phylogeny will be an accurate depiction of the species phylogeny (Moore, 1997). Support for the use of nuclear markers comes from long-wavelength opsin (LW *Rh*), where sequencing of a 590 base pair for 21 species of oak gallwasps (Cook, *et al.*, 2001, Chapter 7), showed both good resolution and high branch support values. However, the extent of intraspecific variation for this marker and the consequences of including alleles from different phylogeographic refuges have, as yet, not been examined.

6.5.4 Gene and species phylogenies

What is of interest to most phylogeneticists is the split of ancestral populations giving rise to new species, whereas what is recorded (in gene phylogenies) is the split between haplotype lineages, which by necessity happened before the population split (reviewed in Edwards and Beerli, 2000). Three are the main causes of discordance between gene and species phylogenies – assuming that estimation of the former is correct: lineage sorting of ancestral polymorphisms, horizontal transfer and gene duplication (Maddison, 1997). Our data suggest that all three processes may have occurred / are occurring in oak gallwasps, potentially causing the deviation of the cytochrome *b* phylogeny from that of the species'; lineage sorting is probably the reason for the non-monophyly of some species-specific haplotypes, whereas for others introgressive hybridisation (a form of horizontal transfer) of mtDNA is a more likely explanation. Finally, the multiple cytochrome *b*-like products recovered from some gallwasp species might be considered as paralogs.

6.5.5 Evolutionary implications of the European oak gallwasp phylogeny

The ancestral lifecycle for oak gallwasps is cyclical parthenogenesis; it involves an obligate alteration between a spring sexual generation and a summer/autumn sexual generation. In all European *Andricus* species each generation is specific to one of two oak sections. Parthenogenetic generation wasps gall white oaks (section *Quercus sensu stricto*), while sexual generation wasps are found either on white oaks or on black oaks (section *Cerris*), but not on both. All the species that use black oaks as hosts for their sexual generation form a single, robustly supported, clade (Figs. 6.3 – 6.6) (see also Cook, *et al.*, 2001, Chapter 7). Thus, it is very likely that there has been just one host plant shift for the sexual generation oak gallwasps from a white oak host to a black oak host, during the European radiation of *Andricus* (Cook, *et al.*, 2001, Chapter 7). Furthermore, this pattern is consistent with biogeographical evidence suggesting that oak gallwasps originated in the New World, where white oaks, but not black oaks, are found (Manos, *et al.*, 1999; Cook, *et al.*, 2001, Chapter 7). Consistent with this hypothesis is the fact that the four non-European *Andricus* species (*A. kashiwaphilus*, *A. mukaigawae* and *A. symbioticus* from Japan and *A. spectabilis* from the US) are all located basally on the tree, suggesting that use of two different oak hosts is a derived state and is only found in European species.

The phylogenies inferred from both methods suggest the existence of five major *Andricus* clades. The species found within each of these clades usually possess galls of similar structure, arguing for a phylogenetic constraint on gall structures, as has been shown previously for gallwasps (Stone and Cook, 1998), and other gall-forming insects (Stern, 1995; Crespi and Worobey, 1998; Nyman, *et al.*, 2000).

6.5.6 What is an insect genus?

Taxonomic categories are arbitrary and genera across insects show a wide range of genetic divergences (Table 6.3), as has been shown for vertebrate taxonomic categories (Johns and Avise, 1998). Underlying reasons for varying levels of genetic divergence at a given taxonomic level include variation in rate of evolution

(morphological or molecular), date of origin of lineage, or simply in the philosophy of the taxonomist ('splitter' or 'lumper', Johns and Avise, 1998; Rokas, *et al.*, 2001b, Chapter 3). Table 6.3 shows a compilation of studies with molecular divergence values for cytochrome *b* in a variety of insect genera. Both within the tribe of oak gallwasps as well as across insects there is substantial variance in the genetic diversity found within a genus. Data contained in Table 6.3 are probably incomplete due to the lack of standardised sampling; for some genera all the species have been sampled (e.g. *Ischnura*) whereas for others only a small, geographically biased, portion has been sampled (e.g. *Andricus*). This lack of sampling standards coupled with the increasing realisation that there is no correspondence of taxonomic categories between different organismal groups and the lack of fossil record as a 'yardstick' of time for most insect groups, make the prediction of usefulness of markers for insect phylogenetic questions extremely difficult.

Within oak gallwasps, the genus *Andricus* is the only one for which the phylogenies presented here strongly indicate that it is paraphyletic (or even polyphyletic), in addition to showing the highest species richness and the highest genetic distances. A plausible explanation for these attributes is that its classification has been based on ancestral morphological characters (plesiomorphies in the cladistic terminology) rather than on derived ones (apomorphies). This would imply that the genus *Andricus* may represent an ancestral morphological ground-plan from within which other genera (or *Andricus* clades) subsequently have evolved. Unfortunately, the morphology-based phylogenetic relationships of oak gallwasp genera which would test the validity of this hypothesis, have not, as yet, been examined.

6.6 Conclusion

A range of interesting molecular phylogenetic questions may be addressed in the study of the oak gallwasp tribe, that may generally apply for many insect groups. For example, lineage sorting of ancestral polymorphisms and genetic structure of populations due to Pleistocene glaciation cycles may have a serious effect on within-genus phylogenetics. Furthermore, the possibility of introgressive hybridization

between species and the reality of, as yet, unmatched generations make the task of phylogenetic reconstruction even more difficult.

In the attempt to resolve the phylogenetic history of a group such as oak gallwasps, where divergence times between taxa in a well-sampled data matrix differ by an order of magnitude, it is very likely that a single locus will not be effective at resolving all the nodes. A combined approach, employing several loci which show different and complementary rates of evolution and possibly different modes of evolution, is probably the only way to obtain well-supported phylogenetic hypotheses (Hillis, *et al.*, 1996; Rokas and Holland, 2000; Rokas, *et al.*, 2001b, Chapters 2, 3, Appendix I). Lessons learned from single-locus studies will be a most useful guide towards achieving this goal.

6.7 Acknowledgements

A version of this chapter is being prepared for publication: Rokas, A., G. Melika, Y. Abe, J.-L. Nieves-Aldrey, J. M. Cook and G. N. Stone. A mitochondrial molecular phylogeny of the European oak gallwasps (Hymenoptera: Cynipidae: Cynipini).

George Melika, Yoshihisa Abe and Jose-Luis Nieves-Aldrey identified and / or provided some of the species used in this study. Graham Stone and James Cook published in 1998 a subset of the dataset used in this chapter, with Graham also providing comments on the manuscript.

Appendix 6.1. Collection locations and type of material used in DNA extraction of the species studied.

Species	Location, Country	Tissue used in DNA extraction
Genus *Andricus*		
A. aestivalis	Tatabanya, Hungary	adult
A. aries	Ascott, U.K.	larva
A. askewii	Çekerek, Turkey	adult
A. burgundus	Madrid, Spain	adult
A. caliciformis	Gödöllõ, Hungary	pupa
A. caputmedusae	Valtice, Czech Republic	adult
	Montevulture, Italy	adult
	Küllüce, Turkey	adult
A. conglomeratus	Gödöllõ, Hungary	adult
A. conificus	Veszprem, Hungary	pupa
	Massa Marittima, Italy	adult
A. coriarius	Mátrafüred, Hungary	larva
	Edessa, Greece	adult
	El Escorial, Spain	adult
A. coronatus	Sopron, Hungary	adult
	Massa Marittima, Italy	adult
	Küllüce, Turkey	adult
A. corruptrix	Oxford, U.K.	adult
A. crispator	Köszeg, Hungary	adult
A. curvator	Bükk Mountains, Hungary	adult
A. cydoniae	Mátrafüred, Hungary	larva
A. dentimitratus (formerly known as *A. viscosus*)	Sopron, Hungary	adult
	Siena, Italy	adult
	Beybesli, Turkey	adult
A. fecundator	Gödöllõ, Hungary	larva

A. gallaetinctoriae	Madenli, Turkey	adult
A. gemmeus	Gödöllõ, Hungary	larva
A. glutinosus	Vienna, Austria	adult
A. grossulariae	Romhány, Hungary	adult
A. hartigi	Szentkut, Hungary	adult
A. hungaricus	Gödöllõ, Hungary	adult
A. inflator	Oxford, U.K.	adult
A. kashiwaphilus	Hyogo, Japan	adult
A. kollari	Gödöllõ, Hungary	adult
A. lignicolus	Randalstown, Ireland	adult
A. lucidus	Gödöllõ, Hungary	adult
	Egirdir, Turkey	adult
	Antalya, Turkey	adult
A. malpighii (formerly known as *A. nudus*)	Bibulano, Italy	larva
A. mayri	Bra, Italy	adult
A. mediterraneae	Madenli, Turkey	adult
A. mitratus	Beybesli, Turkey	adult
	Mátrafüred, Hungary	adult
A. mukaigawae	Osaka, Japan	adult
A. panteli	Chiusi, Italy	adult
	Radicofani, Italy	adult
	Azrou, Morocco	adult
A. polycerus	Ruffeno, Italy	pupa
A quadrilineatus	Oxford, U.K.	adult
A. quercuscalicis	Oxford, U.K.	adult
A. quercusramuli	Mátrafüred, Hungary	adult
A. quercustozae	Sopron, Hungary	adult
A. schroeckingeri	Jászberény, Hungary	adult
A. seckendorffi	Bra, Italy	adult

	Hadim, Turkey	adult
A. solitarius	Gödöllõ, Hungary	adult
	St. Chinian, France	adult
A. spectabilis	Ventura, California – USA	adult
A. symbioticus	Hyogo, Japan	adult
A. theophrastii	Beysehir, Turkey	larva
A. tinctoriusnostrus	Madenli, Turkey	adult
A. tomentosus	Arnissa, Greece	adult
Genus *Cynips*		
C. cornifex	Szentkut, Hungary	pupa
	Komnina, Greece	adult
C. divisa	Gödöllõ, Hungary	adult
C. disticha	B. de Mervent, France	adult
C. longiventris	Istria, Croatia	adult
	Çekerek, Turkey	adult
C. quercus	Szentkut, Hungary	pupa
	Poppi, Italy	adult
	Yeşilyurt, Turkey	adult
C. quercusfolii	Olympiada, Greece	adult
	Mátrafüred, Hungary	adult
Genus *Neuroterus*		
N. lanuginosus	Sirok, Hungary	adult
N. macropterus	Gödöllõ, Hungary	adult
	Arnissa, Greece	adult
N. saliens	Mátrafüred, Hungary	adult
Other genera		
Aphelonyx cerricola	Valtice, Czech Republic	adult
Biorhiza pallida	Zliv, Slovakia	adult
	Villaviciosa, Spain	adult
Callirhytis glandium	Gyöngös, Hungary	adult

Chilaspis nitida	Gödöllõ, Hungary	adult
Plagiotrochus quercusilicis	El Pardo, Spain	adult
Trichagalma serratae	Osaka, Japan	adult
Trigonaspis synaspis	Suluova, Turkey	adult

CHAPTER 7

Evolutionary shifts between host oak species and host plant organs in *Andricus* gallwasps

7.1. Abstract

Wasps in the genus *Andricus* (Hymenoptera: Cynipidae) induce galls on oak trees. Their ancestral lifecycle is cyclical parthenogenesis, with obligate alternation between a spring sexual generation and a summer/autumn parthenogenetic generation. In all European *Andricus* species each generation is specific to one of two oak sections. The parthenogenetic generations of all European *Andricus* species gall white oaks (section *Quercus sensu stricto*), while sexual generations are found either on white oaks or on black oaks (section *Cerris*), but not on both. We estimate phylogenies using both mitochondrial and nuclear gene sequence data and demonstrate that: (1) a lifecycle involving both sexual and parthenogenetic generations on white oaks is ancestral, and (2) there has been just one host plant shift during the European radiation of *Andricus*, giving rise to a clade whose sexual generation develops on black oaks. This pattern is consistent with biogeographical evidence that oak gallwasps originated in the New World, where white oaks, but not black oaks, are found. *Andricus* species are also highly specific as to the oak organ (e.g. leaf, fruit, catkin) that they gall. However, we use a maximum likelihood approach to show that (3) evolutionary changes in gall location occur at a significantly higher rate than evolutionary changes in host plant usage. In combination with other evidence, our results suggest that (4) the ability to induce galls is more dependent on oak genotype than on the characteristics of particular host tissues.

7.2 Introduction

Insect herbivory covers an extremely broad range of insect-plant interactions, of which the most intimate links gall inducers and their hosts (Cornell, 1983; Price, *et al.*, 1987; Stone, *et al.*, 2002). Gall inducers are found in a wide range of insect families, and several taxa (such as the cynipid gallwasps, the cecidomyiid gallmidges, and the pemphigid aphids) contain many gall inducing species. Gall inducers cause host plant tissues to differentiate into novel structures, ranging from endosperm-like inner nutritive tissue to woody outer protective tissues (Shorthouse and Rohfritsch, 1992; Crespi, *et al.*, 1997; Nyman, *et al.*, 1998; Stone and Cook, 1998; Schönrogge, *et al.*, 2000). While the mechanisms involved in manipulation of host plant development remain largely unknown, detailed studies of gallwasps (Hymenoptera: Cynipidae), sawflies (Hymenoptera: Tenthredinidae), aphids (Homoptera: Pemphiginae) and gallmidges (Diptera: Cecidomyiidae) have revealed the interactions between the gall inducers and host tissues to be extremely complex (Weis and Abrahamson, 1986; Wood and Payne, 1988; Fay and Hartnett, 1991; Zantoko and Shukle, 1997; Akimoto, 1998; Brooks and Shorthouse, 1998; Kearsley and Whitham, 1998; Zantoko and Shukle, 1999; Nyman and Julkunen-Tiito, 2000; Schönrogge, *et al.*, 2000).

Gall inducing insects commonly show two types of specificity. First, most are specific to a particular taxon (species, or group of related species) of host plants (Cornell, 1985, 1996; Crespi, *et al.*, 1997; Abrahamson, *et al.*, 1998; Melika, *et al.*, 2000; Nyman and Julkunen-Tiito, 2000). Second, they commonly show a high degree of organ specificity, preferentially attacking just one out of leaves or buds or roots or fruits (Weis, *et al.*, 1988; Shorthouse and Rohfritsch, 1992; Abrahamson, *et al.*, 1998; Melika, *et al.*, 2000). The diversity of host plants and host plant organs exploited by gall inducers raises an interesting question. Is it easier to shift between two organs on the same plant species, or to shift between two host species while attacking the same plant organ? In other words, in a phylogeny of gall inducers, do closely related species gall the same plant organ, or the same host plant species? It is also possible that these two types of shift might occur coincidentally during the radiation of gall inducers (galling of a new host is associated with galling of a new

plant organ), or occur independently and so frequently that neither type of shift shows any phylogenetic pattern.

While the physiological and morphological correlates of gall induction are becoming increasingly well understood in a few gall systems (Weis and Abrahamson, 1986; Wood and Payne, 1988; Fay and Hartnett, 1991; Brooks and Shorthouse, 1998), with the notable exception of the hessian fly (a cecidomyiid gall midge, Zantoko and Shukle, 1997, 1999) too little is known about the metabolic basis of tissue and host specificity to make predictions about phylogenetic patterns in the use of either host organ or host species. From the point of view of a gall inducer, alternative organs on the same plant (for example, roots and flowers) may constitute environments that are just as different in terms of host plant gene expression as the same organs on two different host species. If a galler is critically dependent on an organ-specific tissue (such as endosperm in seeds), we might expect transitions between host plants to be more common than intra-specific transitions between host organs. In contrast, if the ability to gall a given host species is determined by specific genetic loci in gall inducer and host (as it is in the hessian fly, Ratcliffe, *et al.*, 1994; Formusoh, *et al.*, 1996; Zantoko and Shukle, 1997, 1999; Ratcliffe, *et al.*, 2000), plant genotype rather than plant organ may be crucial. For the vast majority of gall inducers, it remains unknown whether host plant or organ specificity results from a precise physiological compatibility only with certain plant tissues, from egg-laying preferences of the gall inducer, or from a combination of both (for example, see Whitham, 1978; Abrahamson and Weis, 1997). To date there has been no phylogenetic comparison of frequencies of gall location and host plant changes in any gall inducing taxon. Here we address this issue by examining evolutionary patterns in gallwasps of the genus *Andricus* (Hymenoptera: Cynipidae) feeding on oak (genus *Quercus*).

The genus *Andricus* is ideal for such an analysis because it is species-rich (with approximately 100 species in Europe, Askew, 1984) and shows considerable diversity in the two key characters (gall location and host plant) (Nieves-Aldrey, 1987; Cook, *et al.*, 1998; Melika, *et al.*, 2000). *Andricus* lifecycles primitively involve two generations each year — a sexual one in the spring, and a parthenogenetic one in the summer/autumn (Askew, 1984; Cook, *et al.*, 1998; Stone,

et al., 2001). Galls of a given generation are also generally highly specific to a particular taxon of oaks. As in some other gall inducers (e.g thrips, Crespi, *et al.*, 1997; Crespi and Worobey, 1998), specificity is usually demonstrated at the level of host species groups, rather than individual species. All *Andricus* host plants belong to the subgenus *Quercus* and this is divided into four sections (Manos, *et al.*, 1999), of which only two (the white oaks of section *Quercus sensu stricto* and the black oaks of section *Cerris*) are present in Europe. All known parthenogenetic generations of European *Andricus* species attack white oaks (e.g. *Q. infectoria, Q. faginea, Q. petraea, Q. pubescens, Q. pyrenaica* and *Q. robur*). In contrast, sexual generations of *Andricus* species gall either white or black oaks (e.g. *Q. cerris, Q. suber*), but never both (see Fig. 7.1). These generalisations hold also for the many *Andricus* species that are currently known from either only a sexual or only a parthenogenetic generation (Nieves-Aldrey, 1987; Melika, *et al.*, 2000).

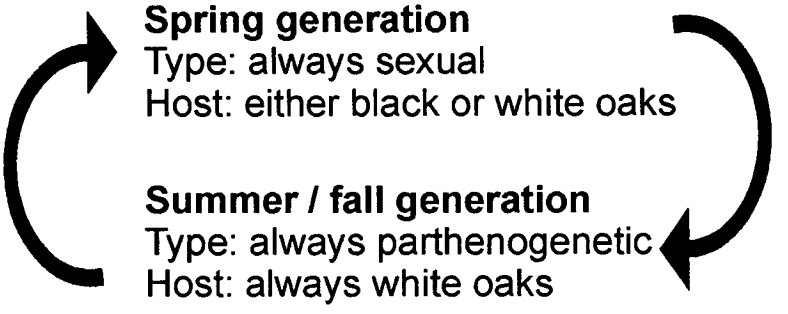


Figure 7.1. The life cycle of European *Andricus* gall wasps.

Each generation of each *Andricus* species is also highly specific with regard to the oak organ galled, and with few exceptions attacks only one of buds, shoots, roots, leaves, catkins, or acorns. The majority of known sexual generation *Andricus* galls are induced on buds or catkins, with fewer species galling stems (Appendix 7.1). Parthenogenetic generation galls are also most abundant on buds, with lower numbers on acorns, and very few species attacking catkins or roots. Leaves are very rarely galled by either sexual or parthenogenetic generations of European *Andricus* species, although members of other European oak cynipid genera, such as *Cynips*,

commonly gall leaves (Appendix 7.1) (as do some north American *Andricus* species - Weld, 1957, 1959, 1960).

In this paper we use sequence data for fragments of two genes (mitochondrial cytochrome *b* and nuclear long-wavelength opsin) to estimate the phylogeny of 34 European cynipid gallwasp species, including 32 from the key genus *Andricus*. We then map the host oak species and host organ attacked onto the phylogeny and ask: (1) How frequently have the sexual generations of *Andricus* gallwasps shifted between white and black oaks? and (2) How frequently have *Andricus* species shifted between alternative oak organs? Taken together, these questions will reveal whether speciation in gallwasps is more commonly associated with shifts between host plants or between host organs.

In addition, we address a phylogeographic issue in the origin of gallwasps attacking black oaks. Oak gallwasps (Cynipini) are far more species-rich in north America (473 species) than in Europe (144) (Askew, 1984) and several genera, including *Andricus, Callirhytis, Cynips, Dryocosmus* and *Neuroterus*, occur in both regions (Weld, 1957, 1959, 1960; Melika, *et al.*, 2000). Phylogeographic analyses of rose-feeding gallwasps of the genus *Diplolepis*, which show a similar geographic variation in diversity, suggest an origin in north America (Plantard, *et al.*, 1998b). A possible explanation for the patterns seen in oak gallwasps is that these too radiated in the Americas and reached Europe via Asia and a Bering land bridge. While white oaks are widespread in north America, black oaks are restricted to Europe and western Asia (Camus, 1936-1954). This predicts that possession of a sexual generation galling black oaks (absent from north America) should be derived from an ancestral state of both generations on white oaks (common in north America).

7.3 Materials and Methods

7.3.1 Study species

We include data for 32 European *Andricus* species and 2 outgroup species (*Biorhiza pallida* and *Cynips quercus*) from closely related cynipid genera (Appendix 7.1). The collection locations and life stages used in DNA extraction for 23 *Andricus* species

and the 2 outgroups are listed in Table 1 of Stone and Cook (1998). Names of 2 species have been modified relative to Stone and Cook (1998) in line with recent revisions (Melika, *et al.*, 2000). The additional species included in this study are *Andricus aestivalis, A. cydoniae, A. malpighii, A. mitratus, A. panteli, A. quadrilineatus, A. quercusramuli, A. schroeckingeri* and *A. tinctoriusnostrus* (Table 7.1).

Table 7.1. Collection locations, life stage used in DNA extraction and generation of the species studied here that are not included in Stone and Cook (1998).

Species	Generation	Collecting location	DNA extract
A. aestivalis	sexual	Tatabanya, Hungary	adult
A. cydoniae	sexual	Mátrafüred, Hungary	larva
A. malpighii	parthenogenetic	Bibulano, Italy	larva
A. mitratus	parthenogenetic	Bebesli, Turkey	adult
A. panteli	parthenogenetic	Poppio, Italy	adult
A quadrilineatus	parthenogenetic	Oxford, U.K.	adult
A. quercusramuli	parthenogenetic	Mátrafüred, Hungary	adult
A. schroeckingeri	sexual	Jászberény, Hungary	adult
A. tinctoriusnostrus	parthenogenetic	Madenli, Turkey	adult

7.3.2 Host plants and gall location

The oak taxa and plant organs galled by the study species are shown in Appendix 7.1. Data were collected throughout the known distributions of each species, extending in some cases from Iberia across Europe into Asia Minor, both by literature survey (Houard, 1912; Buhr, 1965; Sternlicht, 1968; Ambrus, 1974; Chodjai, 1980; Nieves-Aldrey, 1987; Pujade, 1994; Melika, *et al.*, 2000) and our own extensive collections throughout the study area over many years.

Our sample of *Andricus* species can be broken down as follows:

a) 'sexual generation only' species, which gall black oaks (5 species);

b) 'parthenogenetic generation only' species, which gall white oaks (13 species);

c) species with alternating sexual and parthenogenetic generations, both on white oaks (7 species);
d) species with alternating sexual and parthenogenetic generations on black oaks and white oaks respectively (6 species).

For one species (*A. dentimitratus*), a sexual generation is known from adult insects but host plant association and/or gall location data are unconfirmed. It is worth reiterating that there are no *Andricus* species that have a parthenogenetic generation (alone or in alternation with a sexual generation) on black oaks (see Fig. 7.1). Both outgroup species have alternating sexual and asexual generations on white oaks.

7.3.3 Molecular methods

For each species DNA was extracted from a single adult wasp as described in Stone and Cook (1998), or by using the DNeasy Tissue kit (QIAGEN cat. 69504) and following the protocol of the manufacturer for insect DNA extraction. Sequencing was carried out using the Perkin-Elmer BigDye Terminator chemistry and an ABI 377 sequencer. For both genes, PCR products were sequenced in both directions to minimise PCR artefacts, ambiguities and base-calling errors, and the sequences aligned by eye.

7.3.4 Cytochrome b

A 433 base pair fragment of the mitochondrial cytochrome *b* (cyt*b*) gene was amplified using the CB1/CB2 primer combination (Stone and Cook, 1998; Rokas, *et al.*, 2001a; Stone, *et al.*, 2001, Chapter 4, Appendix I). The total volume of 3 PCR reactions for each individual wasp was electrophoresed on a 1% agarose gel. The expected bands were cut from the gel and cleaned with the QIAQuick Gel Extraction kit (QIAGEN cat. 28704) before sequencing. All sequences were 433 bases long with full ORFs, and are deposited in Genbank (accession numbers AJ228448-AJ228454, AJ228456, AJ228458-61, AJ228463-72, AJ228474-75, AJ228478, AJ228481, AJ131065-AJ131068).

7.3.5 Long-wavelength opsin

A 588-590 base pair fragment of the nuclear gene long-wavelength opsin (LW*Rh*) was amplified by PCR using the LW*Rh*F/LW*Rh*R primer combination (Mardulyn and Cameron, 1999), using the conditions described in Rokas *et al.* (2001b, Chapter 3). PCR products were cloned using the TOPO TA Cloning kit (Invitrogen, cat. 4500-01) and 2-4 clones from each specimen were subsequently sequenced. Plasmids containing the fragment of interest were isolated using the QIAprep Spin Miniprep kit (QIAGEN, cat. 27104) and sequenced. All sequences will be available from Genbank on publication. The alignments and consensus trees will be available from TreeBASE (http://www.herbaria.harvard.edu/treebase) on publication.

7.3.6 Phylogenetic analyses

We used a Bayesian Markov chain Monte Carlo (MCMC) analysis to estimate both the trees and the posterior probabilities of nodes in the trees (Larget and Simon, 1999). The posterior probability of a node gives the probability that the node is true, conditional upon the data and the model of sequence evolution, and thus gives a direct indication of confidence in that node (Lutzoni, *et al.*, 2001). The MCMC program used is available from M. Pagel by request.

We estimated phylogenies separately for the cyt*b* and LW*Rh* datasets, because there are several species in the cyt*b* data set that do not appear in the LW*Rh* data set. For each gene, we used MCMC procedures to generate a population of 2,000,000 trees and then sampled 2 independent replicates of 2000 trees to estimate the posterior probabilities of nodes in the trees. The MCMC procedure samples trees from the total 'population' according to their probability of occurrence under the specified model of molecular evolution, while the 2 independent runs provide an indication of the variability between replicates. We use the HKY model of sequence evolution with gamma rate heterogeneity among sites. Tree sampling was delayed until after the first 10,000 trees in each run to avoid including trees sampled before Markov chain convergence (Lutzoni, *et al.*, 2001).

For comparison, we also analysed the data using standard maximum parsimony (MP) and maximum likelihood (ML) approaches. Standard ML phylogenies were built using PAUP* versions 4.0b7-8 (Swofford, 2000), after estimating the best-fit model using MODELTEST 3.0 (Posada and Crandall, 1998). Parameters allowed to vary were base composition, substitution rates and rate heterogeneity among sites. The ML model of sequence evolution was identified using likelihood ratio tests (Huelsenbeck and Rannala, 1997). The parameter values suggested by MODELTEST were then used to perform heuristic searches (with tree bisection and reconnection - TBR) on 100 bootstrap replicates. MP phylogenies were also generated with PAUP*, using heuristic searches, involving TBR branch-swapping with 100 random stepwise additions of taxa. Characters and changes were all weighted equally and the consensus trees were subjected to bootstrapping (500 replicates).

7.3.7 Gall locations in sexual and asexual generations

For descriptive purposes we mapped gall locations for each generation onto a phylogeny and inferred the minimum number of changes in gall location using MACCLADE 4.0 (Maddison and Maddison, 2000). Gall location was treated as an unordered, multistate character with five states (Appendix 7.1, Figure 7.4) and was mapped onto cyt*b* trees because these include more species than, and are highly consistent with, the LW*Rh* trees. Since data are not available for all species in both generations we removed taxa with missing data from the trees. This left 27 species for the asexual generation and 18 species for the sexual generation.

7.3.8 Comparing rates of change in host plant and gall location in the sexual generation

We used log likelihood comparisons to test whether the rate of change in gall location was significantly different from the rate of change in host plant exploitation, using the computer programmes DISCRETE (Pagel, 1994, 1999b) and MULTISTATE (Pagel, unpublished). Both use a continuous time Markov model of character evolution but, while DISCRETE can only analyse binary characters, MULTISTATE can accommodate up to 6 character states. Host plant was treated as an unordered character with two states (black or white oaks) and, therefore, two possible types of change. Gall location was treated as an unordered character with three states (bud, catkin, leaf) and consequently 6 different types of change, considering all possible states and also both possible directions of change. Given our overwhelming evidence that white oaks are the ancestral host plant, we imposed this constraint in the analyses.

Initially, we fitted a binary model (A) to the host plant data and estimated two parameters that describe the rates of change from white (state 0) to black (state 1) oaks. Thus, the rate of forward changes from white to black oaks is denoted q_{01} and the corresponding rate for backwards changes from black to white oaks as q_{10}. We then fitted two different models to the gall location data. The first was a full model (B) with three character states and therefore 6 different rate of change parameters (q_{01}, q_{10}, q_{02} , q_{20}, q_{12}, q_{21}). The second was a simplified model (C) that considered just forwards and backwards changes from any state and therefore has just the two parameters (q_{01} and q_{10}), like model A. If model C was not significantly worse (by comparison of log likelihoods) than model B, then we could use q_{01} and q_{10} from model C to compare overall rates of change in host plant and gall location. To do this we analysed model D, which used the gall location data from model C but the host plant rate parameters (q_{01} and q_{10}) from model A. Comparison of the log likelihoods of models C and D then provided a test of whether rates of host plant and gall location changes were significantly different. Since we were not dealing with nested hypotheses, significance could not be tested using the χ^2 distribution, so we instead

applied the general 'rule of thumb' that 2 log likelihood units constitutes a significant difference (Pagel, 1999b).

We used only the cyt*b* phylogenies because these include more species, and are, in any case, highly consistent with the LW*Rh* trees. We conducted each analysis twice, using the two different consensus trees (1 and 2) generated by the two independent runs of the MCMC phylogenetic analysis. The respective phylogenies were also pruned to remove the two outgroups (which have states not represented by the ingroup), as well as species for which the gall location is unknown for a particular generation (see Appendix 7.1). After pruning, 18 species remained in the trees for character analysis.

7.4 RESULTS

7.4.1 Phylogenies

Figures 7.2 and 7.3 show the 50% majority-rule consensus trees from the MCMC analyses of cyt*b* and LW*Rh* data. There is great consistency in the trees estimated from the two different genes. In fact, the only substantial disagreement is that *A. inflator* clusters with the outgroups in cyt*b* trees but as the basal member of the *Andricus* clade in LW*Rh* trees. Even in this case, the dubious cyt*b* placement is due to a node with only low support (PP = 64%) and may be attributable to long branch attraction (see Fig. 7.2). This reflects a general pattern that cyt*b* resolves intermediate nodes, and especially tips, quite well but provides little resolution at deeper nodes (see also Stone and Cook, 1998, Chapter 6). In contrast, the LW*Rh* trees resolve basal *Andricus* relationships (e.g. the *A. fecundator* clade) rather well (see Figs. 7.2 and 7.3). These differences are probably due to the faster rate of evolution, larger number of taxa and smaller length of sequence in the cyt*b* data set. Nevertheless, we emphasise that there is great congruence between the two analyses. In particular, there are no relationships supported strongly by one gene that are contradicted by the other gene. The main difference is that some deeper nodes that are unresolved or weakly supported by cyt*b*, receive strong support from LW*Rh*. In fact, it appears that

LW*Rh* is an exceptionally informative gene for analysing the relationships of interest in this study.

The two independent samples of 2000 trees from the MCMC analysis yielded toplogically identical consensus trees for LW*Rh* but two slightly different trees for cyt*b*. The cyt*b* trees were essentially very similar to each other, expect that tree 1 recovered two clades with weak support (posterior probabilities of 52 and 55 respectively) that were not recovered in tree 2 (Fig 7.2). We label five clades of *Andricus* species in Figures 7.2 and 7.3. These are consistent with those recognised by Stone and Cook (1998), but now contain more species.

The phylogenies estimated using standard MP and ML analyses were highly consistent with those using the MCMC approach so, for brevity, we do not report them in great detail here. However, the bootstrap values from MP and ML analyses are displayed in Figures 7.2 and 7.3 and comparison of the results of different analysis procedures reveals three main patterns. First, LW*Rh* trees are close to identical by all methods of analysis, with strong support for most nodes with all methods (Fig. 7.3). Second, cyt*b* trees are generally less well-resolved than LW*Rh* trees but still show great consistency between analysis methods (Fig. 7.2); there are no nodes receiving > 50% support with any method that conflict with results from another method. Third, MCMC generally (but not always) returns higher support for nodes that ML or MP. This is most evident for deeper nodes in the cyt*b* trees, which are poorly supported under MP and, to a lesser extent, ML. However, with the exception of the placement of *A. inflator* (discussed above), the nodes supported by MCMC alone in cyt*b* analyses are supported by all methods in the LW*Rh* data set.

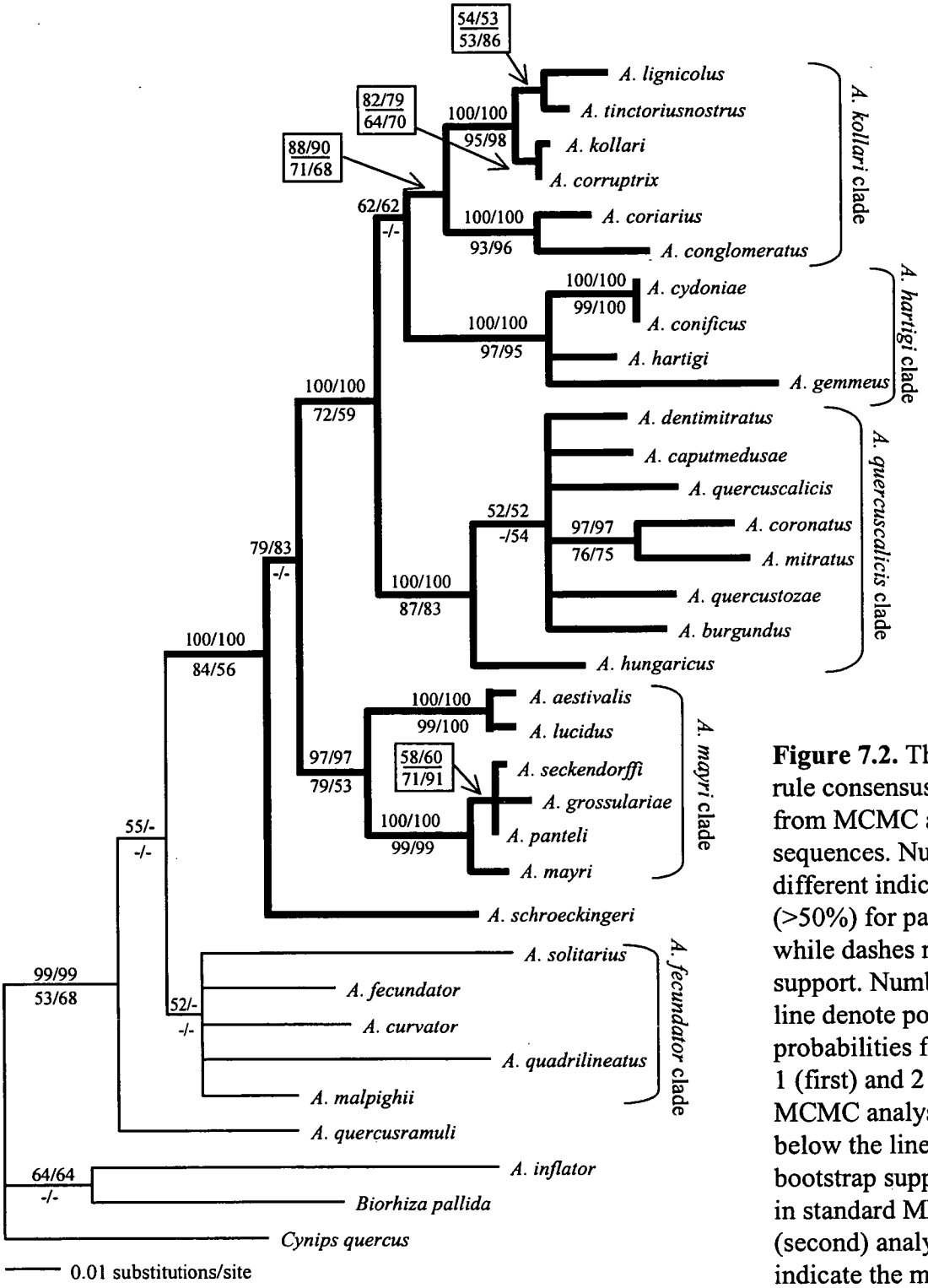


Figure 7.2. The 50% majority rule consensus phylogram from MCMC analyses of cyt*b* sequences. Numbers show different indices of support (>50%) for particular clades while dashes represent <50% support. Numbers above the line denote posterior probabilities for tree samples 1 (first) and 2 (second) in the MCMC analyses. Numbers below the line indicate bootstrap support for the node in standard ML (first) and MP (second) analyses. Thick lines indicate the monophyletic derived clade containing all species that form galls on black oaks.

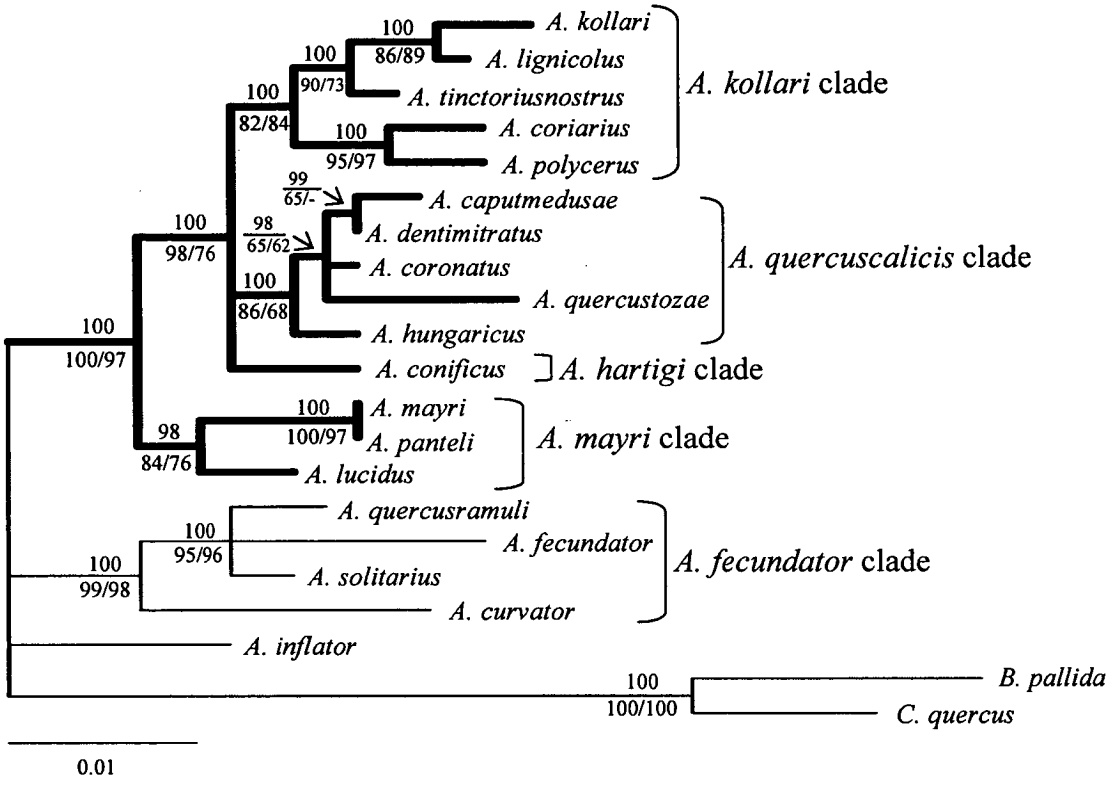


Figure 7.3. The 50% majority rule consensus phylogram from MCMC analyses of LW*Rh* data. Numbers show different indices of support (>50%) for particular clades while dashes represent <50% support. The number above the line denotes the posterior probability of the node in the MCMC analyses (both replicates yielded the same values). The two numbers below the line indicate bootstrap support for the node in standard ML (first) and MP (second) analyses. Thick lines indicate the monophyletic derived clade containing all species that form galls on black oaks.

7.4.2 Host plant associations

The most striking result is that all *Andricus* species whose sexual generations induce galls on black oaks lie in a monophyletic derived clade (Figs. 7.2, 7.3). The monophyly of this clade is supported by all phylogenetic methods and by both data sets. In particular, the MCMC analyses find a posterior probability of 100% for this node in both sequence data sets. This suggests that in the genus *Andricus* there has been only a single host plant shift to black oaks, and that a lifecycle in which sexual generations gall white oaks is the ancestral state.

7.4.3 Gall locations

In the parthenogenetic generation, most basal *Andricus* species induce bud galls, with the exception of *A. quadrilineatus*, which galls catkins (Fig. 7.4A). There must have been a minimum of five changes in gall location, although there could have been more, depending partly on the real branching pattern of the poorly resolved *A. quercuscalicis* clade (Fig. 7.4A). Shifts from buds to catkins are inferred to have happened independently in *A. quadrilineatus* and *A. mayri*, while shifts from buds to acorns have taken place independently in *A. seckendorffi* and at least once in the *A. quercuscalicis* clade. A single colonisation of lenticel buds from shoot buds is inferred for the common ancestor of the *A. hartigi* clade (Fig. 7.4A). Since there were several changes in gall location but no change in host plant use in the parthenogenetic generation, we did not make any further comparison of these rates of change.

In the sexual generation, gall locations are split fairly evenly between catkins and buds and there must have been at least five changes (Fig. 7.4B). Most species gall buds or catkins and changes in both directions are implied. In addition, there were probably independent shifts from catkins to leaves in *A. curvator* and *A. schroeckingeri*.

7.4.4 Comparing rates of change in host plant and gall location in the sexual generation

The models listed in Table 7.2 are labelled according to the cyt*b* phylogeny (1 or 2) and transition parameters used (A-D). The analyses all yielded higher estimates of rates of forward (q_{01}) and backward (q_{10}) change for gall location than for host plant (Table 7.2). Considering the analyses using phylogeny 1 (models 1A-1D), the simple (2 parameter) model 1C for gall location was not significantly worse than the full (6 parameter) model 1B (LR = 0.25). Consequently, we were able to substitute the q_{01} and q_{10} values from the host plant analysis (model 1A) into the simple binary model of location change to generate model 1D. The likelihood of 1D, a model of gall location changes that uses q_{01} and q_{10} from host plant analysis, was significantly worse (6.4 log units) than the likelihood of model 1C (using q_{01} and q_{10} from the gall location analysis. The results were qualitatively identical, and quantitatively very similar, when the same series of analyses (models 2A-2D in Table 7.2) was performed using phylogeny 2). In summary, we can conclude that changes in gall location have been significantly more frequent than changes in host plant use during the radiation of European *Andricus* species.

Table 7.2. Results of ML investigation of rates of change of host plant and gall location in the sexual generation.

Model	Character	Tree	States	Parameters	q_{01}	q_{10}	L
1A	Host	1	binary	1A	0.235	0.039	-18.23
1B	Location	1	multistate	1B	—	—	-15.74
1C	Location	1	binary	1C	0.907	0.911	-15.99
1D	Location	1	binary	FROM 1A	0.235	0.039	-22.39
2A	Host	2	binary	2A	0.238	0.029	-19.57
2B	Location	2	multistate	2B	—	—	-15.76
2C	Location	2	binary	2C	1.03	1.03	-16.02
2D	Location	2	binary	FROM 2A	0.238	0.029	-22.87

7.4.5 A new lifecycle pair?

Two insects previously regarded as separate species - the sexual only *A. cydoniae* and the parthenogenetic only *A. conificus* yielded identical DNA sequences for the cyt*b* fragment. Each 'species' was formerly known only from a single generation, and our finding is consistent with them being alternate generations of a single *Andricus* species. The two 'species' are both quite rare but have similar geographic distributions and, although further studies are required, it seems likely that they belong to the same species (see below).

7.5 Discussion

7.5.1 Host oak associations

Previous studies have demonstrated that gallwasps are extremely sensitive to genotypic variation in host oaks. In America, where gallwasp host plants belong either to section *Lobatae* (red oaks) or section *Quercus sensu stricto* (white oaks) (Manos, *et al.*, 1999), no cynipids form galls on members of both sections (Cornell, 1985, 1996; Abrahamson, *et al.*, 1998). A European gallwasp, *A. kollari*, appears to show ecotypic specialisation to two different sexual generation hosts that is specific enough to limit gene flow between the ecotypes (Stone, *et al.*, 2001, Appendix I). Sensitivity to host genotype can also be fine enough for gallwasps to discern levels of introgression in a hybrid zone between two oak species, one of which is the natural host (Boecklen and Spellenberg, 1990; Aguilar and Boecklen, 1991; Moorehead, *et al.*, 1993). While these studies show that gallwasp-oak interactions are highly sensitive to the oak genotype, they provide no information about the number of evolutionary transitions between oak taxa.

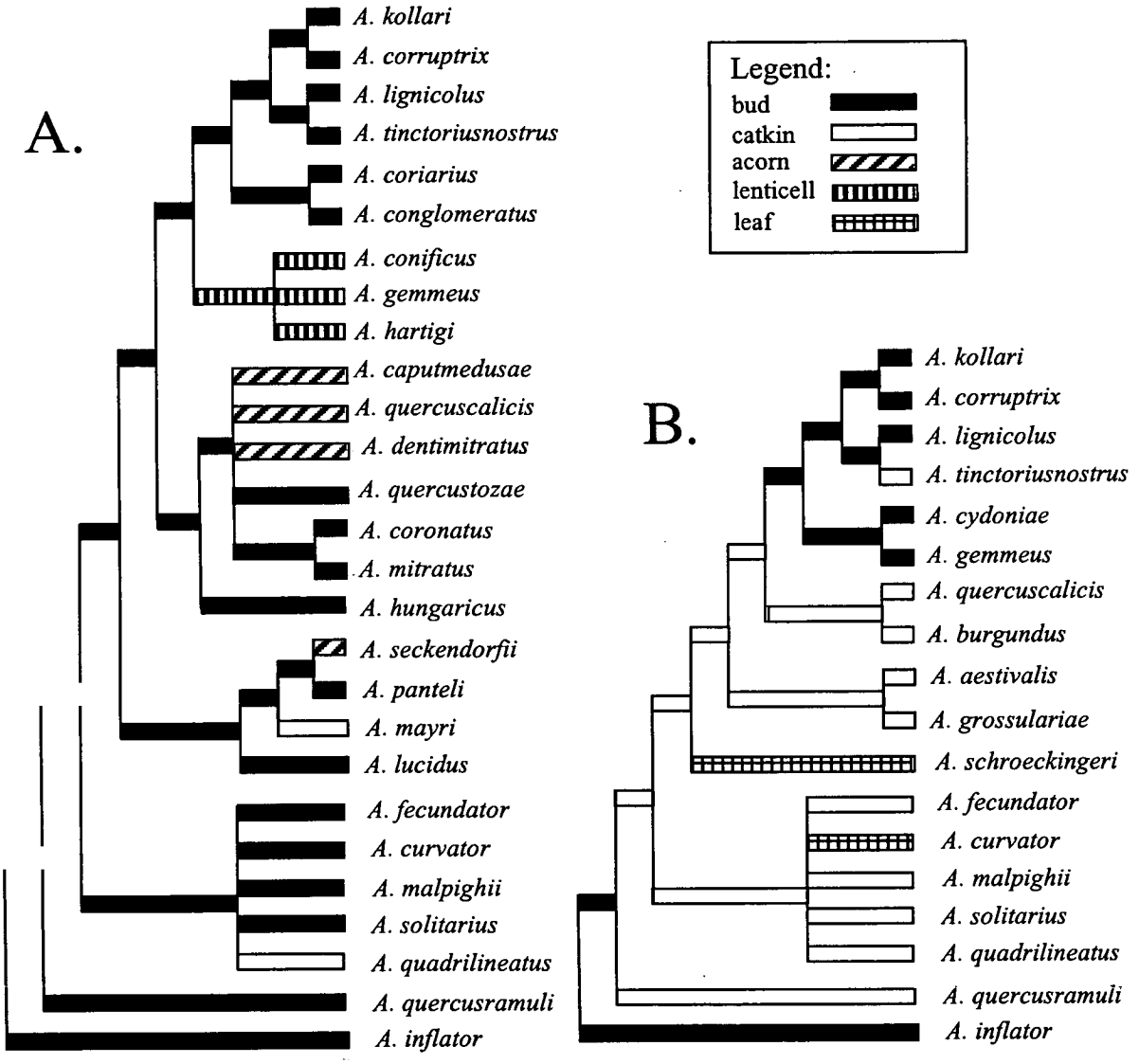


Figure 7.4. The phylogenetic distributions of known gall locations for the parthenogenetic (A) and sexual (B) generations of *Andricus*. The phylogenies used are pruned versions of Figure 7.2, from which the species with unknown gall locations have been removed. Each figure shows one of the most parsimonious reconstructions, which involves five changes in each generation.

Our analyses show that exploitation of black oaks is a derived state in *Andricus,* and, furthermore, that the transition of sexual generations to black oaks has occurred only once. For one member of the host alternating clade, *A. dentimitratus*, the sexual generation insect is known, but its oak host remains unknown. Based on its phylogenetic position (see Fig. 7.2), we predict its host to be a black oak. Population genetic analyses indicate that some of the 'parthenogenetic only' species (*A. caputmedusae, A. coriarius, A. lucidus, A. panteli, A. quercustozae* and *A. seckendorffi*) in the host alternating clade are actually cyclical parthenogens whose sexual generation is as yet undiscovered (Atkinson, 2000; Atkinson, *et al.*, 2001; Stone, *et al.*, 2001, Appendix I). Again, we predict that the sexual generations of these species will be found to develop on black oaks.

Host associations in European *Andricus* are thus extremely conserved, a pattern also revealed by phylogenetic analyses of gall inducer/host plant interactions in both thrips (Crespi and Worobey, 1998) and aphids (Stern, 1995). The rarity of host shifts in gallwasps in general is also supported by the observation that, with the exception of *Andricus* and *Callirhytis*, all genera of oak cynipids which alternate generations use the same host oak section in both generations (Weld, 1957, 1959, 1960; Folliot, 1964; Askew, 1984; Nieves-Aldrey, 1987; Melika, *et al.*, 2000). This is also true for the Pediaspidini, the sister group of the Cynipini (Folliot, 1964; Melika, *et al.*, 2000). We therefore predict that a suitable analysis will show that host transitions between white and red oaks in the American oak cynipid fauna have also been rare.

Gall induction on black oaks must have evolved several times in other oak cynipids. One independent colonisation is represented by the genus *Synophrus*, which lies within the predominantly inquiline tribe Synergini, and is well separated phylogenetically from the oak gallwasp tribe Cynipini (Ronquist, 1994, 1995; Liljeblad and Ronquist, 1998). In the Cynipini, gall induction on black oaks is found in seven European genera (*Andricus*, *Aphelonyx*, *Callirhytis*, *Chilaspis*, *Dryocosmus*, *Neuroterus* and *Plagiotrochus*, Nieves-Aldrey, 1987; Melika, *et al.*, 2000). Of these, four (*Andricus, Callirhytis, Dryocosmus* and *Neuroterus*) are also found on other oak taxa in north America. If north America is indeed the centre of origin of these Holarctic genera, then colonisation of black oaks must have occurred independently in each genus. *Aphelonyx, Chilaspis* and *Plagiotrochus* are not found in north

America (Weld, 1957, 1959, 1960), and (with one possible exception) are known only from black oaks (Houard, 1912; Buhr, 1965; Sternlicht, 1968; Ambrus, 1974; Chodjai, 1980; Nieves-Aldrey, 1987; Pujade, 1994; Melika, *et al.*, 2000). These genera presumably represent endemic Palaearctic radiations that evolved only after the divergence of black oaks. Inference of the number of additional colonisation events of black oaks represented by these genera must await more extensive phylogenetic analyses of the Cynipini as a whole.

7.5.2 Evolutionary changes in gall location

In contrast to the conservation of host plant association, the organ galled by particular *Andricus* species has changed at least five times in each generation during the radiation of the genus in Europe. Changes have occurred between a range of host organs and reversions seem very likely to have occurred in the sexual, and probably also the parthenogenetic, generation (see Fig. 7.4). Furthermore, formal comparison of rates of change in the sexual generation has shown that locations shifts are more frequent than host plant shifts. Overall, this suggests that the mechanisms associated with gall induction are less specific to particular oak tissues than they are to oak taxa and is consistent with the ability of a few gallwasp species to induce structurally similar galls on different oak tissues. For example, galls of the sexual generation of *Neuroterus quercusbaccarum* are commonly found on either leaves or catkins, *Andricus curvator* is able to induce its sexual generation galls on leaves or buds, and *Andricus lucidus* is able to induce galls on buds or, rarely, acorns. Folliot (1964) showed that females of some *Andricus* species caged on an oak organ type where their galls were never found in nature would, eventually, oviposit and that their larvae could, in some cases, induce successful galls. These results suggest that gall location is probably more a function of female oviposition behaviour than of specificity in the gall induction process.

Within *Andricus*, the conserved exploitation of lenticel buds by the group of species in the *A. hartigi* clade is particularly interesting (Fig. 7.4A). Relatively few oak cynipids gall buds on mature branches and tree trunks and members of the *A. hartigi* clade are noteworthy for the diversity of unique gall structures they show

(Stone and Cook, 1998). It is tempting to suggest that the structural diversity in the *A. hartigi* clade is related to this exploitation of a novel gall location within the genus.

7.5.3 Independent evolutionary trajectories of sexual and parthenogenetic generations

No parthenogenetic *Andricus* gall is able to develop on black oaks, despite the fact that the sexual generations of some species possess the necessary traits. A consistent (but inverse) pattern is seen in some European members of the genus *Callirhytis*, in which the parthenogenetic generation galls develop only within the acorns of black oaks, while the sexual generation galls develop under the bark of white oaks (Nieves-Aldrey, 1987; Nieves-Aldrey, 1992). As for *Andricus*, western Palaearctic members of *Callirhytis* are probably derived from north American ancestors, whose parthenogenetic generation galls develop in the acorns of red and white oaks (Weld, 1957, 1959, 1960; Cornell, 1985). In *Callirhytis,* only the parthenogenetic generation has made the transition to black oaks.

Other traits of gallwasps, including gall structure and gall location, also show different patterns of evolutionary change in sexual and parthenogenetic generations (Stone and Cook, 1998). Taken together, these patterns suggest that the evolutionary trajectories of the two generations of the gallwasp lifecycle are to some extent independent, and agree with general conclusions concerning the evolution of components of complex lifecycles (Moran, 1994). This raises an interesting issue in the cases of *Chilaspis nitida* and *Dryocosmus cerriphilus*, two oak cynipids in which both sexual and parthenogenetic generations gall black oaks. If the outgroup for these species had both generations on white oaks, then there are two alternative mechanisms by which the derived lifecycle may have evolved. First, there may have been two independent colonisation events, one in each generation. Second, a single generation may have evolved novel traits allowing colonisation of black oaks, followed by expression of the same traits in the other generation.

7.5.4 Possible consequences of a host plant shift

Farrell (Farrell, 1998) and von Dohlen and Moran (2000) have used sister group comparisons to show that host shifts from gymnosperms to angiosperms were associated with large increases in species richness of beetles and aphids, respectively. While these studies focus on the diversification of higher taxonomic levels (tribes and above) since the Cretaceous, the same methods may be used to investigate more recent events at lower taxonomic levels. Inspection of Figure 7.2 and 7.4 suggests that the host-alternating clade of *Andricus* may have higher species richness than its sister group, the *A. fecundator* clade. However, any general conclusion concerning changes in species richness associated with colonisation of black oaks requires better estimates of species richness in the respective clades, as well as further independent comparisons between sister taxa. One candidate is the genus *Neuroterus,* which has substantial numbers of European species that exploit black oaks, as well as many species restricted to white oaks. Unfortunately, however, there has been no phylogenetic study of this genus to date. If host alternation is generally associated with an increase in the speciation rate in gallwasps, it would represent an intriguing contrast with aphids, in which transition to complex, host-alternating lifecycles may represent an evolutionary dead-end (Moran, 1988).

A striking feature of the host-alternating clade of *Andricus* is that it contains all of the sampled species for which only a parthenogenetic generation is known, and all of those for which only a sexual generation is known. Species without matching generations either possess cryptic generations yet to be discovered, or have secondarily lost them. Facultative secondary loss of the sexual generation has been demonstrated for *Andricus quadrilineatus* (Folliot, 1964) and obligate loss in a Japanese *Andricus* species complex (Abe, 1986). However, the frequency of genuine lifecycle simplification remains unknown since population genetic studies suggest that a number of apparently purely parthenogenetic species may in fact possess cryptic sexual generations (Atkinson, 2000; Atkinson, *et al.*, 2001; Stone, *et al.*, 2001, Appendix I). Nevertheless, the possibility of a link between colonisation of black oaks by the sexual generation and the probability of loss of the same generation remains intriguing.

Evidence from gall-inducing aphids shows that there may be strong fitness trade-offs (in terms of performance on the host plant) associated with maintaining a host-alternating phenotype (Moran, 1991) and this alone might favour collapse to a single host lifecycle. Demographic considerations suggest a second possible reason for life cycle simplification — populations of host-alternating *Andricus* species can persist only where suitable white and black oaks grow together. This requirement generates severe geographic range restrictions, which can be revealed dramatically when the distribution of one or both hosts is increased by human activity (Hails and Crawley, 1991; Stone and Sunnucks, 1993; Sunnucks and Stone, 1996; Atkinson, 2000; Rokas, *et al.*, 2001a, Chapter 4, Appendix I; Stone, *et al.*, 2001). Using the very general approximation that 2.3% sequence divergence in mtDNA indicates 1 million years since separation (Brower, 1994), the host-alternating *Andricus* clade has been diversifying for about 10 million years (Stone and Cook, 1998). Over this time scale, the distributions of oak species in Europe have varied enormously with the advance and retreat of quaternary ice sheets (Huntley and Birks, 1983; Bennett, 1986; Toumi and Lumaret, 1998; Hewitt, 1999, 2000). Significantly for host alternators, patterns of northward range expansion also differ substantially between oak taxa, such that areas in which black oaks and white oaks occur together are both relatively restricted in area and unstable in time. This demographic constraint may have selected for repeated lifecycle simplification in members of the host-alternating clade.

7.6 Acknowledgements

A version of this chapter has been submitted for publication: Cook, J.M., A. Rokas, M. Pagel and G. N. Stone. Evolutionary shifts between host oak species and host plant organs in *Andricus* gallwasps.

James Cook, Antonis Rokas and Graham Stone had an equal contribution to the production of this work, both in generating the data as well as analysing them. Mark Pagel provided the Bayesian phylogenetic analyses and the maximum likelihood ancestral state reconstructions.

Appendix 7.1. Host plants and gall locations of species included in the study.

Wasp species	Sexual generation galls		Parthenogenetic generation galls	
	Location	**Host oak subgenus**	**Location**	**host oak subgenus**
genus *Andricus* HARTIG, 1840				
A. aestivalis GIRAUD, 1859	Catkin	*Cerris*	—	—
A. burgundus GIRAUD, 1859	Catkin	*Cerris*	—	—
A. caputmedusae (HARTIG, 1843)	—	—	Acorn	*Quercus*
A. conglomeratus (GIRAUD, 1859)	—	—	Bud	*Quercus*
A. conificus (HARTIG, 1843)	—	—	Bud (lenticel)	*Quercus*
A. coriarius (HARTIG, 1843)	—	—	Bud	*Quercus*
A. coronatus (GIRAUD, 1859)	—	—	Bud	*Quercus*
A. corruptrix (SCHLECHTENDAL, 1870)	Bud	*Cerris*	Bud	*Quercus*
A. curvator HARTIG, 1840	Bud/leaf	*Quercus*	Bud	*Quercus*
A. cydoniae GIRAUD, 1859	Bud	*Cerris*	—	—
A. dentimitratus REJTO, 1887 — formerly known as *A. viscosus* NIEVES ALDREY, 1986	*	*Cerris**	Acorn	*Quercus*
A. fecundator (HARTIG, 1840)	Catkin	*Quercus*	Bud	*Quercus*
A. gemmeus (GIRAUD, 1859)	Bud	*Cerris*	Bud (lenticel)	*Quercus*

A. grossulariae GIRAUD, 1859	Catkin	*Cerris*	—	—
A. hartigi (HARTIG, 1843)	—	—	Bud (lenticel)	*Quercus*
A. hungaricus (HARTIG, 1843)	—	—	Bud	*Quercus*
A. inflator HARTIG, 1840	Bud	*Quercus*	Bud	*Quercus*
A. kollari (HARTIG, 1843)	Bud	*Cerris*	Bud	*Quercus*
A. lignicolus (HARTIG, 1840)	Bud	*Cerris*	Bud	*Quercus*
A. lucidus (HARTIG, 1843)	—	—	Bud	*Quercus*
A. malpighii (ADLER, 1881) - formerly known as *A. nudus* ADLER, 1881	Catkin	*Quercus*	Bud	*Quercus*
A. mayri (WACHTL, 1879)	—	—	Catkin	*Quercus*
A. mitratus (MAYR, 1870)	—	—	Bud	*Quercus*
A. panteli (KIEFFER)	—	—	Bud	*Quercus*
A. polycerus GIRAUD	—	—	Bud	*Quercus*
A. quadrilineatus HARTIG, 1840	Catkin	*Quercus*	Catkin	*Quercus*
A. quercuscalicis (BURGSDORF, 1783)	Catkin	*Cerris*	Acorn	*Quercus*
A. quercusramuli (LINNAEUS, 1761)	Catkin	*Quercus*	Bud	*Quercus*
A. quercustozae (BOSC, 1792)	—	—	Bud	*Quercus*
A. schroeckingeri WACHTL, 1876	Leaf	*Cerris*	—	—
A. seckendorffi (WACHTL, 1879)	—	—	Acorn	*Quercus*

A. solitarius (BOYER DE FONSCOLOMBE, 1832)	Catkin	*Quercus*	Bud	*Quercus*
A. tinctoriusnostrus DE STEFANI	Catkin	*Cerris*	Bud	*Quercus*
Outgroups				
Biorhiza pallida (OLIVIER, 1791)	Bud	*Quercus*	Root	*Quercus*
Cynips quercus (FOURCROY, 1785)	Bud	*Quercus*	Leaf	*Quercus*

CHAPTER 8

Distribution and diversity of *Wolbachia* in gallwasps (Hymenoptera; Cynipidae)

8.1 Abstract

Bacteria of the genus *Wolbachia* are widespread endosymbionts of arthropods and nematodes that are responsible for a number of reproductive alterations in their hosts. Although our knowledge on *Wolbachia* is advancing rapidly, we still lack fundamental ecological data regarding its prevalence in different taxonomic groups, its association with particular life histories and the relative importance of the various pathways via which infection occurs. Gallwasps (Hymenoptera; Cynipidae) represent an attractive system for addressing these questions because they show different reproductive modes (cyclical parthenogenesis, arrhenotoky and thelytoky) and two linked trophic niches (gall-inducers, and inquilines that have lost the gall-inducing ability and instead inhabit the galls of other species). In this study, we have: (i) PCR-screened for the presence of *Wolbachia* in 609 wasps from 64 gallwasp species; and (ii) sequenced two *Wolbachia* genes (*wsp* and *ftsZ*) from individuals of all the infected species. We show that: (i) the prevalence of *Wolbachia* differs between lineages with different reproductive modes; (ii) *Wolbachia* is not associated with thelytoky in the cyclical parthenogenetic gallwasp lineage (in contrast, thelytoky in the arrhenotokous lineages is, most likely, *Wolbachia*-induced); and (iii) horizontal transfer of *Wolbachia* is likely to have occurred between gall-inducers and their associated inquilines.

8.2 Introduction

Bacteria of the genus *Wolbachia* are a widespread group of endosymbionts belonging to the alpha-proteobacteria (O'Neill, *et al.*, 1992). They are inherited cytoplasmically (i.e. transmitted from mother to daughter) and alter reproduction in their arthropod hosts in a number of ways, including cytoplasmic incompatibility, male killing, feminisation and imposition of parthenogenesis (see O'Neill, *et al.*, 1997, and chapters therein). In addition to immediate reproductive modifications, *Wolbachia* infection has a range of longer term evolutionary impacts on host taxa (O'Neill, *et al.*, 1997; Werren, 1997; Hurst and Schilthuizen, 1998; Werren, 1998; Stouthamer, *et al.*, 1999).

Levels of *Wolbachia* infection have been examined in random assemblages of arthropods and in distinct taxonomic groups. Screening studies of a wide variety of insects have revealed that *Wolbachia* bacteria are extremely common, infecting 16-22% of insects (e.g. Werren, *et al.*, 1995b; West, *et al.*, 1998; Werren and Windsor, 2000), with one study indicating that the percentage might be even higher (Jeyaprakash and Hoy, 2000). Data from fruit-flies (Bourtzis, *et al.*, 1996), mites (Breeuwer and Jacobs, 1996), crustaceans (Bouchon, *et al.*, 1998), ants (Wenseleers, *et al.*, 1998), stalk-eyed flies (Hariri, *et al.*, 1998), rose gallwasps (Plantard, *et al.*, 1999) and mosquitoes (Kittayapong, *et al.*, 2000) suggest that infection levels may vary both within and between different assemblages (West, *et al.*, 1998; Werren and Windsor, 2000). *Wolbachia* infection levels may also vary intra-specifically and do so in many arthropods (e.g. Jiggins, *et al.*, 2001a), including gallwasps (Plantard, *et al.*, 1998a; Rokas, *et al.*, 2001a, Chapter 4, Appendix I).

Gallwasps are a family of wasps (Hymenoptera: Cynipidae) that parasitise herb and tree species, inducing gall formation (Askew, 1984; Stone, *et al.*, 2002). The 1369 described species found in Cynipidae are currently divided, on the basis of morphology, into six tribes (see Fig. 8.1) (Liljeblad and Ronquist, 1998). Tribe Cynipini species (oak gallwasps) gall mostly oaks (*Quercus*) and other Fagaceae, tribe Pediaspidini species gall only maple trees (*Acer*), one tribe galls roses (rose gallwasps, tribe Diplolepidini), one tribe galls the genera *Acacia* and *Prosopis* from the Fabaceae (tribe Eschatocerini), whereas another tribe galls eudicot herb species

(tribe 'Aylacini'; this tribe is paraphyletic and will be shown in inverted commas). Most members of the sixth tribe (tribe Synergini) have lost the ability to induce galls themselves but instead develop as inquilines inside the galls of other cynipids (Ronquist, 1994). Inquilines do not kill their gall-inducing host directly, though attack by some species usually results in death of the gall-inducer (Schönrogge, *et al.*, 1995; Stone, *et al.*, 2002).

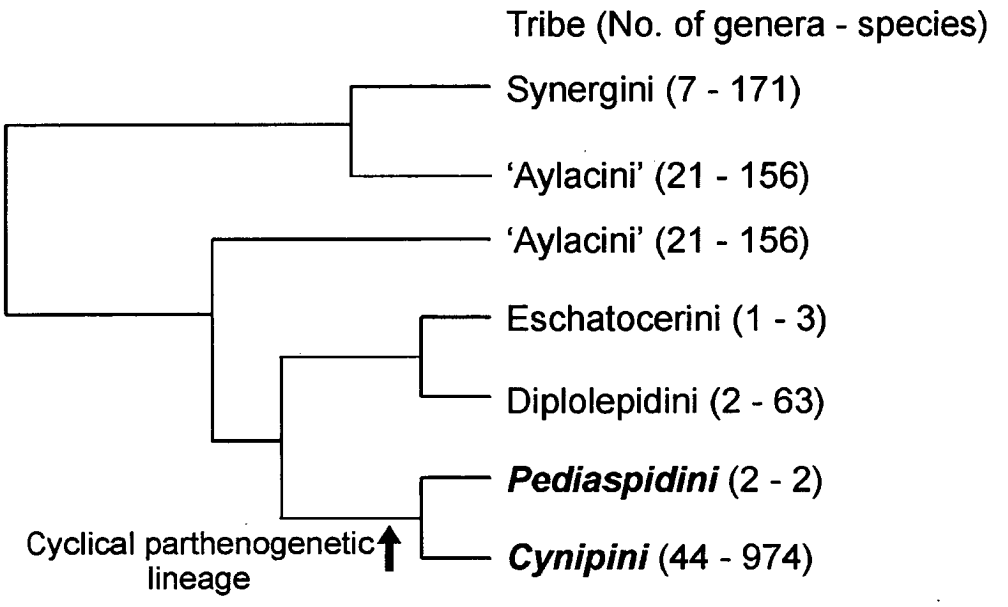


Figure 8.1. The phylogenetic relationships between the different tribes within Cynipidae (data from Liljeblad and Ronquist 1998). The cyclical parthenogenetic tribes are shown in bold-italics. All the other tribes are arrhenotokous. The number of genera and species within each tribe are indicated in the parentheses after each tribe' s name.

As for other Hymenoptera, gallwasps are haplodiploid, i.e. female wasps are diploid and males haploid. Three different modes of reproduction exist within the Cynipidae; arrhenotoky, cyclical parthenogenesis and thelytoky (see lifecycle column in Appendix 8.1). Arrhenotoky (unfertilized eggs develop parthenogenetically into males; fertilized eggs develop into females) is probably the ancestral mode of reproduction for the whole family and is currently found in representatives of four tribes of gallwasps ('Aylacini', Synergini, Diplolepidini and

Eschatocerini). Cyclical parthenogenesis (the strict alternation of one arrhenotokous and one thelytokous generation - symbolized as CP) is restricted to two tribes, the Pediaspidini and the Cynipini, in a single monophyletic lineage (Liljeblad and Ronquist, 1998; Stone, *et al.*, 2002). Additionally, a number of species in both arrhenotokous and cyclical parthenogen lineages have reverted to thelytoky (or parthenogenesis; unfertilized eggs give rise to female offspring).

Loss of sex (i.e. a transition from ancestral arrhenotoky to secondary thelytoky) is widespread in the 'Aylacini' and Diplolepidini, and in all cases studied to date thelytoky is tightly correlated with *Wolbachia* infection (Plantard, *et al.*, 1998a; Plantard, *et al.*, 1999). Cytological evidence from *Diplolepis rosae* (a thelytokous rose gallwasp infected with *Wolbachia*) suggests that the mechanism of thelytoky is gamete duplication (Stille and Davring, 1980), the same mechanism via which *Wolbachia* bacteria induce parthenogenesis in other hymenopterans (Stouthamer and Kazmer, 1994; Stouthamer, 1997). In gamete duplication, the two identical sets of chromosomes do not separate during the anaphase of the first mitotic division and the net result of the first mitotic cycle is a single nucleus containing two copies of the same set of chromosomes (Stouthamer and Kazmer, 1994).

In the Cynipini, a number of species are known only from one or other generation, leading to the suggestion that CP species may have lost a generation from their lifecycle and reverted either to thelytoky or to arrhenotoky only. Species without matching generations either possess cryptic generations yet to be discovered, or have secondarily lost them. Population genetic analyses indicate that at least some of these apparently thelytokous species are actually CP species for which an arrhenotokous generation exists but is as yet undiscovered (Atkinson, 2000; Atkinson, *et al.*, 2001; Stone, *et al.*, 2001, Appendix I, Atkinson, *et al.*, in preparation). Facultative secondary loss of the arrhenotokous generation has however been demonstrated for *Andricus quadrilineatus* by Folliot (1964), and obligately in one *Andricus* species complex by Abe (1986). The frequency of genuine lifecycle simplification in oak cynipids remains unknown. Although *Wolbachia* has been shown to induce thelytoky in other gallwasp tribes (see above), it is not known if it is responsible for the thelytokous species in the CP lineage.

Pathways via which *Wolbachia* is transmitted in natural populations have been analyzed in few assemblages. Most studies performed on the phylogenetic congruence of hosts and their *Wolbachia* strains strongly support frequent horizontal transfer of strains between different hosts (e.g. Werren, *et al.*, 1995a; Schilthuizen and Stouthamer, 1997). Direct support for the possibility of horizontal transfer comes from the experimental demonstration that *Wolbachia* can be transmitted by infection (Huigens, *et al.*, 2000). However, the route by which this horizontal transfer occurs is less clear. One possibility is that parasitoids may be the agents of transfer (Werren, *et al.*, 1995a), although an early attempt to test this hypothesis on a leaf-miner and an aphid food web provided negative results (West, *et al.*, 1998). More recent work on a *Drosophila*-Hymenoptera web shows cases of sharing of the same *Wolbachia* strains between parasitoids and their insect hosts (Vavre, *et al.*, 1999). The different results obtained by these two studies may be, at least partially, explained by the fact that until recently a fast-evolving gene which would accurately discriminate among the different *Wolbachia* strains was lacking (Zhou, *et al.*, 1998); West *et al.* (1998) used the slow-evolving gene *ftsZ*, whereas Vavre *et al.* (Vavre, *et al.*, 1999) used the fast-evolving gene *wsp*. The detailed natural history knowledge of the host-parasitoid system found within the family of gallwasps (consisting of the gall-inducers, their inquilines as well as other parasitoid taxa) represents an attractive ecological assemblage for the study of *Wolbachia* transmission (e.g. Askew, 1961; Askew, 1984).

In an attempt to reveal the extent and pattern of *Wolbachia* infection in the monophyletic clade of gallwasps, we have PCR-assayed 609 wasps of 64 different species belonging to 3 gallwasp tribes (Appendix 8.1) and sequenced two *Wolbachia* genes (*wsp* and *ftsZ*) from individuals of those species infected. By combining the data presented in this study with published data on two other tribes of Cynipidae (Plantard, *et al.*, 1999) we are able to address several questions on *Wolbachia* ecology. Specifically: (i) what is the percentage of *Wolbachia* infection in the family of gallwasps; (ii) is there any variation in infection rates for the gallwasp lineages with different reproductive modes (cyclical parthenogenesis and arrhenotoky); (iii) is *Wolbachia* associated with thelytoky in the cyclical parthenogenetic gallwasp lineage; and (iv) is there any evidence for horizontal transfer between different

gallwasp species and, in particular, between gall-inducers and their cynipid inquilines.

8.3 Materials and Methods

8.3.1 Collection and DNA extraction

609 specimens of 64 gallwasp species (Appendix 8.1) were sampled by rearing or dissection of galls collected throughout Europe during an on-going set of projects on this taxon (Cook, *et al.*, 1998; Stone and Cook, 1998; Atkinson, 2000; Atkinson, *et al.*, 2001, Chapters 3, 4 and Appendix I; Rokas, *et al.*, 2001a ; Rokas, *et al.*, 2001b; Stone, *et al.*, 2001). Oak gallwasps can be identified on the basis of the gall they induce, and all species used in this study were also identified from sequence data from a 433 base pair fragment of the mitochondrial cytochrome *b* gene (Cook, *et al.*, 1998; Stone and Cook, 1998; Rokas, *et al.*, 2001a; Rokas, *et al.*, 2001b, Chapters 3, 4, 6, 7 and Appendix I; Stone, *et al.*, 2001) (data not shown – see Chapter 6). Inquiline taxa were identified using morphology-based keys recently developed by one of the authors (J.-L. Nieves-Aldrey). For 18 species 20-30 specimens from many different localities were collected in order to test for intra-specific geographic variation in *Wolbachia* infection. Galls from many gallwasp species are multilocular (more than one offspring emerge from a single gall). To minimise screening of siblings in such cases we used one female from each gall.

DNA was extracted from fresh individuals using the DNeasy Tissue kit (QIAGEN cat. 69504) and following the protocol of the manufacturer for insect DNA extraction. To avoid contamination, prior to DNA extraction each female wasp was soaked in 5% bleach for a few seconds and then serially rinsed in drops of sterile water (as in West, *et al.*, 1998). With each DNA extraction three control extractions were performed using a *Nasonia Wolbachia*-positive strain, a *Nasonia Wolbachia*-negative strain and a no-DNA sample (Werren, *et al.*, 1995b; West, *et al.*, 1998).

8.3.2 Wolbachia *screening*

Screening for *Wolbachia* was performed by PCR using the *Wolbachia*-specific primers for the *ftsZ* cell-cycle gene ftsZF1(Werren, *et al.*, 1995b) and WOLG-R (Rokas, *et al.*, 2001a, Chapter 4, Appendix I). All PCRs were performed in a PTC-200 DNA engine (MJ Research). Screening PCRs were attempted for sample DNA extractions at dilutions ranging from one tenth to one hundredth. Control PCRs were always performed. The PCR cycle for *ftsZ* was: one cycle of 94 °C for 3 min, 55 °C for 90 sec and 72 °C for 5 min, followed by 35 cycles of 94 °C for 30 sec, 55 °C for 90 sec, 72 °C for 5 min and a final extension step at 72 °C for 5 min. The PCRs were performed in 25μl volumes and consisted of 1μl of DNA sample, 2.5μl of 10x PARR Buffer (HYBAID), 1μl of $MgCl_2$ (25 mM), 0.5μl of dNTPs (10mM), 0.35μl of each primer (20mM), 0.25μl of Taq (Promega) and 19.05μl of distilled, deionized H_2O. 15μl of each reaction were electrophoresed on a 1% ethidium bromide-stained agarose gel. To check that samples negative for *Wolbachia* were not because of: (i) failed DNA extraction; (ii) presence of PCR inhibitors; and (iii) incorrect DNA concentration, control PCRs with the general eukaryotic 28S rDNA primers 28Sf and 28Sr were performed as described in Werren *et al.* (1995b). Control PCRs for a number of specimens were also performed with the mitochondrial cytochrome *b* primers CB1 and CB2 (e.g. Rokas, *et al.*, 2001a, Chapter 4, Appendix I). These primers were shown to be more accurate indicators of the DNA quality of the specimen than 28S primers (data not shown), perhaps due to their lower generality as universal primers. All specimens that failed to amplify the 28S rDNA and/or cytochrome *b* fragments were excluded from subsequent analysis.

Jeyaprakash and Hoy (2000) have recently suggested that *Taq* DNA polymerase often fails to amplify *Wolbachia* genes and that this has caused underestimation of the incidence of *Wolbachia* in arthropods in general. They argued that a greater number of infections was detected when a mixture of the enzymes *Taq* and *Pwo* was used for PCR amplification. In this study we have attempted to estimate the prevalence of *Wolbachia* in gallwasps by employing the 'standard' protocol (which uses *Taq* DNA polymerase) for *Wolbachia* screening first

established by Werren *et al.* (1995b) and subsequently used by many authors (e.g. Wenseleers, *et al.*, 1998; West, *et al.*, 1998; Werren and Windsor, 2000). Our choice of PCR screening protocol is based on the fact that we wanted to compare levels of *Wolbachia* prevalence between different taxonomic groups; therefore only use of the 'standard' protocol would allow us meaningful comparisons. Furthermore, Werren and Windsor (2000) obtained similar results using two different PCR assays (*ftsZ* and 16S rDNA); given the evolutionary conservation of 16S rDNA it is unlikely that, at least, the 16S primers failed to detect *Wolbachia* in a significant portion of the sample. Nevertheless, a comparative study of the sensitivity of the different protocols and primers available is needed to evaluate robustly potential biases/errors in currently used assays.

8.3.3 PCR amplification and sequencing

Wolbachia diversity in infected populations was assessed by sequencing of two *Wolbachia* genes, *wsp* and *ftsZ*. *wsp* is the most polymorphic gene so far isolated from *Wolbachia* (Zhou, *et al.*, 1998) and hence the most likely to distinguish between two closely related *Wolbachia* strains, whereas *ftsZ* has been the most widely used *Wolbachia* gene. Two recent studies have presented direct (Werren and Bartos, 2001) and indirect (Jiggins, *et al.*, 2001b) evidence of recombination in the *wsp* gene. This poses a serious obstacle for phylogenetic reconstruction since the characters (nucleotides) of recombinant sequences do not share a single phylogenetic history. Both groups (Jiggins, *et al.*, 2001b; Werren and Bartos, 2001) have suggested that fine-scale phylogenies of *wsp* sequences should be treated with caution. Although we agree with these authors' conclusions, the *wsp* sequences that will be presented here are fully concordant with the *ftsZ* results and the conclusions drawn in our study (see results and discussion) do not rely heavily on an accurate phylogenetic reconstruction of the *wsp* fragment.

Wolbachia-infected individuals were sequenced for *wsp* using the 81F and 691R primers following methods previously described (Zhou, *et al.*, 1998; Rokas, *et al.*, 2001a, Chapter 4, Appendix I). For *ftsZ* sequencing, PCR products were cloned using the TOPO TA Cloning kit (Invitrogen, cat. 4500-01) and 4-6 clones from each

specimen (10-15 for those specimens that had more than one *wsp* product) were subsequently sequenced. Plasmids containing the fragment of interest were isolated using the QIAprep Spin Miniprep Kit (QIAGEN, cat. 27104). Plasmid DNA was subsequently quantified and sequenced. Sequencing was carried out using the Perkin-Elmer BigDye Terminator chemistry and an ABI 377 sequencer. Sequencing reactions were performed in both directions to minimise PCR artefacts, ambiguities and base-calling errors.

8.3.4 Analysis of Sequence Data

For both datasets, representative sequences of all known *Wolbachia* strains were included. Sequences were aligned using CLUSTALW (Thompson, *et al.*, 1994) with the default settings. The *ftsZ* fragment was 687 base pairs (bp) long for all the species examined and contained no gaps. The *wsp* fragment varied in length between 540 and 594 bp, with the aligned dataset being 614 bp long. For the *wsp* dataset, nucleotide positions 74-109, 217-255 and 519-585 were excluded from the analysis, due to ambiguity in the identification of positional nucleotide homology in these regions. All the sequences will be available from Genbank on publication. The two alignments and their corresponding phylogenies will be are available electronically from TreeBASE (http://www.herbaria.harvard.edu/treebase/) on publication.

Phylogenies for *wsp* and *ftsZ* were estimated by neighbour-joining (NJ) analysis of the sequence data, using the phylogenetic analysis package PAUP*, version 4.0b8 (Swofford, 2000). Gaps were coded as missing. Distances were calculated using maximum likelihood (ML) and the best-fit ML model for each locus identified using MODELTEST 3.0 (Posada and Crandall, 1998). MODELTEST utilises likelihood ratio tests (Huelsenbeck and Rannala, 1997; Lewis, 1998) to identify the ML model of sequence evolution (Posada and Crandall, 1998). The models with the best-fit for the two datasets were rather similar. For *wsp*, the best-fit model included unequal base frequencies, a transition/transversion ratio of 3.29 and rate heterogeneity among sites (gamma distribution shape parameter α=0.3548). The best-fit model for *ftsZ* included unequal base frequencies, a different rate between transitions and transitions (but with the two types of transitions, A↔G and T↔C,

showing ratios of 7.41 and 18.7 respectively, relative to transversions) and rate heterogeneity among sites (gamma distribution shape parameter α=0.13). The parameter values suggested by MODELTEST for each dataset were used to specify the distance matrix for neighbour-joining analysis on 1000 bootstrap replicates.

8.3.5 Analysis of infection patterns within and among gallwasp lineages

For the statistical analysis of infection patterns we have assumed that gallwasp species are acquiring infections of *Wolbachia* independently of each other (rather than through common ancestry). This assumption is justified by current knowledge on transmission of *Wolbachia* between different arthropod hosts (Werren, *et al.*, 1995a; Schilthuizen and Stouthamer, 1997; Huigens, *et al.*, 2000) and has been employed in many published studies (e.g. Wenseleers, *et al.*, 1998; Kittayapong, *et al.*, 2000; Werren and Windsor, 2000).

8.4 Results

8.4.1 The distribution of Wolbachia *infection in the Cynipidae*

Eleven of the 64 species screened (17.2%) were found to be infected with *Wolbachia* (these are: in the tribe Synergini *C. cerris*, *S. crassicornis*, *S. diaphanus*, *S. gallaepomiformis*, *S. reinhardi*, *S. umbraculus*; in the tribe Cynipini *A. solitarius*, *B. pallida*, *N. macropterus*, *C. glandium*, *P. quercusilicis*). Three different *wsp* sequences (two of them differed only by a 3 bp indel) were isolated from *A. solitarius* (but only one *ftsZ* sequence). Given that the variability of the *ftsZ* gene in the group A *Wolbachia* clade is very low, it is probable that *A. solitarius* is infected by two or possibly three *Wolbachia* group A strains. All the other ten gallwasp species were infected with a single strain. All species harboured group A *Wolbachia*, with the exception of *C. cerris* which was infected with a group B *Wolbachia*. This prevalence of group A *Wolbachia* in gallwasps is in agreement with previous surveys showing that Hymenoptera are more frequently infected with group A *Wolbachia*

(West, *et al.*, 1998; Werren and Windsor, 2000). Seven of the eleven infected species shared the presence of *Wolbachia* in all samples screened. The remaining four species are discussed below.

8.4.2 Intra-specific variation in Wolbachia *infection*

For 18 out of the 64 gallwasp species screened in this study, many individuals from different localities were examined for *Wolbachia* infection. 13 of these species were not infected with *Wolbachia*. Of the five infected species only one (*N. macropterus*) was found to be infected for all individuals and locations. The other four species show varying levels of infection (see Appendix 8.1). In *B. pallida* (85.4%) and *S. reinhardi* (48%) the variation is among different collection sites (i.e. presence or absence of *Wolbachia* infection is geographically localised). *B. pallida* is not infected in central Spain, but is infected in south and north Spain and in the rest of Europe. *S. reinhardi* is infected in Hungary and Austria and uninfected in France and Italy. For the other two species the data are inconclusive in regard to variation among sites, although they show variation within sites. *S. gallaepomiformis* (85.3%) is infected throughout Europe - only a few individuals in certain Hungarian sites are not infected, while *S. umbraculus* (16.7%) infected individuals may be found in Hungary, France and the UK but not in Spain.

8.4.3 Wolbachia *infection and lifecycle variation within the Cynipidae*

8.4.3.1 Patterns in the CP lineage

The limited distribution of *Wolbachia* in the CP lineage shows that this symbiont can play no general role in the alternation between thelytoky and arrhenotoky in the CP lifecycle. Furthermore, infection with *Wolbachia* is also not correlated with presence of only a known thelytokous generation in this lineage. Only two out of the apparently twenty thelytokous species in the tribe Cynipini are infected with *Wolbachia* (see Table 8.1); this ratio does not differ significantly from the ratio (3/31) of *Wolbachia* infection observed in cyclical parthenogenetic species (χ^2=0.20, d.f.=1, p>0.05).

Table 8.1. Numerical ratio and percentage of gallwasp species infected with *Wolbachia* for the various tribes and reproductive modes of Cynipidae. The ancestral reproductive mode in each tribe is indicated in parentheses (see Appendix 8.1 for abbreviations).

Tribe / Reproductive mode	***Wolbachia*-infected species / Total number of species**	**% of *Wolbachia*-infected / Total**	**Study**
Cynipini (CP)	5/53	9.4%	This study
'Aylacini' (AT)	4/9	44.4%	Plantard, *et al.*, 1999; this study
Synergini (AT)	6/10	60.0%	This study
Diplolepidini (AT)	11/19	57.9%	Plantard, *et al.*, 1999
Arrhenotokous	21/38	55.3%	Plantard, *et al.*, 1999; this study
Cyclical parthenogenetic	3/31	9.7%	This study
Parthenogenetic	2/20	10.0%	This study

8.4.3.2 Patterns across gallwasp tribes

Infection with *Wolbachia* was unevenly spread between lineages with different reproductive modes; arrhenotokous lineages (combined) are more likely to be infected than the cyclical parthenogen lineage (χ^2=20.58, d.f.=1, p<0.001) (Table 8.1). There was no significant difference among the arrhenotokous tribes ('Aylacini' *versus* Synergini, χ^2=0.05, d.f.=1, p>0.05 / 'Aylacini' *versus* Diplolepidini, χ^2=0.07, d.f.=1, p>0.05 / Synergini *versus* Diplolepidini, χ^2=0.08, d.f.=1, p>0.05).

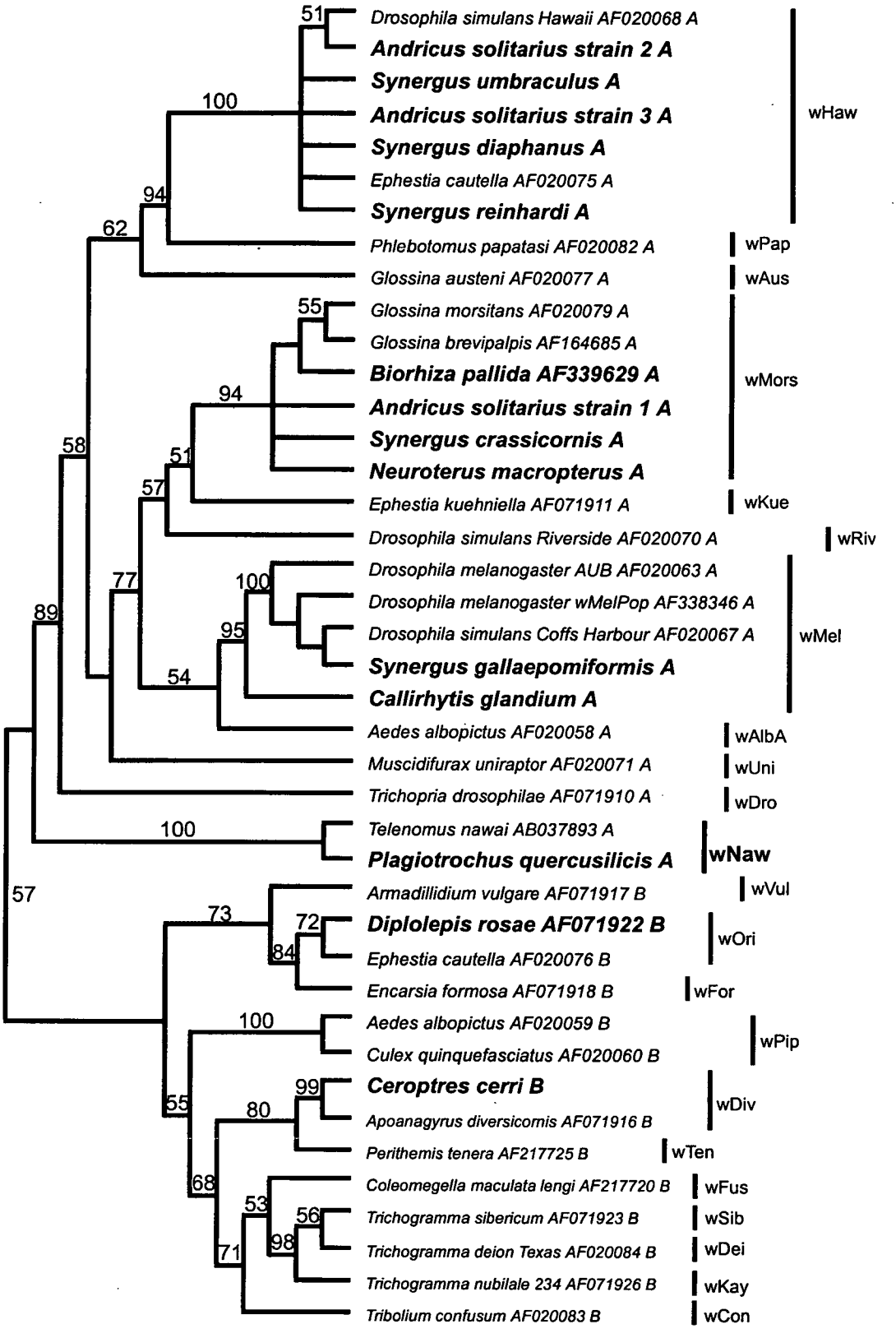
8.4.4 Molecular phylogenetic analysis of two Wolbachia *genes (*ftsZ *and* wsp*)*

The phylogenetic tree for *ftsZ*, in accordance with previous studies (e.g. Werren, *et al.*, 1995a), was unresolved within group A *Wolbachia* strains, which includes the majority of strains infecting gallwasps (tree not shown). Five out of the eleven infected gallwasp species have identical *ftsZ* sequences (*S. diaphanus*, *S. reinhardi*, *S. umbraculus*, *N. macropterus* and *A. solitarius*), with two more differing only in one nucleotide position (*S. crassicornis* and *S. gallaepomiformis*).

In contrast, the *wsp* tree is very structured (Fig. 8.2) and provides a much clearer picture of the different *Wolbachia* strains infecting gallwasps. The majority of infected gallwasp species are infected with group A *Wolbachia* (as is also the case for *ftsZ*) and due to the high variability of *wsp* a more precise distinction between the various strains of group A *Wolbachia* infecting gallwasps is possible. Two groups of identical sequences can be identified. Four gallwasp species (*S. reinhardi*, *S. diaphanus*, *S. umbraculus* and *A. solitarius* strain 3) have identical *wsp* sequences and together with *A. solitarius* strain 2 belong to the wHaw strain. Three other species (*S. crassicornis*, *A. solitarius* strain 1 and *N. macropterus*) also have identical sequences and together with *Biorhiza pallida* belong to the wMors strain. *S. gallaepomiformis* and *C. glandium* belong to the wMel strain and *C. cerris* to the group B *Wolbachia* wDiv strain. Species from different cynipid tribes (Cynipini and Synergini) are commonly infected with the same strain of *Wolbachia* (strains wHaw, wMors and wMel).

An interesting case is the *Wolbachia wsp* sequence found in *P. quercusilicis*, which is very similar to the one carried by the hymenopteran parasitoid *Teleonomus nawai* (1.11% distance) (Arakaki, *et al.*, 2000) and belongs to group A *Wolbachia*. The *Plagiotrochus* - *Teleonomus* clade is very divergent from all the other sequences in this dataset (distances range between 11% and 31%) and its placement, in our phylogenetic reconstructions, appears unstable. In most reconstructions the clade appears either basal to the group A *Wolbachia* or unresolved (a polytomy is created between the *Plagiotrochus* - *Teleonomus* clade, group A and group B *Wolbachia* clades). Less frequently, the clade is placed as a sister group of the wOri strain in

group B. The placement of the *Plagiotrochus* - *Teleonomus* clade (especially when the clade appears within group B *Wolbachia*) is possibly an artefact of the fast rate of evolution of *wsp* resulting in error in phylogenetic inference (long branch attraction). Irrespective of the exact phylogenetic placement of this clade, the *Plagiotrochus* - *Teleonomus wsp* sequences clearly represent a new strain of *Wolbachia* which we name wNaw. Our conclusion is based on the group threshold of 2.5% variation established by Zhou *et al.* (1998) (and utilised in subsequent studies (van Meer, *et al.*, 1999; Jeyaprakash and Hoy, 2000)) and the naming on the nomenclature followed there-in.



gure 8.2. The phylogenetic relationships between the different *Wolbachia wsp* sequences as reconstructed ing a neighbour-joining algorithm on a maximum likelihood estimated distance matrix. Name of the host thropod species is followed by Genbank accession number and group designation. The *Wolbachia* strains to hich sequences belong to are also shown. The tree shown is rooted on the group A - group B split for easy sualisation. Values above branches indicate bootstrap support (all compatible groupings with bootstrap values low 50% are shown, but without the bootstrap values). Sequences isolated from gallwasp species and the oposed new *Wolbachia* strain are shown in larger font and bold.

8.5 Discussion

8.5.1 Prevalence of Wolbachia *infection within species*

If *Wolbachia* infects only a small percentage of a population (or species), screening a few individuals may lead to an underestimation of the real percentage of species in which it occurs (Werren, *et al.*, 1995b; Jiggins, *et al.*, 2001a). In our analysis we have included rare species represented by only one or a few specimens and therefore, our result that *Wolbachia* infection is low in the CP lineage and high in the arrhenotokous lineages only provides a lower limit estimate of infection levels. The prevalence of *Wolbachia* in a given taxon may be correlated with its phenotypic effects on the taxon (Jiggins, *et al.*, 2001a) and the severity of conflict between the bacteria and the host. For example, Jiggins *et al.* (2001a) recently reported, in a screen of *Acraea* butterflies, that only 15% of each infected species harboured *Wolbachia*, the main reason being that the phenotype of these *Wolbachia* infections is male killing (which leads to sex-ratio distortion). Male killing bacteria (not only of the genus *Wolbachia*) typically occur in 1 - 30% of females of an infected taxon (Hurst and Jiggins, 2000) and are not expected to be fixed in a population, since populations consisting entirely of sexual females are driven to extinction. In contrast, strains that cause cytoplasmic incompatibility are typically fixed (or near fixation) in a population (and are expected to be so Turelli, 1994). Strains responsible for thelytoky and feminisation show a broader range of values (e.g. Stouthamer, 1997; Bouchon, *et al.*, 1998; Plantard, *et al.*, 1999). Although we do not know what phenotypic effects *Wolbachia* infection has in certain gallwasp tribes (see below), in species for which we screened many individuals from a wide geographical distribution few showed intra-specific variation, with 9/11 species having infection levels higher than 75% (see Appendix 8.1). These results suggest that although infected species are rare, once infected they show high prevalence of *Wolbachia*. In turn, this suggests that screening of many individuals from a given species probably did not have a serious effect on our *Wolbachia* infection estimates.

8.5.2 Wolbachia *is not associated with apparent thelytoky in the CP lineage*

Our data show that *Wolbachia* can have no role either in the alternation between arrhenotoky and thelytoky in the CP species, or in apparent loss of the arrhenotokous generation in the CP lineage. As described in the introduction, population genetic analyses suggest that at least some of the thelytokous species in the CP are actually cyclical parthenogens with an undiscovered arrhenotokous generation (Atkinson, *et al.*, 2001). Nevertheless, purely thelytokous species in the CP lineage are known (Folliot, 1964; Abe, 1986). It is notable that lifecycle simplification in these cases does not involve loss of sex in an arrhenotokous generation, but loss of a whole generation from an effectively bivoltine (i.e. two-generation) lifecycle. This pattern also argues against a role for *Wolbachia*, and the true cause of thelytoky in the CP lineage remains unknown.

So far, the parthenogenesis-inducing phenotype of *Wolbachia* is restricted to two haplodiploid insect orders; Hymenoptera and Thysanoptera (Stouthamer, 1997; Arakaki, *et al.*, 2001). Within the Hymenoptera, the distribution of the parthenogenesis-induction phenotype is not widespread, but limited to two parasitoid wasp super-families, the Chalcidoidea and the Cynipoidea (Stouthamer, 1997). The data presented here suggest that *Wolbachia*-induced parthenogenesis may be restricted even within these two taxa.

This lack of association between thelytoky and *Wolbachia* in the CP lineage is in contrast with Plantard *et al.* 's (1999) study on two other cynipid tribes Diplolepidini and 'Aylacini' (both with ancestral arrhenotokous lifecycles), where infected species are always secondarily thelytokous. Lack of correlation between a thelytokous lifecycle and *Wolbachia* infection has also been reported in other thelytokous hymenopterans (one bee species and 6 species of ant) (Wenseleers and Billen, 2000).

With the exception of the thelytoky phenotype in the Diplolepidini and 'Aylacini' (Plantard, *et al.*, 1998a; Plantard, *et al.*, 1999), little is known of the phenotypic effect(s) of *Wolbachia* infection in the other gallwasp tribes. Infected species in the tribe Synergini are expected to be thelytokous (as is the case for

infected species in all arrhenotokous gallwasp tribes). All of the infected *Synergus* species are known to possess males (Nieves-Aldrey, 1985; Nieves-Aldrey and Pujade-Villar, 1985), and it is probable that such species show geographical variation in reversal to thelytoky, as demonstrated for the rose gallwasp *Diplolepis spinosissimae* by Plantard *et al.* (1998a). The slow lifecycle of gallwasps and the obligate development of the larva within oak tissues make the mating and curing experiments needed to examine the wide range of *Wolbachia*-associated effects (such as for example, cytoplasmic incompatibility) extremely difficult (Plantard, *et al.*, 1998a; Plantard, *et al.*, 1999; Rokas, *et al.*, 2001a, Chapter 4, Appendix I).

8.5.3 Variation in prevalence of Wolbachia *among lineages*

In this study we show that gallwasps in the arrhenotokous lineages are much more frequently infected with *Wolbachia* than gallwasps in the cyclical parthenogenetic lineage (55.3% *versus* 12.0% - Table 8.1). Why? In general, differences in the prevalence of *Wolbachia* infection may be the end product of either (i) higher occurrence of *Wolbachia* infection in certain lineages than in others; and/or (ii) higher *Wolbachia* persistence levels in certain lineages than in others (see also Wenseleers, *et al.*, 1998). More specifically, the infection patterns observed in gallwasps may be explained either by a differential ability of *Wolbachia* in spreading into lineages with different reproductive modes or, alternatively, *Wolbachia* persistence may be selectively favoured in the arrhenotokous lineages (these scenarios are not mutually exclusive). Unfortunately experimental discrimination between these alternatives is very difficult. Levels of *Wolbachia* infection show high variation between different taxonomic groups and, occasionally, high variation within the same group (Table 8.2). For example, in ants there is a weak association between *Wolbachia* prevalence and mode of colony foundation (see Table 8.2) (Wenseleers, *et al.*, 1998). The basis for this correlation (as for gallwasps) remains unknown.

Table 8.2. Per cent prevalence of *Wolbachia* infection in various lineages. The group studied, its taxonomic status, the % of *Wolbachia* infection as well as the source study are indicated.

Group	Taxonomic status	% Prevalence of *Wolbachia*	Study
Fruit-flies (*Drosophila*)	Genus	8/48 - 16.7%	Bourtzis, *et al.*, 1996
Stalk-eyed flies (Diptera: Diopsidae)	Family	4/17 - 23.5%	Hariri, *et al.*, 1998
Ants (Hymenoptera: Formicidae)	Family	Whole family: 25/50 - 50.0% Dependently-founding species: 11/22 - 50% Independently-founding species: 1/8 - 13%	Wenseleers, *et al.*, 1998
Aphids (Hemiptera: Aphidoidea) Parasitoids and hyperparasitoids (Hymenoptera) Leaf-mining moths (Lepidoptera: Gracillariidae) Butterflies (Lepidoptera)	Various	Aphids: 0/4 - 0.0% Aphid parasitoids: 0/19 - 0.0% Aphid hyperparasitoids: 1/8 - 12.5% Leaf-miners: 8/21 - 38.1% Leaf-miner parasitoids: 5/18 - 27.8% Butterflies: 4/13 - 30.8%	West, *et al.*, 1998
Predatory (Acari: Phytoseiidae) and spider mites (Acari: Tetranychidae)	2 different families	Both families: 10/27 - 37% Spider mites: 6/16 - 37.5% Predatory mites: 4/11 - 36.4%	Breeuwer and Jacobs, 1996
Mosquitoes (Diptera: Culicidae)	13 different genera	25/89 - 28.1% - Frequency of infected species differed significantly among genera	Kittayapong, *et al.*, 2000

Crustaceans (Arthropoda: Crustacea)	5 different orders	Isopoda: 22/63 - 34.9% Amphipoda: 0/12 - 0% Tanaidacea: 0/1 - 0% Cumacea: 0/3 - 0% Decapoda: 0/4 - 0%	Bouchon, *et al.*, 1998
Rose gallwasps	Tribe	11/19 - 57.9%	Plantard, *et al.*, 1999
Oak gallwasps	Tribe	5/53 - 9.4%	This study

8.5.4 Horizontal transfer of Wolbachia between gall inducers and their associated inquilines

Ten species from the tribe Synergini were screened for *Wolbachia* infection in an attempt to understand whether there is evidence for horizontal transmission of *Wolbachia* between inquilines and gallwasps. Inquiline species show some taxonomic specificity in the galls they attack. Members of the *Periclistus* genus attack galls from the Diplolepidini, whereas *Ceroptres* and *Synergus* attack Cynipini galls. In contrast, *Synophrus* is apparently a gall-inducer itself, although recent observations indicate that it may be a true, cryptic inquiline (Nieves-Aldrey, unpublished data). Based on the 100% similarity shown in their *wsp* sequences, two horizontal transfer cases may be identified in this dataset. The first case involves three inquiline species (*S. reinhardi*, *S. diaphanus*, *S. umbraculus*) and one gall-inducer (*A. solitarius* strain 3). The second case involves one inquiline (*S. crassicornis*) and two gall-inducers (*A. solitarius* strain 1 and *N. macropterus*). To date none of these inquiline species have been reared from galls of *A. solitarius*. Nevertheless, all four *Synergus* species are known to attack a wide range of host galls in the genus *Andricus*, and it is highly probable that they do attack *A. solitarius* (Nieves-Aldrey and Pujade-Villar, 1985; Nieves-Aldrey, 1988; Schönrogge, *et al.*, 1995). These cases could represent direct transfer of *Wolbachia* from one species to the other (as, for example, seen in Huigens, *et al.*, 2000) or might also involve other, as yet unidentified, non-gallwasp parasitoid species. While some inquilines develop in the inner cell of the gall-inducer (and so are in potentially intimate contact with it), others develop in different regions of the host gall, and are not in direct contact with the gall-inducer, perhaps reducing the potential for transfer of *Wolbachia*. *S. umbraculus* for example does inhabit the inner cell of the gall-inducer in *Andricus* galls (Schönrogge, *et al.*, 1995). Inner cell inquilines are often termed lethal inquilines (Stone, *et al.*, 2002) because their developing cells eventually crowd out and kill the gall-inducer. This mortality is dependent on the number of inquiline larvae, however, and it is possible, at least in principle, for both the gall-inducer and the inquiline to develop in close proximity. An alternative transmission route is by parasitoids, some of which attack both inquilines and gall-inducers in oak galls (e.g.

Schönrogge, *et al.*, 1995). The extent and pattern of *Wolbachia* infection within oak gall parasitoids remains unknown. Another interesting, as yet untested, alternative hypothesis is that the *Wolbachia* infection may be transmitted by the gall tissues themselves, which are in intimate contact with all gall occupants and richly vascularised (Stone, *et al.*, 2002).

Some of the best-known work on insect-based food webs an community structure concerns insect galls (Godfray, 1994) and this study is a first attempt to utilise this knowledge for the study of *Wolbachia* transmission (see also Schilthuizen and Stouthammer, 1998). Oak gall communities in particular hold much promise as a model system, given that the communities harboured in a single gall range from a single insect to hundreds of insects, of up to 20 species, on 5 trophic levels (Askew, 1984). Demonstrating the presence of *Wolbachia* bacteria in the gall-inducers as well as in their associated inquilines is a significant step towards elucidating the dynamics of *Wolbachia* infections in this insect community.

8.6 Acknowledgements

A version of this chapter has been submitted for publication: Rokas, A., R. J. Atkinson, J.-L. Nieves-Aldrey, S. A. West and G. N. Stone. Distribution and diversity of *Wolbachia* in gallwasps (Hymenoptera; Cynipidae).

Jose-Luis Nieves-Aldrey identified the inquiline species used in this study. Rachel Atkinson, Stuart West and Graham Stone made extensive comments on previous versions of this chapter. I acknowledge the important contribution of Olivier Plantard 's published work on *Wolbachia* in gallwasps.

Appendix 8.1. List of gallwasp species screened for *Wolbachia* infections. Countries and localities from which specimens were obtained are also listed. Numbers next to localities indicate the number of specimens screened from that locality. Ratio indicates the proportion of individuals of a given species that were infected with *Wolbachia*. The LC column states the lifecycle of a given species. CP: cyclical parthenogenesis, T: thelytoky, AT: arrhenotoky, ATg: only the arrhenotokous generation has been found (this lifecycle is present only in the CP lineage).

No.	Species	Country and Locality	Ratio	LC
Tribe Cynipini (oak gallwasps)				
1	*Andricus aestivalis*	**Hungary** - Érsekvadkert-3.	0/3	ATg
2	*Andricus amenti* (formerly known as *A. giraudianus*)	**Hungary** - Szentkút-1.	0/1	CP
3	*Andricus aries*	**Hungary** - Isaszeg-2.	0/2	T
4	*Andricus askewii*	**Turkey** - Çekerek-1, Tokat-1.	0/2	T
5	*Andricus caputmedusae*	**Hungary** - Balaton-5, Eger-2, Matráfüred-2, Várpalota-3, Sirok-2, Szoloske-1. **Italy** - Piedimonte Matese-2, Monte S. Angelo-3, Castagnola-2, Monte Vulture-2, Gildone-1, Gargáno-2. **Austria** - Vienna-2. **Turkey** – Egirdir-2, Küllüce-4, Suluova-1, Tokat-1, Kizezoglu-1.	0/38	CP

6	*Andricus conglomeratus*	**Hungary** - Matráfüred-3.	0/3	T
7	*Andricus conificus*	**Italy** - San Venanzo-1, Massa Marittima-1. **Hungary** - Várpalota-2.	0/4	T
8	*Andricus coronatus*	**Hungary** - Balaton-7, Szeghalom-2, Várpalota-2, Szentkút-1. **Italy** - Chiusi-2, Massa Marittima-2, San Venanzo-2, Bombiana-1, Greve-1, Monte S. Angelo-1. **Turkey** - Madenli-2, Küllüce-2, Suluova-2. **Greece** - Edessa-4, Pisoderi-2, Arnissa-1, Agras-1.	0/35	T
9	*Andricus coriarius*	**Hungary** - Matráfüred-5, Gödöllő -1, Tatabánya-1, Várpalota-1. **Italy** - Rieti-2, Molise-2, Valpiana-1, Castelli-1. **Turkey** - Gezende-2, Suluova-2, Tokat-1, Niksar-1, Küllüce-1, Beysehir-1. **Greece** - Edessa-2, Florina-1, Pisoderi-1. **Slovakia** - Plástovce-1. **Spain** - El Escorial-2, Llerida-1.	0/30	CP
10	*Andricus corruptrix*	**Hungary** - Gödöllő-5.	0/5	CP
11	*Andricus crispator*	**Hungary** - Kőszeg-2.	0/2	ATg
12	*Andricus curvator*	**Spain** - Madrid-4. **U.K.** - Walsingham-2. **Hungary** - Miskolc-1.	0/7	CP
13	*Andricus dentimitratus* (formerly known as *A. viscosus*)	**Spain** - Barcelona-5-3. **Slovenia** - ?-2. **Hungary** - Kőszeg-4, Devecser-1. **Italy** - Chiusi-2, Gargáno-5, Massa Marittima-2, Siena-1. **Turkey** - Beybesli-6.	0/28	CP
14	*Andricus fecundator*	**U.K.** - Leeds-1.	0/1	CP
15	*Andricus gallaetinctoriae*	**Hungary** - Miskolc-5. **Italy** - Rocca di Ruffeno-2. **Turkey** - Madenli-2.	0/9	CP
16	*Andricus glutinosus*	**Hungary** - Szarnalum-5, Eger-1. **Austria** - Vienna-2.	0/8	T

17	*Andricus grossulariae*	**Spain** - Piedralaves-2. **Hungary** - Devecser-3.	0/5	ATg
18	*Andricus hartigi*	**Italy** - Lame-1, Casina-1, Massa Marittima-2, Greve-1. **Hungary** - Sirok-1.	0/6	T
19	*Andricus hungaricus*	**Hungary** - Karcag-4. **Austria** - Vienna-2.	0/6	T
20	*Andricus inflator*	**Finland** - Turku-1.	0/1	CP
21	*Andricus kollari*	**Hungary** - Hortobágy-5, Matráfüred-2.	0/7	CP
22	*Andricus lignicolus*	**Hungary** - Gödöllő-5. **U.K.** - Hampstead Heath-2.	0/7	CP
23	*Andricus lucidus*	**Hungary** - Matráfüred-5, Isaszeg-2, Szob-2, Jászberény-1, Balaton-1, Sopron-1. **Italy** - Casina-1, Lame-1, Piedimonte Matese-1, Gargáno-1, Chiusi-1, Radicofani-1, Vernasca-1. **Czech** - Valtice-1. **France** - Nantes-2. **Turkey** - Antalya-5, Egirdir-3, Beysehir-3, Aglasun-1, Suluova-1.	0/35	CP
24	*Andricus malpighii* (formerly known as *A. nudus*)	**Italy** - Bibulano-1.	0/1	CP
25	*Andricus mayri*	**Hungary** - Szob-3.	0/3	T
26	*Andricus mediterraneae*	**Turkey** - Madenli-5.	0/5	T
27	*Andricus mitratus*	**Slovenia** - ?-5. **Hungary** - Matráfüred-1. **Turkey** - Beybesli-1.	0/7	T

28	*Andricus panteli*	**Turkey** - Madenli-5, Egirdir-4, Gezende-2, Aglasun-2. **Morocco** - Azrou-6. **Italy** - Greve-2, Chiusi-2, Poppi-2, Bombiana-1, Radicofani-1, San Venanzo-1, Monte S. Angelo-3.	0/31	CP
29	*Andricus polycerus*	**Hungary** - Isaszeg-5. **Italy** - Poppi-2.	0/7	T
30	*Andricus quercuscalicis*	**Hungary** - Tiszakürt-1. **Ireland** - Dublin-1. **France** - Crécy-1. **Germany** - Rottenbach-1.	0/4	CP
31	*Andricus quercusramuli*	**U.K.** - Oxford-5. **Hungary** - Matráfüred-2.	0/7	CP
32	*Andricus quercustozae*	**Hungary** - Sopron-5. **Italy** - Greve-2, Gildone-1, Felitto-1, Jelsi-1. **Morocco** - Azrou-2. **Turkey** - Gezende-2, Madenli-2, Küllüce-2, Refahiye-3, Yeniyol-1. **Spain** - Prado del Rey-1. **Greece** - Arnissa-3, Pisoderi-2.	0/28	CP
33	*Andricus seckendorfii*	**Italy** - Bra 95-5, Massa Marittima-3, Chianti-2, Chiusi-1. **Turkey** - Hadim-8, Madenli-2.	0/21	CP
34	*Andricus solitarius*	**France** - St. Chinian-3.	3/3	CP
35	*Andricus testaceipes*	**Hungary** - Sopron-2.	0/2	CP
36	*Andricus theophrastus*	**Turkey** - Beysehir-3.	0/3	T
37	*Andricus tinctoriusnostrus*	**Turkey** - Madenli-2. **Lebanon** - Shakra-1.	0/3	T
38	*Andricus tomentosus*	**Turkey** - Madenli-1. **Greece** - Arnissa-1.	0/2	T
39	*Aphelonyx cerricola*	**Italy** - ?-3, Piombino-2, Chianti-2, Rieti-1. **Hungary** - Várpalota-5, Kemence-1, Kőszeg-1. **Slovakia** - Cifáre-1. **Turkey** - Beysehir-3.	0/19	T

40	*Biorhiza pallida*	See Rokas *et al.* (2001) for countries and localities.	172/206	CP
41	*Callirhytis glandium*	**Hungary** - Gyöngös-2.	2/2	CP
42	*Chilaspis nitida*	**Hungary** - Unjuned-7.	0/7	CP
43	*Cynips cornifex*	**Hungary** - Matráfüred-2. **Italy** - Colombaro-1.	0/3	T
44	*Cynips disticha*	**Hungary** - Kemence-2. **France** - B. de Mervent-1.	0/3	CP
45	*Cynips divisa*	**Hungary** - Gödöllő-2, Nagygyanté-1, Sopron-1.	0/4	CP
46	*Cynips longiventris*	**Hungary** - Devecser-3, Pácin-2, Nagygjante-2. **Slovakia** - Ciáre-1, Sikenica-2. **Croatia** - Istria-1. **France** - Nantes-1, St. Porquier-1. **U.K.** - Hartsholme-1.	0/14	CP
47	*Cynips quercus*	**Hungary** - Eger-2, Veszprém-1, Szentkút-1, Sopron-1. **France** - Angoulême-1, Le Barp-1. **Italy** - Rieti-1, Poppi-1, Anconella-2, Bibulano-1, Pianavia-1, The Matese-1. **Greece** - Komnina-2, Arnissa-1, Pisoderi-1, Prespa-1, Stagira-1, Edessa-1. **Turkey** - Tefenni-1, Aglasun-1, Guzeloluk-1, N. of Antalya-1, Sekerek-1, Yeşilyurt-1, Suluova-1, Kirezoğlu-1, Refahiye-1, Küllüce-1.	0/31	CP
48	*Cynips quercusfolii*	**Hungary** - Mesterszállás-5, Matráfüred-1. **France** - Montereau-2. **U.K.** - Oxford-2. **Croatia** - Petrovina-1. **Greece** - Olympiada-1. **Turkey** - Black Sea-1.	0/13	CP
49	*Neuroterus lanuginosus*	**Hungary** - Sirok-3.	0/3	T
50	*Neuroterus macropterus*	**Hungary** - Matráfüred-4, Sopron-3, Gödöllő-2, Várpalota-3, Kemence-3. **Slovakia** - Plástovce-2. **Italy** - Castelletto-1. **Greece** - Arnissa-1.	19/19	T

51	*Neuroterus quercusbaccarum*	**Hungary** - Matráfüred-5. **U.K.** - Auchtermuchty-2.	0/7	CP
52	*Plagiotrochus quercusilicis*	**Spain** - El Pardo-3.	3/3	T
53	*Trigonaspis synaspis*	**Turkey** - Beysehir-3, Suluova-2.	0/5	CP

Tribe Synergini (inquilines)

54	*Ceroptres cerris*	**Hungary** - Matráfüred-2.	2/2	AT
55	*Periclistus brandti*	**Sweden** - Gottsunda-1.	0/1	AT
56	*Synergus crassicornis*	**Spain** - Casa del Campo, Madrid-3.	3/3	AT
57	*Synergus diaphanus*	**Hungary** - Sopron-1.	1/1	AT
58	*Synergus gallaepomiformis*	**Hungary** - Szentendre-2, Karcag-1, Balaton-1, Szentkút-1 (-), Sopron-2, Szeghalom-2 (+/-), Eger-1 (-),Matráfüred-1 (-),Veszprém-1 (-). **Slovakia** - Prasnik-1. **U.K.** - Elsfield-4, Cambridge-4, Birnwood Forest-4. **France** - Rennes-3, Forêt Decouvres-2, Montain-1, Nantes-1. **Spain** - Cercedilla-2.	29/34	AT
59	*Synergus hayeanus*	**Hungary** - Karcag-1, Gödöllő-2, Gyöngös-1, Veszprém-1.	0/5	AT
60	*Synergus incrassatus*	**Hungary** - Matráfüred-1.	0/1	AT
61	*Synergus reinhardi*	**Hungary** - Gödöllő-2 (+), Balaton-2 (+), Eger-1 (+), Felsőtárkány-1, Szentendre-1 (+), Veszprém-5 (+). **France** - Rennes-4, Nantes-5, Crécy-1, Castres-1. **Italy** - Fellizzano-1. **Austria** - Vienna-1 (+).	12/25	AT

62	*Synergus umbraculus*	**Hungary** - Matráfüred-7 (+/-). **France** - Guérande-1 (+). **U.K.** - London-1 (+). **Spain** - Cercedilla-9.	3/18	AT
63	*Synophrus politus*	**Hungary** - Gödöllő-1, Sopron-1, Szoloske-1, Tatabánya-1, Kemence-1. **Italy** - Lame-5, Greve-1, Piombino-1. **Greece** - Arnissa-4, Edessa-1. **Spain** - Madrid-2. **Turkey** - Madenli-4.	0/23	AT

Tribe Aylacini (herb gallwasps)

64	*Panteliella bicolor*	**Sweden** - Gottsunda-1.	0/1	AT

CHAPTER 9

Concluding remarks

The work presented in this thesis has focused on two main areas: (a) identification of useful molecular markers for phylogenetic reconstruction, and (b) application of molecular phylogenetics and 'tree-thinking' to analyse patterns of evolution in gallwasps. In this concluding section, I will briefly consider ideas for future work on these two areas and what interesting questions have arisen from the results described in the preceding chapters.

9.1 Mind the data; molecular markers and phylogenetics

Reconstructing the 'Tree of Life' or in Darwin' s words obtaining 'fairly true genealogical trees of each great kingdom of Nature' is one of the most important tasks for evolutionary biologists to fulfil in the years to come. Whether the final product of this exercise will be a nice and neatly arranged bifurcating 'tree' as Darwin dreamed of or a complex (and far from genealogical) 'bush' is difficult to assess from current data, but what' s obvious is that information encoded in molecules (DNA and amino acid sequences) has been instrumental in the first steps of its reconstruction. The tremendous technological advances made over the last few years in the field of molecular biology have offered yet another boost to research in molecular phylogenetics by offering new kinds of markers (rare genomic changes - RGCs, Chapter 2). Further testing of RGCs and of their power in phylogenetic applications, more accurate estimates regarding their homoplasy levels as well as statistical modelling of their evolution are three areas that have already started attracting more attention.

9.2 Oak gallwasps, fragmented landscapes and the ever-changing climate

In the last two and a half million years, Earth' s climate has undergone major changes, which lead to changes in the distribution (or in the extinction) of many plant and animal taxa. In Chapters 4 and 5 I analysed how these climate changes coupled with the nature of the European landscape have shaped the genetic diversity and pattern of distribution of two oak gallwasp species. Perhaps the most important conclusion drawn from these studies is the effect exerted by their hosts, the oaks, in determining distribution patterns. Studies on more oak gallwasp species, focusing both on groups of species sharing the same host preferences as well as on groups of species differing in their host specificity, will allow a better understanding of whether species' reactions to environmental disturbances are largely determined by their host specificity or whether these reactions are more species-specific (and perhaps more stochastic) than currently assumed.

9.3 Oak gallwasps and the New World

Although oak gallwasps have a Holarctic distribution, most of the work described in this thesis (with the small exception of four species used in Chapter 6) has focused on the European oak gallwasp fauna. Answering major questions regarding the origin, biogeography and more generally the evolution of the whole tribe will require a better understanding of the oak gallwasp fauna in the Nearctic as well as in the eastern Palaearctic. Central America harbours the highest diversity of oak species, but its oak gallwasp fauna has not been very well studied, despite having a much higher species richness compared to Europe; it is perhaps indicative of the high diversity of these regions to note that California alone has more gallwasp species than the whole of Europe.

Work on the Nearctic and eastern Palaearctic will allow us to understand whether the European oak gallwasp fauna is the result of a single radiation or whether there have been multiple independent radiations from the east, whether they happened alongside the diversification of oaks in these regions, or whether they are

much more recent products of evolution as well as how gall structure and host plant and host organ preferences have evolved in the oak floras of the two regions. For example, are gall structure, host organ and host plant specificity conserved features in the American world of oak cynipids as evidence suggests for European taxa or not? Unfortunately, the taxonomic status of many oak gallwasp taxa in these regions (especially in the Americas) is not very robust and serious taxonomic revisions may be required before molecular phylogenetic analysis and reconstruction of evolutionary patterns can commence. It is an encouraging sign that considerable progress in currently being made.

9.4 Oak gallwasps and their micro-community

Oak galls harbour a spectacular micro-community of inquiline gallwasps and insect parasitoids. The high levels of diversity observed and the relative ease of their sampling makes communities associated with gallwasps attractive models for addressing questions on the evolution of insect communities in general. For example, the work described in Chapters 4 and 5 on *Biorhiza pallida* and *Andricus quercustozae* respectively, could be taken further by examining, firstly, whether the community associated with the gall-inducer is the same throughout the Pleistocene refuge areas and secondly, whether invasion of the gall-inducer into central and northern Europe was associated with acquisition of new inquilines and parasitoids or whether inquilines and parasitoids have tracked the gall-inducer on its range expansion. Additionally, one could examine whether the oak cynipid gall structures are correlated with the load of parasitism on the given gall and whether novelties in gall structure are avoiding parasite load or whether they represent 'evolutionary relicts'. The micro-community associated with oak gallwasps may also offer an interesting model for investigating the spread of bacterial endosymbionts, such as *Wolbachia* (Chapter 8).

9.5 Gallwasps and microbes: beyond *Wolbachia*

Wolbachia' s cunning ways of spreading into insect populations on the one hand (such as induction of parthenogenesis in haplodiploid taxa, mating incompatibility between infected and uninfected individuals, feminisation of male individuals in isopods or the killing of males seen in butterflies and beetles) and its emerging importance in the pathogenesis of important filarial human diseases on the other (*Wolbachia'* s relation with the nematode host may be more adequately described as symbiosis, since elimination of *Wolbachia* has detrimental consequences for its nematode host – this is not the case for insect hosts) pose interesting evolutionary questions and have rightly attracted the attention of many scientists. However, *Wolbachia* is just one of the many bacterial endosymbionts that may be found within insects and more generally invertebrate species. More and more phenotypes previously thought to be exclusively associated with *Wolbachia* (e.g. male-killing, parthenogenesis) are now being found to be induced by other endosymbionts as well. Although gallwasps may be an attractive system for understanding how *Wolbachia*, and more generally endosymbionts cross species barriers (Chapter 8), the limited number of generations per year and its obligate association with its oak host make the more elaborate manipulation and crossing experiments required to demonstrate effects associated with endosymbiont presence or absence difficult to execute.

9.6 Conclusion

'What has been achieved is but the first step; we still stand in the presence of riddles. But riddles with the hope of solution, what more can a scientist desire?'

Thus spoke Hans Spemann back in the 1930s about the progress in the understanding of developmental processes. Perhaps the same argument can be made regarding a much humbler task, understanding the evolution of oak gallwasps. Powerful markers in the form of DNA sequences and allozyme polymorphisms have been developed and / or standardised in the gallwasp system; preliminary results from the first few species suggest that they are likely to prove very powerful. Perhaps more

importantly, our understanding of the ecology, taxonomy and distribution of oak gallwasps and their associated community is ever increasing and more and more researchers from heterogeneous backgrounds are drawn into the 'black hole' of gallwasp research. It is only by such an active and diverse community of researchers that fascinating questions in gallwasp research (e.g. how galls are formed, the evolution of cyclical parthenogenesis) may one day be answered.

REFERENCES

Abe, Y. (1986). Taxonomic status of the *Andricus mukaigawae* complex and its speciation with geographical parthenogenesis (Hymenoptera, Cynipidae). *Appl. Entomol. Zool.* **21:** 436-447.

Abrahamson, W. G., Melika, G., Scrafford, R. and Csoka, G. (1998). Gall-inducing insects provide insights into plant systematic relationships. *Am. J. Bot.* **85:** 1159-1165.

Abrahamson, W. G. and Weis, A. E. (1997). 'Evolutionary ecology across three trophic levels: goldensrods, gallmakers and natural enemies,' Princeton University Press, Princeton, New Jersey.

Aguilar, J. M. and Boecklen, W. J. (1991). Patterns of herbivory in the *Quercus grisea* X *Quercus gambelii* species complex. *Oikos* **64:** 498-504.

Aguinaldo, A. M. A., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A. and Lake, J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387:** 489-493.

Akimoto, S. I. (1998). Heterogeneous selective pressures on egg-hatching time and the maintenance of its genetic variance in a *Tetraneura* gall-forming aphid. *Ecol. Entomol.* **23:** 229-237.

Ambrus, B. (1974). 'Cynipid Galls,' Academic Press, Budapest, Hungary.

Arakaki, N., Miyoshi, T. and Noda, H. (2001). *Wolbachia*-mediated parthenogenesis in the predatory thrips *Franklinothrips vespiformis* (Thysanoptera: Insecta). *Proc. R. Soc. London B* **268:** 1011-1016.

Arakaki, N., Noda, H. and Yamagishi, K. (2000). *Wolbachia*-induced parthenogenesis in the egg parasitoid *Telenomus nawai*. *Entomol. Exp. Appl.* **96:** 177-184.

Ascher, J. S., Danforth, B. N. and Ji, S. (2001). Phylogenetic utility of the major opsin in bees (Hymenoptera: Apoidea): a reassessment. *Mol. Phylog. Evol.* **19:** 76-93.

Askew, R. R. (1961). On the biology of the inhabitants of oak galls of Cynipidae in Britain. *Trans. R. Entomol. Soc.London* **14:** 237-268.

Askew, R. R. (1984). The biology of gall wasps. *In* 'The Biology of Galling Insects' (Ananthakrishnan, T. N., Ed. pp. 223-271, Oxford and IBH Publishing Co., New Delhi.

Atkinson, R. J. (2000). The genetic analysis of natural history, reproductive strategy and population structure in European oak gallwasps (Hymenoptera: Cynipidae), pp. 286. University of Oxford, Oxford.

Atkinson, R. J., McVean, G. A. T. and Stone, G. N. (2001). Use of population genetic data to infer oviposition behaviour: Species-specific patterns in four oak gallwasps (Hymenoptera: Cynipidae). *Proc. R. Soc. London B*: in press.

Averof, M. and Akam, M. (1995). Insect-crustacean relationships: insights from comparative developmental and molecular studies. *Philos. Trans. R. Soc. London B* **347:** 293-303.

Averof, M., Rokas, A., Wolfe, K. H. and Sharp, P. M. (2000). Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* **287:** 1283-1286.

Avise, J. C. (2000). 'Phylogeography: the history and formation of species,' Harvard University Press.

Avise, J. C. and Johns, G. C. (1999). Proposal for a standardised temporal scheme of biological classification for extant species. *Proc. Natl. Acad. Sci. U. S. A.* **96:** 7358-7363.

Balavoine, G. (1998). Are platyhelminthes coelomates without a coelom? An argument based on the evolution of *Hox* genes. *Am. Zool.* **38:** 843-858.

Ballard, J. W. O., Hatzidakis, J., Karr, T. L. and Kreitman, M. (1996). Reduced variation in *Drosophila simulans* mitochondrial DNA. *Genetics* **144:** 1519-1528.

Bayes, T. and Price, R. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Phil. Trans. R. Soc. London* **53:** 370-418.

Beckenbach, A. T., Wei, Y. W. and Liu, H. (1993). Relationships in the *Drosophila obscura* species group, inferred from mitochondrial cytochrome oxidase II sequences. *Mol. Biol. Evol.* **10:** 619-634.

Beebe, N. W., Ellis, J. T., Cooper, R. D. and Saul, A. (1999). DNA sequence analysis of the ribosomal DNA ITS2 region for the *Anopheles punctulatus* group of mosquitoes. *Insect Mol. Biol.* **8:** 381-390.

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N. and Bonhomme, F. (2000). 'GENETIX 4.02, logiciel sous Windows TM pour la génétique des populations,' Laboratoire Génome, Populations, Interactions, CNRS UPR 9060, Université de Montpellier II, Montpellier.

Belshaw, R. and Quicke, D. L. J. (1997). A molecular phylogeny of the Aphidiinae (Hymenoptera: Braconidae). *Mol. Phylog. Evol.* **7:** 281-293.

Bennett, K. D. (1986). The rate of spread and population increase of forest trees during the postglacial. *Philos. Trans. R. Soc. London B* **314:** 523-531.

Bensasson, D., Zhang, D.-X., Hartl, D. L. and Hewitt, G. M. (2001). Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol. Evol.* **16:** 314-321.

Bensasson, D., Zhang, D. X. and Hewitt, G. M. (2000). Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. *Mol. Biol. Evol.* **17:** 406-415.

Björklund, M. (1999). Are third positions really that bad? A test using vertebrate cytochrome b. *Cladistics* **15:** 191-197.

Boecklen, W. J. and Spellenberg, R. (1990). Structure of herbivore communities in 2 oak (*Quercus* spp) hybrid zones. *Oecologia* **85:** 92-100.

Boore, J. L. (1999). Animal mitochondrial genomes. *Nucleic Acids Res.* **27:** 1767-1780.

Boore, J. L., Daehler, L. L. and Brown, W. M. (1999). Complete sequence, gene arrangement, and genetic code of mitochondrial DNA of the cephalochordate *Branchiostoma floridae* (amphioxus). *Mol. Biol. Evol.* **16:** 410-418.

Boore, J. L., Lavrov, D. and Brown, W. M. (1998). Gene translocation links insects and crustaceans. *Nature* **393:** 667-668.

Bouchon, D., Rigaud, T. and Juchault, P. (1998). Evidence for widespread *Wolbachia* infection in isopod crustaceans: molecular identification and host feminization. *Proc. R. Soc. London B* **265:** 1081-1090.

Boursot, P., Auffray, J.-C., Britton-Davidian, J. and Bonhomme, F. (1993). The evolution of house mice. *Annu. Rev. Ecol. Syst.* **24:** 119-152.

Bourtzis, K., Nirgianaki, A., Markakis, G. and Savakis, C. (1996). *Wolbachia* infection and cytoplasmic incompatibility in *Drosophila* species. *Genetics* **144:** 1063-1073.

Bradley, R. D. and Hillis, D. M. (1997). Recombinant DNA sequences generated by PCR amplification. *Mol. Biol. Evol.* **14:** 592-593.

Brakefield, P. M. (2001). The structure of a character and the evolution of patterns. *In* 'The Character Concept in Evolutionary Biology' (Wagner, G. P., Ed. pp. 343-361, Academic Press.

Breeuwer, J. A. J. and Jacobs, G. (1996). *Wolbachia*: Intracellular manipulators of mite reproduction. *Exp. Appl. Acarol.* **20:** 421-434.

Bromham, L. D. and Degnan, B. M. (1999). Hemichordates and deuterostome evolution: robust molecular phylogenetic support for a hemichordate + echinoderm clade. *Evol. Dev.* **1:** 166-171.

Brooks, S. E. and Shorthouse, J. D. (1998). Developmental morphology of stem galls of *Diplolepis nodulosa* (Hymenoptera : Cynipidae) and those modified by the inquiline

Periclistus pirata (Hymenoptera : Cynipidae) on *Rosa blanda* (Rosaceae). *Can. J. Bot.* **76:** 365-381.

Brower, A. V. Z. (1994). Rapid morphological radiation and convergence among races of the butterfly *Heliconius erato* inferred from patterns of mitochondrial DNA evolution. *Proc. Natl. Acad. Sci. U. S. A.* **91:** 6491-6495.

Brower, A. V. Z. and DeSalle, R. (1994). Practical and theoretical considerations for choice of a DNA sequence region in insect molecular systematics, with a short review of published studies using nuclear gene regions. *Ann. Entomol. Soc. Am.* **87:** 702-716.

Brower, A. V. Z. and DeSalle, R. (1998). Patterns of mitochondrial versus nuclear DNA sequence divergence among nymphalid butterflies: the utility of *wingless* as a source of characters for phylogenetic inference. *Insect Mol. Biol.* **7:** 73-82.

Buckler IV, E. S., Ippolito, A. and Holtsford, T. P. (1997). The evolution of ribosomal DNA: Divergent paralogues and phylogenetic implications. *Genetics* **145:** 821-832.

Buhr, H. (1965). 'Bestimmungstabellen der Gallen (Zoo und Phytocecidien) an Pflanzen Mittel- und Nordeuropas,' V.E.B. Gustav Fischer Verlag, Jena, Germany.

Burger, W. G. (1975). The species concept in *Quercus*. *Taxon* **24:** 45-50.

Campbell, A., Mrazek, J. and Karlin, S. (1999). Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl. Acad. Sci. U. S. A.* **96:** 9184-9189.

Camus, A. (1936-1954). Les Chênes. Monographie du genre *Quercus* and monographie du genre *Lithocarpus*. *Encycl. Écon. de Sylvic.* Volumes **6, 7**.

Cariello, N. F., Thylli, W. G., Swenberg, J. A. and Skopek, T. R. (1991). Deletion mutagenesis during polymerase chain reaction: Dependence on DNA polymerase. *Gene* **99:** 105-108.

Castresana, J., Feldmaier-Fuchs, G., Yokobori, S., Satoh, N. and Paabo, S. (1998). The mitochondrial genome of the hemichordate *Balanoglossus carnosus* and the evolution of deuterostome mitochondria. *Genetics* **150:** 1115-1123.

Castroviejo, S. Ed. (1986). 'Flora Iberica: Plantas vasculatres de la Península Ibérica el Islas Baleares,' CSIC, Madrid.

Caterino, M. S., Cho, C. and Sperling, F. A. H. (2000). The current state of insect molecular systematics: a thriving tower of Babel. *Annu. Rev. Entomol.* **45:** 1-54.

Cavalli-Sforza, L. L. and Edwards, A. W. F. (1967). Phylogenetic analysis: models and estimation procedures. *Evolution* **32:** 550-570.

Chippindale, P. T., Dave, V. K., Whitmore, D. H. and Robinson, J. V. (1999). Phylogenetic relationships of north American damselflies of the genus *Ischnura* (Odonata: Zygoptera: Coenagrionidae) based on sequences of three mitochondrial genes. *Mol. Phylog. Evol.* **11:** 110-121.

Cho, S., Mitchell, A., Regier, J. C., Mitter, C., Poole, R. W., Friedlander, T. P. and Zhao, S. (1995). A highly conserved nuclear gene for low-level phylogenetics: Elongation factor-1 a recovers morphology-based tree for heliothine moths. *Mol. Biol. Evol.* **12:** 650-656.

Chodjai, M. (1980). Study of the gall forming cynipid Hymenoptera of Iranian oak forests. *J. Entomol. Soc. Iran.*: Suppl. 3.

Clarke, T. E. (1998). Molecular reevaluation of the *Nebria gregaria* infragroup and the implications for the existence of an ice age refugium on the Queen Charlotte Islands, University of Victoria, British Columbia, Canada.

Cline, J., Braman, J. C. and Hogrefe, H. H. (1996). PCR fidelity of *Pfu* DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res.* **24:** 3546-3551.

Cook, J. M., Rokas, A., Pagel, M. and Stone, G. N. (2001). Evolutionary shifts between host oak species and host plant organs in *Andricus* gallwasps. *Evolution*: submitted.

Cook, J. M., Stone, G. N. and Rowe, A. (1998). Patterns in the evolution of gall structure and life cycles in oak gallwasps (Hymenoptera: Cynipidae). *In* 'The Biology of Gall-Inducing Arthropods' (Csoka, G., Mattson, W. J., Stone, G. N. and Price, P. W., Eds.), pp. 261-279, U.S. Department of Agriculture, Forest Service, North Central Research Station, St. Paul, Minnesota.

Cooper, S. J. B., Ibrahim, K. M. and Hewitt, G. M. (1995). Postglacial expansion and genome subdivision in the European grasshopper *Chorthippus parallelus*. *Mol. Ecol.* **4:** 49-60.

Cornell, H. V. (1983). The secondary chemistry and complex morphology of galls formed by the Cynipinae (Hymenoptera) - Why and how. *Am. Midland Nat.* **110:** 225-234.

Cornell, H. V. (1985). Local and regional richness of cynipine gallwasps on California oaks. *Ecology* **66:** 1247-1260.

Cornell, H. V. (1996). Oak species attributes and host size influence cynipine was species richness. *Ecology* **76:** 1582-1592.

Cornell, H. V. and Washburn, J. O. (1979). Evolution of the richness area correlation for cynipid gallwasps on oak trees. A comparison of two geographic areas. *Evolution* **33:** 257-274.

Crease, T. J. and Taylor, D. J. (1998). The origin and evolution of variable-region helices in V4 and V7 of the small-subunit ribosomal RNA of branchiopod crustaceans. *Mol. Biol. Evol.* **15:** 1430-1446.

Crespi, B. and Worobey, M. (1998). Comparative analysis of gall morphology in australian gall thrips: the evolution of extended phenotypes. *Evolution* **52:** 1686-1696.

Crespi, B. J., Carmean, D. A. and Chapman, T. W. (1997). Ecology and evolution of galling thrips and their allies. *Annu. Rev. Entomol.* **42:** 51-71.

Crespi, B. J., Carmean, D. A., Mound, L. A., Worobey, M. and Morris, D. (1998). Phylogenetics of social behavior in Australian gall-forming thrips: evidence from mitochondrial DNA sequence, adult morphology and behavior, and gall morphology. *Mol. Phylog. Evol.* **9:** 163-180.

Crozier, R. H. and Crozier, Y. C. (1993). The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organisation. *Genetics* **133:** 97-117.

Crozier, R. H., Crozier, Y. C. and Mackinlay, A. G. (1989). The CO-I and CO-II region of honeybee mitochondrial DNA - evidence for variation in insect mitochondrial evolutionary rates. *Mol. Biol. Evol.* **6:** 399-411.

Cruickshank, R. H., Johnson, K. P., Smith, V. S., Adams, R. J., Clayton, D. H. and Page, R. D. M. (2001). Phylogenetic analysis of partial sequences of elongation factor 1 alpha identifies major groups of lice (Insecta : Phthiraptera). *Mol. Phylog. Evol.* **19:** 202-215.

Csóka, G. (1997). 'Plant galls,' Forest Research Institute, Budapest.

Csóka, G., Stone, G., Atkinson, R. and Schönrogge, K. (1998). The population genetics of postglacial invasions of northern Europe by cynipid gall wasps (Hymenoptera: Cynipidae). *In* 'The Biology of Gall-Inducing Arthropods' (Csóka, G., Mattson, W. J., Stone, G. N. and Price, P. W., Eds.), pp. 280-294, U.S. Department of Agriculture, Forest Service, North Central Research Station, St. Paul, Minnesota.

Cummings, M. P., Otto, S. P. and Wakeley, J. (1995). Sampling properties of DNA sequence data in phylogenetic analysis. *Mol. Biol. Evol.* **12:** 814-822.

Dalla Torre, K. W. and Kieffer, J. J. (1910). 'Cynipidae,' Friedlander & Sohn, Berlin.

Danforth, B. N. and Ji, S. (1998). Elongation factor-1a occurs as two copies in bees: implications for phylogenetic analysis of EF-1a sequences in insects. *Mol. Biol. Evol.* **15:** 225-235.

Darwin, C. (1859). 'On the Origin of Species,' John Murray, London.

de Rosa, R., Grenier, J. K., Andreeva, T., Cook, C. E., Adoutte, A., Akam, M., Carroll, S. B. and Balavoine, G. (1999). *Hox* genes in brachiopods and priapulids and protostome evolution. *Nature* **399:** 772-776.

Deschamps-Cottin, M., Aubert, J., Barascud, B. and Descimon, H. (2000). Hybridization and introgression between full-fledged species. The case of *Parnassius apollo* and *P. phoebus* (in French). *C. R. Acad. Sci. III – Sci. Vie* **323:** 327-337.

Dodds, K. S. (1939). Oogenesis in *Neuroterus baccarum* L. *Genetica* **21:** 177-193.

Doncaster, L. (1910). Gametogenesis of the gall-fly *Neuroterus lenticularis* (*Spathegaster baccarum*). Part I. *Proc. R. Soc. London B* **82:** 88-113.

Doncaster, L. (1911). Gametogenesis of the gall-fly *Neuroterus lenticularis* (*Spathegaster baccarum*). Part II. *Proc. R. Soc. London B* **83:** 476-489.

Doncaster, L. (1916). Gametogenesis and sex-determination in the gall-fly, *Neuroterus lenticularis* (*Spathegaster baccarum*). Part III. *Proc. R. Soc. London B* **89:** 183-200.

Donnelly, P. and Tavare, S. (1995). Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29:** 401-421.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science* **284:** 2124-2129.

Dowton, M. and Austin, A. D. (1995). Increased genetic diversity in mitochondrial genes is correlated with the evolution of parasitism in the Hymenoptera. *J. Mol. Evol.* **41:** 958-965.

Dowton, M. and Austin, A. D. (1999). Models of analysis for molecular datasets for the reconstruction of basal hymenopteran relationships. *Zool. Scr.* **28:** 69-74.

Doyle, J. J., Doyle, J. L., Ballenger, J. A. and Palmer, J. D. (1996). The distribution and phylogenetic significance of a 50-kb chloroplast DNA inversion in the flowering plant family Leguminosae. *Mol. Phylog. Evol.* **5:** 429-438.

Drown, D. M. and Brown, J. M. (1998). Molecular phylogeny of north American oak-galling Cynipini (Hymenoptera: Cynipidae) supports need for generic division. *In* 'The Biology of Gall-Inducing Arthropods' (Csoka, G., Mattson, W. J., Stone, G. N. and Price, P. W., Eds.), pp. 241-246, U.S. Department of Agriculture, Forest Service, North Central Research Station, St. Paul, Minnesota.

Dumolin-Lapegue, S., Demesure, B., Fineschi, S., Le Corre, V. and Petit, R. J. (1997). Phylogeographic structure of white oaks throughout the European continent. *Genetics* **146:** 1475-1487.

Dumolin-Lapegue, S., Pemonge, M. H., Gielly, L., Taberlet, P. and Petit, R. J. (1999). Amplification of oak DNA from ancient and modern wood. *Mol. Ecol.* **8:** 2137-2140.

Eady, R. D. and Quinlan, J. (1963). Hymenoptera: Cynipoidea. Key to families and subfamilies and Cynipinae (including galls). *Handbooks for the Identification of British Insects* **8(1a):** 1-81.

Edwards, A. W. F. (1992). 'Likelihood (Expanded Edition),' The John Hopkins University Press, Baltimore, Maryland.

Edwards, A. W. F. and Cavalli-Sforza, L. L. (1964). Reconstruction of evolutionary trees. *In* 'Phenetic and phylogenetic classification' (Heywood, V. H. and McNeill, J., Eds.), pp. 67-76, Systematics Association, London.

Edwards, S. V. and Beerli, P. (2000). Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* **54:** 1839-1854.

Emerson, S. B. and Hastings, P. A. (1998). Morphological correlations in evolution: consequences for phylogenetic analysis. *Q. Rev. Biol.* **73:** 141-162.

Farrell, B. D. (1998). 'Inordinate fondness' explained: why are there so many beetles? *Science* **281:** 555-559.

Fay, P. A. and Hartnett, D. C. (1991). Constraints on growth and allocation patterns of *Silphium integrifolium* (Asteraceae) caused by a cynipid gall wasp. *Oecologia* **88:** 243-250.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17:** 368-376.

Felsenstein, J. (1993). 'PHYLIP (Phylogeny Inference Package),' Distributed by the Author, Department of Genetics, University of Washington, Seattle.

Felsenstein, J. (2001). The troubled growth of statistical phylogenetics. *Syst. Biol.* **50:** 465-467.

Ferris, C., King, R. A., Vainola, R. and Hewitt, G. M. (1998). Chloroplast DNA recognizes three refugial sources of European oaks and suggests independent eastern and western immigrations to Finland. *Heredity* **80:** 584-593.

Ferris, C., Oliver, R. P., Davy, A. J. and Hewitt, G. M. (1993). Native oak chloroplasts reveal an ancient divide across Europe. *Mol. Ecol.* **2:** 337-344.

Fitch, W. M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science* **155:** 279-284.

Folliot, R. (1964). Contribution a l' etude de la biologie des Cynipides gallicoles (Hymenoptera: Cynipoidea). *Ann. Sci. Natur., Zool. (Paris), Series 12* **6:** 407-564.

Formusoh, E. S., Hatchett, J. H., Black, W. C. and Stuart, J. J. (1996). Sex-linked inheritance of virulence against wheat resistance gene H9 in the Hessian fly (Diptera: Cecidomyiidae). *Ann. Entomol. Soc. Am.* **89:** 428-434.

Forterre, P. and Philippe, H. (1999). Where is the root of the universal tree of life? *Bioessays* **21:** 871-879.

Franck, P., Garnery, L., Loiseau, A., Oldroyd, B. P., Hepburn, H. R., Solignac, M. and Cornuet, J.-M. (2001). Genetic diversity of the honeybee in Africa: microsatellite and mitochondrial data. *Heredity* **86:** 420-430.

Friedlander, T. P., Regier, J. C. and Mitter, C. (1992). Nuclear gene sequences for higher level phylogenetic analysis: 14 promising candidates. *Syst. Biol.* **41:** 483-490.

Friedlander, T. P., Regier, J. C. and Mitter, C. (1994). Phylogenetic information content of five nuclear gene sequences in animals: initial assessment of character sets from concordance and divergence studies. *Syst. Biol.* **43:** 511-525.

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

Garcia-Fernandez, J. and Holland, P. W. H. (1994). Archetypal organization of the amphioxus Hox gene cluster. *Nature* **370:** 563-566.

Gatesy, J. and O'Leary, M. A. (2001). Deciphering whale origins with molecules and fossils. *Trends Ecol. Evol.* **16:** 562-570.

Gillespie, J. H. (1986). Variability of evolutionary rates of DNA. *Genetics* **113:** 1077-1091.

Gleeson, D. M., Rowell, D. M., Tait, N. N., Briscoe, D. A. and Higgins, A. V. (1998). Phylogenetic relationships among Onychophora from Australasia inferred from the mitochondrial cytochrome oxidase subunit I gene. *Mol. Phylog. Evol.* **10:** 237-248.

Godfray, H. C. J. (1994). 'Parasitoids: Behavioural and Evolutionary Ecology,' Princeton University Press, Princeton, N.J.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36:** 182-198.

Goldman, N., Anderson, J. P. and Rodrigo, A. G. (2000). Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* **49:** 652-670.

Graybeal, A. (1994). Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. *Syst. Biol.* **43:** 174-193.

Gupta, R. S. (1998). What are archaebacteria: life's third domain or monoderm prokaryotes related to gram-positive bacteria? A new proposal for the classification of prokaryotic organisms. *Mol. Microbiol.* **29:** 695-707.

Hafner, M. S., Sudman, P. D., Villablanca, F. X., Spradling, T. A., Demastes, J. W. and Nadler, S. A. (1994). Disparate rates of molecular evolution in cospeciating hosts and parasites. *Science* **265:** 1087-1090.

Hails, R. S. and Crawley, M. J. (1991). The population dynamics of an alien insect - *Andricus quercuscalicis* (Hymenoptera, Cynipidae). *J. Anim. Ecol.* **60:** 545-562.

Handt, O., Hoss, M., Krings, M. and Paabo, S. (1994). Ancient DNA: methodological challenges. *Experientia* **50:** 534-539.

Hariri, A. R., Werren, J. H. and Wilkinson, G. S. (1998). Distribution and reproductive effects of *Wolbachia* in stalk-eyed flies (Diptera: Diopsidae). *Heredity* **81:** 254-260.

Harvey, P. H. and Pagel, M. D. (1991). 'The Comparative Method in Evolutionary Biology,' Oxford University Press, Oxford.

Hasegawa, M., Kishino, M. and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22:** 160-174.

Hashimoto, T., Sánchez, L. B., Shirakura, T., Müller, M. and Hasegawa, M. (1998). Secondary absence of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny. *Proc. Natl. Acad. Sci. U. S. A.* **95:** 6860-6865.

Hennig, W. (1966). 'Phylogenetic systematics,' University of Illinois Press, Chicago.

Hestmark, G. (2000). Temptations of the tree. *Nature* **408:** 911.

Hewitt, G. M. (1996). Some genetic consequences of ice ages, and their role in divergence and speciation. *Biol. J. Linn. Soc.* **58:** 247-276.

Hewitt, G. M. (1999). Post-glacial re-colonization of European biota. *Biol. J. Linn. Soc.* **68:** 87-112.

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

Hillis, D. M. (1996). Inferring complex phylogenies. *Nature* **383:** 130-131.

Hillis, D. M. (1998). Phylogenetic analysis of frequency data in molecular ecological studies. *In* 'Advances in Molecular Ecology' (Carvalho, G. R., Ed.) pp. 25-38, IOS Press.

Hillis, D. M. (1999). SINEs of the perfect character. *Proc. Natl. Acad. Sci. U. S. A.* **96:** 9979-9981.

Hillis, D. M. and Dixon, M. T. (1991). Ribosomal DNA: molecular evolution and phylogenetic inference. *Q. Rev. Biol.* **66:** 411-453.

Hillis, D. M., Moritz, C. and Mable, B. K. Eds. (1996). 'Molecular Systematics,' Sinauer, Sunderland, MA.

Hillis, D. M. and Wiens, J. J. (2000). Molecules versus morphology in systematics: conflicts, artefacts and misconceptions. *In* 'Phylogenetic Analysis of Morphological Data' (Wiens, J. J., Ed.) pp. 1-19.

Hoelzer, G. A. (1997). Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees revisited. *Evolution* **51:** 622-626.

Houard, C. (1912). The animal galls of southern Africa. *Ann. Soc. Entomol. France* **81:** 1-43.

Howland, D. E. and Hewitt, G. M. (1995). Phylogeny of the Coleoptera based on mitochondrial cytochrome oxidase I sequence data. *Insect Mol. Biol.* **4:** 203-215.

Huelsenbeck, J. P. (1998). Systematic bias in phylogenetic analysis: is the Strepsiptera problem solved? *Syst. Biol.* **47:** 519-537.

Huelsenbeck, J. P. and Rannala, B. (1997). Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276:** 227-232.

Huelsenbeck, J. P., Rannala, B. and Masly, J. P. (2000). Accommodating phylogenetic uncertainty in evolutionary studies. *Science* **288:** 2349-2350.

Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17:** 754-755.

Huigens, M. E., Luck, R. F., Klaassen, R. H. G., Maas, M. F. P. M., Timmermans, M. J. T. N. and Stouthamer, R. (2000). Infectious parthenogenesis. *Nature* **405:** 178-179.

Huntley, B. and Webb, T. (1989). Migration - species' response to climatic variations caused by changes in the earth's orbit. *J. Biogeogr.* **16:** 5-19.

Huntley, B. H. and Birks, J. B. (1983). 'An Atlas of Past and Present Pollen Maps for Europe: 0 -13,000 Years Ago,' Cambridge University Press, Cambridge.

Hurst, G. D. D. and Jiggins, F. M. (2000). Male-killing bacteria in insects: mechanisms, incidence, and implications. *Emerging Infect. Dis.* **6:** 329-336.

Hurst, G. D. D. and Schilthuizen, M. (1998). Selfish genetic elements and speciation. *Heredity* **80:** 2-8.

Jalas, J. and Suominen, J. (1987). 'Atlas Florae Europaeae. Distribution of Vascular Plants in Europe,' Cambridge University Press, Cambridge.

Jansen, R. K. and Palmer, J. D. (1987). A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). *Proc. Natl. Acad. Sci. U. S. A.* **84:** 5818-5822.

Jermiin, L. S. and Crozier, R. H. (1994). The cytochrome *b* region in the mitochondrial DNA of the ant *Tetraponera rufoniger*: sequence divergence in Hymenoptera may be associated with nucleotide content. *J. Mol. Evol.* **38:** 282-294.

Jeyaprakash, A. and Hoy, M. A. (2000). Long PCR improves *Wolbachia* DNA amplification: wsp sequences found in 76% of 63 arthropod species. *Insect Mol. Biol.* **9:** 393-405.

Jiggins, F. M., Bentley, J. K., Majerus, M. E. N. and Hurst, G. D. D. (2001a). How many species are infected with *Wolbachia*? Cryptic sex ratio distorters revealed to be common by intensive sampling. *Proc. R. Soc. London B* **268:** 1123-1126.

Jiggins, F. M., von der Schulenburg, J. H. G., Hurst, G. D. D. and Majerus, M. E. N. (2001b). Recombination confounds interpretations of *Wolbachia* evolution. *Proc. R. Soc. London B* **268:** 1423-1427.

Johns, G. C. and Avise, J. C. (1998). A comparative summary of genetic distances in the vertebrates from the mitochondrial cytochrome *b* gene. *Mol. Biol. Evol.* **15:** 1481-1490.

Johnstone, R. A. and Hurst, G. D. D. (1996). Maternally inherited male-killing microorganisms may confound interpretation of mitochondrial DNA variability. *Biol. J. Linn. Soc.* **58:** 453-470.

Kambysellis, M. P., Ho, K. F., Craddock, E. M., Piano, F., Parisi, M. and Cohen, J. (1995). Pattern of ecological shifts in the diversification of Hawaiian *Drosophila* inferred from a molecular phylogeny. *Curr. Biol.* **5:** 1129-1139.

Kearsley, M. J. C. and Whitham, T. G. (1998). The developmental stream of cottonwoods affects ramet growth and resistance to galling aphids. *Ecology* **79:** 178-191.

Keeling, P. J. and Doolittle, W. M. (1997). Widespread and ancient distribution of a noncanonical genetic code in diplomonads. *Mol. Biol. Evol.* **14:** 895-901.

Kido, Y., Aono, M., Yamaki, T., Matsumoto, K., Murata, S., Saneyoshi, M. and Okada, N. (1991). Shaping and reshaping of salmonid genomes by amplification of transfer RNA-derived retroposons during evolution. *Proc. Natl. Acad. Sci. U. S. A.* **88:** 2326-2330.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16:** 111-120.

King, R. A. and Ferris, C. (1998). Chloroplast DNA phylogeography of *Alnus glutinosa* (L.) Gaertn. *Mol. Ecol.* **7:** 1151-1161.

Kinsey, A. C. (1920). Phylogeny of cynipid genera and biological characteristics. *Bull. Am. Mus. Nat. Hist.* **42:** 357-402.

Kinsey, A. C. (1923). The gall wasp genus *Neuroterus* (Hymenoptera). *Indiana Univ. Studies* **10:** 1-150.

Kinsey, A. C. (1930). The gall wasp genus *Cynips*; a study in the origin of species. *Indiana Univ. Studies* **16:** 1-577.

Kinsey, A. C. (1936). The origin of higher categories in *Cynips*. *Indiana Univ. Publications, Science Series* **4:** 1-334.

Kinsey, A. C. (1937). Supra-specific variation in nature and in classification. *Am. Nat.* **71:** 206-222.

Kitching, I. J., Forey, P. L., Humphries, C. J. and Williams, D. M. (1998). 'Cladistics: The Theory and Practice of Parsimony Analysis,' Oxford University Press.

Kittayapong, P., Baisley, K. J., Baimai, V. and O'Neill, S. L. (2000). Distribution and diversity of *Wolbachia* infections in southeastern Asian mosquitoes (Diptera: Culicidae). *J. Med. Entomol.* **37:** 340-345.

Kobayashi, M., Furuya, H. and Holland, P. W. H. (1999). Dicyemids are higher animals. *Nature* **401:** 762.

Konarov, V. L. Ed. (1936). 'Flora of the USSR,' Izdatel'stva Akademii Nauk SSSR, Moscow-Leningrad.

Kopp, A., Duncan, I. and Carroll, S. B. (2000). Genetic control and evolution of sexually dimorphic characters in *Drosophila*. *Nature* **408:** 553-559.

Koulianos, S. and Schmid-Hempel, P. (2000). Phylogenetic relationships among bumble bees (*Bombus*, Latreille) inferred from mitochondrial cytochrome *b* and cytochrome oxidase I sequences. *Mol. Phylog. Evol.* **14:** 335-341.

Krimbas, C. and Powell, J. (1999). Inversion polymorphisms in *Drosophila*. *In* 'Evolutionary Genetics: From Molecules to Morphology' (Singh, R. S. and Krimbas, C. B., Eds.), pp. 284-299, Cambridge University Press.

Kristensen, N. P. (1999). Phylogeny of endopterygote insects, the most successful lineage of living organisms. *Eur. J. Entomol.* **96:** 237-253.

Kurabayashi, A. and Ueshima, R. (2000). Complete sequence of the mitochondrial DNA of the primitive opisthobranch gastropod *Pupa strigosa*: systematic implication of the genome organization. *Mol. Biol. Evol.* **17:** 266-277.

Lang, B. F., Gray, M. W. and Burger, G. (1999). Mitochondrial genome evolution and the origin of eukaryotes. *Annu. Rev. Genet.* **33:** 351-397.

Larget, B. and Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* **16:** 750-759.

Lecointre, G., Philippe, H., Lê, H. L. V. and Le Guyader, H. (1994). How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences. *Mol. Phylog. Evol.* **3:** 292-309.

Lee, M. S. Y. (1999). Molecular phylogenies become functional. *Trends Ecol. Evol.* **14:** 177-178.

Lewis, P. O. (1998a). A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Mol. Biol. Evol.* **15:** 277-283.

Lewis, P. O. (1998b). Maximum likelihood as an alternative to parsimony for inferring phylogeny using nucleotide sequence data. *In* 'Molecular Systematics of Plants II: DNA Sequencing' (Soltis, D. E., Soltis, P. S. and Doyle, J. J., Eds.), pp. 132-163, Kluwer.

Liljeblad, J. and Ronquist, F. (1998). A phylogenetic analysis of higher-level gall wasp relationships (Hymenoptera: Cynipidae). *Syst. Entomol.* **23:** 229-252.

Liu, H. and Beckenback, A. T. (1992). Evolution of the mitochondrial cytochrome oxidase II gene among 10 orders of insects. *Mol. Phylog. Evol.* **1:** 41-52.

Luikart, G., Allendorf, F. W., Cornuet, J. M. and Sherwin, W. B. (1998). Distortion of allele frequency distributions provides a test for recent population bottlenecks. *J. Hered.* **89:** 238-247.

Lund, J. N., Ott, J. R. and Lyon, R. J. (1998). Heterogony in *Belonocnema treatae* Mayr (Hymenoptera : Cynipidae). *Proc. Entomol. Soc. Wash.* **100:** 755-763.

Lutzoni, F., Pagel, M. and Reeb, V. (2001). Major fungal lineages are derived from lichen symbiotic ancestors. *Nature* **411:** 937-940.

Lyman, D. F., Monteiro, F. A., Escalante, A. A., Cordon-Rosales, C., Wesson, D. M., Dujardin, J. P. and Beard, C. B. (1999). Mitochondrial DNA sequence variation among triatomine vectors of Chagas' disease. *Am J Trop Med Hyg* **60:** 377-386.

Lyon, R. J. (1984). New cynipid wasps from California (Hymenoptera, Cynipidae). *Pan-Pacific Entomol.* **60:** 289-296.

Lyon, R. J. (1993). Synonymy of 2 genera of cynipid gall wasps and description of a new genus (Hymenoptera, Cynipidae). *Pan-Pacific Entomol.* **69:** 133-140.

Lyon, R. J. (1996). New cynipid wasps from the southwestern United States (Hymenoptera: Cynipidae). *Pan-Pacific Entomol.* **72:** 181-192.

Maddison, D. R. and Maddison, W. P. (2000). 'MacClade 4: Analysis of phylogeny and character evolution.,' Sinauer Associates, Sunderland, MA.

Maddison, W. P. (1997). Gene trees in species trees. *Syst. Biol.* **46:** 523-536.

Malloch, G., Fenton, B. and Butcher, B. D. J. (2000). Molecular evidence for multiple infections of a new subgroup of *Wolbachia* in the European raspberry beetle *Byturus tomentosus*. *Mol. Ecol.* **9:** 77-90.

Mani, M. S. (1964). 'Ecology of Plant Galls,', The Hague.

Manos, P. S., Doyle, J. J. and Nixon, K. C. (1999). Phylogeny, biogeography and processes of molecular differentiation in *Quercus* subgenus *Quercus* (Fagaceae). *Mol. Phylog. Evol.* **12:** 333-349.

Mardulyn, P. and Cameron, S. A. (1999). The major opsin in bees (Insecta: Hymenoptera): a promising nuclear gene for higher level phylogenetics. *Mol. Phylog. Evol.* **12:** 168-176.

Mardulyn, P. and Whitfield, J. B. (1999). Phylogenetic signal in the COI, 16S, and 28S genes for inferring relationships among genera of Microgastrinae (Hymenoptera; Braconidae): Evidence of a high diversification rate in this group of parasitoids. *Mol. Phylog. Evol.* **12:** 282-294.

Mason, D. J., Butlin, R. K. and Gacesa, P. (1995). An unusual mitochondrial DNA polymorphism in the *Chorthippus - Biguttulus* species group(Orthoptera, Acrididae). *Mol. Ecol.* **4:** 121-126.

Mau, B. and Newton, M. A. (1997). Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo methods. *J. Comput. Graph. Stat.* **6:** 122-131.

Maynard Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favorable gene. *Genet. Res.* **23:** 23-35.

Mayol, M. and Rossello, J. A. (2001). Why nuclear ribosomal DNA spacers (ITS) tell different stories in *Quercus*. *Mol. Phylog. Evol.* **19:** 167-176.

Mayr, E. (1982). 'The growth of biological thought : diversity, evolution, and inheritance,' Belknap Press, Cambridge, MA.

Melika, G. and Abrahamson, W. G. (1997a). Descriptions of four new species of cynipid gall wasps of the genus *Neuroterus* Hartig (Hymenoptera:Cynipidae) with redescriptions of some known species from the eastern United States. *Proc. Entomol. Soc. Wash.* **99:** 560-573.

Melika, G. and Abrahamson, W. G. (1997b). Synonymy of two genera (*Eumayria* and *Trisoleniella*) of cynipid gall wasps and description of a new genus, *Eumayriella* (Hymenoptera: Cyniphidae). *Proc. Entomol. Soc. Wash.* **99:** 666-675.

Melika, G. and Abrahamson, W. G. (2000). Review of the cynipid gall wasps of the genus *Loxaulus* Mayr (Hymenoptera : Cynipidae) with descriptions of new species. *Proc. Entomol. Soc. Wash.* **102:** 198-211.

Melika, G., Csoka, G. and Pujade-Villar, J. (2000). Check-list of oak gall wasps of Hungary, with some taxonomic notes (Hymenoptera: Cynipidae, Cynipinae, Cynipini). *Annls. hist.-nat. Mus. natn. Hung.* **92:** 265-296.

Melika, G., Stone, G. N. and Csoka, G. (1999). Description of an oak gall-wasp, *Neuroterus ambrusi* sp n. (Hymenoptera, Cynipidae) from Hungary. *Acta Zool. Acad. Sci. Hung.* **45:** 333-341.

Mindell, D. P., Sorenson, M. D. and Dimcheff, D. E. (1998). Multiple independent origins of mitochondrial gene order in birds. *Proc. Natl. Acad. Sci. U. S. A.* **95:** 10693-10697.

Mirol, P. M., Mascheretti, S. and Searle, J. B. (2000). Multiple nuclear pseudogenes of mitochondrial cytochrome b in *Ctenomys* (Caviomorpha, rodentia) with either great similarity to or high divergence from the true mitochondrial sequence. *Heredity* **84:** 538-547.

Mitchell, A., Mitter, C. and Regier, J. C. (2000). More taxa or more characters revisited: combining data from nuclear protein-coding genes for phylogenetic analysis of Noctuiodea (Insecta: Lepidoptera). *Syst. Biol.* **49:** 202-224.

Miyamoto, M. M. (1999). Perfect SINEs of evolutionary history? *Curr. Biol.* **9:** R816-R819.

Mogie, M. (2000). Historical baggage in biology: the case of higher and lower species. *Bioessays* **22:** 868-869.

Moore, W. S. (1995). Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. *Evolution* **49:** 718-726.

Moore, W. S. (1997). Mitochondrial-gene trees versus nuclear-gene trees, a reply to Hoelzer. *Evolution* **51:** 627-629.

Moorehead, J. R., Taper, M. L. and Case, T. J. (1993). Utilization of hybrid oak hosts by a monophagous gall wasp: how little host character is sufficient? *Oecologia* **95:** 385-392.

Moran, N. A. (1988). The evolution of host plant alternation in aphids: evidence for specialization as a dead end. *Am. Nat.* **132:** 681-706.

Moran, N. A. (1991). Phenotype fixation and genotypic diversity in the complex lifecycle of the aphid *Pemphigus betae*. *Evolution* **45:** 957-970.

Moran, N. A. (1994). Adaptation and constraint in the complex lifecycles of animals. *Annu. Rev. Ecol. Syst.* **25:** 573-600.

Murata, S., Takasaki, N., Saitoh, M. and Okada, N. (1993). Determination of the phylogenetic relationships among pacific salmonids by using short interspersed elements (SINEs) as temporal mandmarks of evolution. *Proc. Natl. Acad. Sci. U. S. A.* **90:** 6995-6999.

Naylor, G. J. P. and Brown, W. M. (1998). Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Syst. Biol.* **47:** 61-76.

Nei, M. (1972). Genetic distance between populations. *Am. Nat.* **106:** 283-292.

Nei, M. and Kumar, S. (2000). 'Molecular evolution and phylogenetics,' Oxford University Press, New York.

Neigel, J. E. and Avise, J. C. (1986). Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. *In* 'Evolutionary processes and theory' (Nevo, E. and Karlin, S., Eds.), pp. 515-534, Academic Press, New York.

Nieves-Aldrey, J. L. (1987). Estado actual del conocimiento de la subfamilia Cynipinae (Hymenoptera, Parasitica, Cynipidae) en la Peninsula Iberica. *Eos* **63:** 179-195.

Nieves-Aldrey, J. L. (1992). Revisíon de las especies europeas del género *Callirhytis* (Hymenoptera, Cynipidae). *Graellsia* **48:** 171-183.

Nieves-Aldrey, J. L. (1994). Revision of the west-European genera of the tribe Aylacini Ashmead (Hymenoptera, Cynipidae). *J. Hymenoptera Res.* **3:** 175-206.

Nieves-Aldrey, J.-L. (1985). La colección de cinípidos gallícolas del instituto español de entomologia. *Graellsia* **41:** 113-124.

Nieves-Aldrey, J.-L. (1988). Los Cinípidos gallícolas e inquilinos de la sierra de Guadarrama y zonas adyacentes. *Eos* **64:** 125-163.

Nieves-Aldrey, J.-L. and Pujade-Villar, J. (1985). Sombe las especies Ibéricas de la Sección I (Mayr 1872) del género *Synergus* Htg. *Eos* **41:** 219-237.

Nikaido, M., Rooney, A. P. and Okada, N. (1999). Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: hippopotamuses are the closest extant relatives of whales. *Proc. Natl. Acad. Sci. U. S. A.* **96:** 10261-10266.

Nordlander, G. (1984). What do we know about parasitic cynipoids (Hymenoptera)? *Entomologisk Tidskrift* **105:** 36-40.

Nuttall, G. H. F. (1904). 'Blood Immunity and Blood Relationship,' Cambridge University Press, Cambridge.

Nyman, T. and Julkunen-Tiito, R. (2000). Manipulation of the phenolic chemistry of willows by gall-inducing sawflies. *Proc. Natl. Acad. Sci. U. S. A.* **97:** 13184-13187.

Nyman, T., Roininen, H. and Vuorinen, J. A. (1998). Evolution of different gall types in willow-feeding sawflies (Hymenoptera: Tenthredinidae). *Evolution* **52:** 465-474.

Nyman, T., Widmer, A. and Roininen, H. (2000). Evolution of gall morphology and host-plant relationships in willow-feeding sawflies (Hymenoptera: Tenthredinidae). *Evolution* **54:** 526-533.

O'Brien, S. J., Menotti-Raymond, M., Murphy, W. J., Nash, W. G., Wienberg, J., Stanyon, R., Copeland, N. G., Jenkins, N. A., Womack, J. E. and Graves, J. A. M. (1999). The promise of comparative genomics in mammals. *Science* **286:** 458-481.

O'Brien, S. J. and Stanyon, R. (1999). Ancestral primate viewed. *Nature* **402:** 365-366.

O'Hara, R. J. (1988). Homage to Clio, or, toward an historical philosophy for evolutionary biology. *Systematic Zoology* **37:** 142-155.

O'Hara, R. J. (1997). Population thinking and tree thinking in systematics. *Zool. Scr.* **26:** 323-329.

O'Neill, L. S., Giordano, R., Colbert, M. E. A., Karr, L. T. and Robertson, M. H. (1992). 16S rRNA phylogenetic analysis of the bacterial endosymbionts associated with cytoplasmic incompatibility in insects. *Proc. Natl. Acad. Sci. U. S. A.* **89:** 2699-2702.

O'Neill, S. L., Hoffmann, A. A. and Werren, J. H. Eds. (1997). 'Influential Passengers: Inherited Microorganisms and Arthropod Reproduction,' Oxford University Press.

Pääbo, S., Irwin, D. M. and Wilson, A. C. (1990). DNA damage promotes jumping between templates during enzymatic amplification. *Journal of Biological Chemistry* **265:** 4718-4721.

Page, R. D. M., Lee, P. L. M., Becher, S. A., Griffiths, R. and Clayton, D. H. (1998). A different tempo of mitochondrial DNA evolution in birds and their parasitic lice. *Mol. Phylog. Evol.* **9:** 276-293.

Pagel, M. (1994). Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. Ser. B-Biol. Sci.* **255:** 37-45.

Pagel, M. (1997). Inferring evolutionary processes from phylogenies. *Zool. Scr.* **26:** 331-348.

Pagel, M. (1999a). Inferring the historical patterns of biological evolution. *Nature* **401:** 877-884.

Pagel, M. (1999b). The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.* **48:** 612-622.

Palmer, J. D. and Delwiche, C. F. (1998). The origin and evolution of plastids and their genomes. *In* 'Molecular Systematics of Plants II: DNA Sequencing' (Soltis, D. E., Soltis, P. S. and Doyle, J. J., Eds.), pp. 375-409, Kluwer.

Palumbi, S. R. (1989). Rates of molecular evolution and the fraction of nucleotide positions free to vary. *J. Mol. Evol.* **29:** 180-187.

Palumbi, S. R. (1996). Nucleic Acids II: The Polymerase Chain Reaction. *In* 'Molecular Systematics' (Hillis, D. M., Moritz, C. and Mable, B. K., Eds.), pp. 205-247, Sinauer, Sunderland, MA.

Palumbi, S. R., Cipriano, F. and Hare, M. P. (2001). Predicting nuclear gene coalescence from mitochondrial data: the three-times rule. *Evolution* **55:** 859-868.

Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Mol. Biol. Evol.* **5:** 568-583.

Parker, P. G., Snow, A. A., Schug, M. D., Booton, G. C. and Fuerst, P. A. (1998). What molecules can tell us about populations: choosing and using a molecular marker. *Ecology* **79:** 361-382.

Patterson, C. Ed. (1987). 'Molecules and morphology in evolution: conflict or compromise?,' Cambridge University Press, Cambridge.

Perna, N. T. and Kocher, T. D. (1996). Mitochondrial DNA: molecular fossils in the nucleus. *Curr. Biol.* **6:** 128-129.

Philippe, H. (1997). Rodent monophyly: pitfalls of molecular phylogenies. *J. Mol. Evol.* **45:** 712-715.

Philippe, H., Chenuil, A. and Adoutte, A. (1994a). Can the cambrian explosion be inferred through molecular phylogeny? *Development Supplement*: 15-25.

Philippe, H. and Laurent, J. (1998). How good are deep phylogenetic trees? *Curr. Opin. Genet. Dev.* **8:** 616-623.

Philippe, H., Sorhannus, U., Baroin, A., Perasso, R., Gasse, F. and Adoutte, A. (1994b). Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. *J. Evol. Biol.* **7:** 247-265.

Plantard, O. (1997). Ecologie des communautes de parasitoides associes aux Cynipidae galligenes (Hymenoptera): Role des caracteristiques des galles, de la structure des populations et de la phylogenie des hotes sur leur cortege parasitaire., pp. 286. University of Paris, VI, Paris.

Plantard, O., Rasplus, J. Y., Mondor, G., le Clainche, I. and Solignac, M. (1998a). *Wolbachia*-induced thelytoky in the rose gallwasp *Diplolepis spinosissimae* (Giraud) (Hymenoptera: Cynipidae), and its consequences on the genetic structure of its host. *Proc. R. Soc. London B* **265:** 1075-1080.

Plantard, O., Rasplus, J.-Y., Mondor, G., Le Clainche, I. and Solignac, M. (1999). Distribution and phylogeny of *Wolbachia* inducing thelytoky in Rhoditini and 'Aylacini' (Hymenoptera: Cynipidae). *Insect Mol. Biol.* **8:** 185-191.

Plantard, O., Shorthouse, J. D. and Rasplus, J.-Y. (1998b). Molecular phylogeny of the genus Diplolepis (Hymenoptera: Cynipidae). *In* ' The Biology of Gall-Inducing Arthropods' (Csoka, G., Mattson, W. J., Stone, G. N. and Price, P. W., Eds.), pp. 247-260, U.S. Department of Agriculture, Forest Service, North Central Research Station, St. Paul, Minnesota.

Poe, S. and Wiens, J. J. (2000). Character selection and the methodology of morphological phylogenetics. *In* 'Phylogenetic Analysis of Morphological Data' (Wiens, J. J., Ed.) pp. 20-36, Smithsonian Institution Press.

Posada, D. and Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14:** 817-818.

Powell, J. R. (1983). Interspecific cytoplasmic gene flow in the absence of nuclear gene flow: evidence from Drosophila. *Proc. Natl. Acad. Sci. U. S. A.* **80:** 492-495.

Price, P. W., Fernandes, G. W. and Waring, G. L. (1987). Adaptive nature of insect galls. *Environ. Entomol.* **16:** 15-24.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155:** 945-959.

Pujade, J. (1994). Formes cinipo-cedidogenes detectades o que poden detectar-se, en les flors i els fruits de les fagacies a Andorra (Hymenoptera: Cynipidae: Cynipinae). *Ann. Inst. Est. And. Barcelona*:

Pujade-Villar, J. (1992). *Andricus kollari* (Hartig) (Insecta: Hymenoptera:Cynipidae) Part II: Considerations sobre el seu cicle biológic. *La Sitja del Llop* **3:** 12.

Raff, R. A., Field, K. G., Olsen, G. J., Giovannoni, S. J., Lane, D. J., Ghiselin, M. T., Pace, N. R. and Raff, E. C. (1989). Metazoan phylogeny based on analysis of 18S ribosomal RNA. *In* 'The Hierarchy of Life. Molecules and Morphology in Phylogenetic Analysis' (Fernholm, B., Bremer, K. and Jörnvall, H., Eds.), pp. 247-260, Excerpta Medica/Elsevier, Amsterdam.

Rambaut, A. and Grassly, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* **13:** 235-238.

Rand, D. M. (1994). Thermal habit, metabolic rate and the evolution of mitochondrial DNA. *Trends Ecol. Evol.* **9:** 125-131.

Ratcliffe, R. H., Cambron, S. E., Flanders, K. L., Bosque-Perez, N. A., Clement, S. L. and Ohm, H. W. (2000). Biotype composition of Hessian fly (Diptera : Cecidomyiidae)

populations from the southeastern, midwestern, and northwestern United States and virulence to resistance genes in wheat. *J. Econ. Entomol.* **93:** 1319-1328.

Ratcliffe, R. H., Safranski, G. G., Patterson, F. L., Ohm, H. W. and Taylor, P. L. (1994). Biotype status of Hessian fly (Diptera, Cecidomyiidae) populations from the eastern United States and their response to 14 Hessian fly resistance genes. *J. Econ. Entomol.* **87:** 1113-1121.

Raubeson, L. A. and Jansen, J. K. (1992). A rare chloroplast-DNA structural mutation is shared by all conifers. *Biochem. Syst. Ecol.* **20:** 17-24.

Raymond, M. and Rousset, F. (1995). Genepop (version-1.2) - population-genetics software for exact tests and ecumenicism. *J. Hered.* **86:** 248-249.

Richardson, B. J., Baverstock, P. R. and Adams, M. (1986). 'Allozyme electrophoresis: A handbook for animal systematics and population studies,' Academic Press.

Rivera, M. C. and Lake, J. A. (1992). Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* **257:** 74-76.

Rohfritsch, O. (1992). Patterns in gall development. *In* 'Biology of Insect-Induced Galls' (Shorthouse, J. D. and Rohfritsch, O., Eds.), pp. 60-87, Oxford University Press.

Rokas, A., Atkinson, R. J., Brown, G. S., West, S. A. and Stone, G. N. (2001a). Understanding patterns of genetic diversity in the oak gallwasp *Biorhiza pallida*: demographic history or a *Wolbachia* selective sweep? *Heredity* **87:** 294-304.

Rokas, A. and Holland, P. W. H. (2000). Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.* **15:** 454-459.

Rokas, A., Kathirithamby, J. and Holland, P. W. H. (1999). Intron insertion as a phylogenetic character: the *engrailed* homeobox of Strepsiptera does not indicate affinity with Diptera. *Insect Mol. Biol.* **8:** 527-530.

Rokas, A., Nylander, J. A. A., Ronquist, F. and Stone, G. N. (2001b). A maximum likelihood analysis of eight phylogenetic markers in gallwasps (Hymenoptera: Cynipidae); implications for insect phylogenetic studies. *Mol. Phylog. Evol.* **21:** in press.

Ronquist, F. (1994). Evolution of parasitism among closely related species: phylogenetic relationships and the origins of inquilism in gall wasps (Hymenoptera, Cynipidae). *Evolution* **48:** 241-266.

Ronquist, F. (1995). Phylogeny and early evolution of the Cynipoidea (Hymenoptera). *Syst. Entomol.* **20:** 309-335.

Ronquist, F. (1999). Phylogeny, classification and evolution of the Cynipoidea. *Zool. Scr.* **28:** 139-164.

Rosenberg, M. S. and Kumar, S. (2001). Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc. Natl. Acad. Sci. U. S. A.* **98:** 10751-10756.

Rozas, J. and Rozas, R. (1999). DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15:** 174-175.

Samuel, R., Bachmair, A., Jobst, J. and Ehrendorfer, F. (1998). ITS sequences from nuclear rDNA suggest unexpected phylogenetic relationships between Euro-Mediterranean, East Asiatic and North American taxa of *Quercus* (Fagaceae). *Plant Syst. Evol.* **211:** 129-139.

Scharf, S. J., Long, C. M. and Erlich, H. A. (1988). Sequence analysis of the HLA-DR beta and HLA-DQ beta loci from three *Pemphigus vulgaris* patients. *Hum. Immun.* **22:** 61-69.

Schilthuizen, M. and Stouthamer, R. (1997). Horizontal transmission of parthenogenesis-inducing microbes in *Trichogramma* wasps. *Proc. R. Soc. London B* **264:** 361-366.

Schilthuizen, M. and Stouthammer, R. (1998). Distribution of *Wolbachia* among the guild associated with the parthenogenetic gall wasp *Diplolepis rosae*. *Heredity* **81:** 270-274.

Schönrogge, K., Harper, L. J. and Lichtenstein, C. P. (2000). The protein content of tissues in cynipid galls (Hymenoptera : Cynipidae): Similarities between cynipid galls and seeds. *Plant Cell Env.* **23:** 215-222.

Schönrogge, K., Stone, G. N. and Crawley, M. J. (1995). Spatial and temporal variation in guild structure - parasitoids and inquilines of Andricus quercuscalicis (Hymenoptera, Cynipidae) in its native and alien ranges. *Oikos* **72:** 51-60.

Schönrogge, K., Walker, P. and Crawley, M. J. (1998). Invaders on the move: parasitism in the sexual galls of four alien gall wasps in Britain (Hymenoptera : Cynipidae). *Proc. R. Soc. Lond. Ser. B-Biol. Sci.* **265:** 1643-1650.

Shedlock, A. M. and Okada, N. (2000). SINE insertions: powerful tools for molecular systematics. *Bioessays* **22:** 148-160.

Shoemaker, J. S., Painter, I. S. and Weir, B. S. (1999). Bayesian statistics in genetics: a guide for the uninitiated. *Trends Genet.* **15:** 354-358.

Shorthouse, J. and Rohfritsch, O. Eds. (1992). 'Biology of insect-induced galls,' Oxford University Press.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H. and Flook, P. (1994). Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Ann. Entomol. Soc. Am.* **87:** 651-701.

Slowinski, J. B. (2001). Molecular polytomies. *Mol. Phylog. Evol.* **19:** 114-120.

Slowinski, J. B. and Page, R. D. M. (1999). How should species phylogenies be inferred from sequence data? *Syst. Biol.* **48:** 814-825.

Sober, E. (1988). 'Reconstructing the Past: Parsimony, Evolution and Inference,' MIT Press, Cambridge, MA.

Solignac, M., Vautrin, D. and Rousset, F. (1994). Widespread occurence of the proteobacteria *Wolbachia* and partial cytoplasmic incompatibility in *Drosophila melanogaster*. *C. R. Acad. Sci. Paris, Serie III* **317:** 461-470.

Soller, R., Wohltmann, A., Witte, H. and Blohm, D. (2001). Phylogenetic relationships within terrestrial mites (Acari : Prostigmata, Parasitengona) inferred from comparative DNA sequence analysis of the mitochondrial cytochrome oxidase subunit I gene. *Mol. Phylog. Evol.* **18:** 47-53.

Stellwag, E. J. (1999). Hox gene duplication in fish. *Semin. Cell Dev. Biol.* **10:** 531-540.

Stern, D. L. (1995). Phylogenetic evidence that aphids, rather than plants, determine gall morphology. *Proc. R. Soc. London B* **260:** 85-89.

Sternlicht, M. (1968). 'The oak galls of Israel'.

Stille, B. and Davring, L. (1980). Meiosis and reproductive strategy in the parthenogenetic gall wasp *Diplolepis rosae* (L.) (Hymenoptera, Cynipidae). *Hereditas* **92:** 353-362.

Stone, G. N., Atkinson, R., Rokas, A., Csóka, G. and Nieves-Aldrey, J.-L. (2001). Differential success in northwards range expansion between ecotypes of the marble gallwasp *Andricus kollari*: a tale of two lifecycles. *Mol. Ecol.* **10:** 761-778.

Stone, G. N. and Cook, J. M. (1998). The structure of cynipid oak galls: patterns in the evolution of an extended phenotype. *Proc. R. Soc. London B* **265:** 979-988.

Stone, G. N., Schönrogge, K., Atkinson, R. J., Bellido, D. and Pujade-Villar, J. (2002). The population biology of oak gallwasps (Hymenoptera: Cynipidae). *Annu. Rev. Entomol.* **47:** 633-668.

Stone, G. N., Schonrogge, K., Crawley, M. J. and Fraser, S. (1995). Geographic and between-generation variation in the parasitoid communities associated with an invading Ggallwasp, *Andricus quercuscalicis* (Hymenoptera, Cynipidae). *Oecologia* **104:** 207-217.

Stone, G. N. and Sunnucks, P. (1993). Genetic consequences of an invasion through a patchy environment - the cynipid gallwasp *Andricus quercuscalicis* (Hymenoptera: Cynipidae). *Mol. Ecol.* **2:** 251-268.

Stouthamer, R. (1997). *Wolbachia*-induced parthenogenesis. *In* 'Influential Passengers: Inherited Microorganisms and Arthropod Reproduction' (O'Neill, S. L., Hoffmann, A. A. and Werren, J. H., Eds.), pp. 102-124, Oxford University Press, Oxford.

Stouthamer, R., Breeuwer, J. A. J. and Hurst, G. D. D. (1999). *Wolbachia pipientis*: Microbial manipulator of arthropod reproduction. *Annu. Rev. Microbiol.* **53:** 71-102.

Stouthamer, R. and Kazmer, D. J. (1994). Cytogenetics of microbe-associated parthenogenesis and its consequence for gene flow in *Trichogramma* wasps. *Heredity* **73:** 317-327.

Strimmer, K. and von Haeseler, A. (1996). Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13:** 964-969.

Sturtevant, A. H. (1921). A case of rearrangement of genes in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* **7:** 235-237.

Sturtevant, A. H. (1942). 'The Classification of the genus *Drosophila*,' University of Texas Publ. no. 4213.

Sturtevant, A. H. and Dobzhansky, T. (1936). Inversions in the third chromosome of wild races of Drosophila pseudoobscura, and their use in the study of the history of the species. *Proc. Natl. Acad. Sci. U. S. A.* **22:** 448-450.

Sturtevant, A. H. and Tan, C. C. (1937). The comparative genetics of *Drosophila pseudoobscura* and *D. melanogaster*. *J. Genet.* **34:** 415-432.

Sullivan, J., Swofford, D. L. and Naylor, G. J. P. (1999). The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. *Mol. Biol. Evol.* **16:** 1347-1356.

Sunnucks, P. J. and Stone, G. N. (1996). Genetic structure of invading insects and the case of the knopper gallwasp. *In* 'Frontiers of Population Ecology' (Floyd, R. B., Sheppard, A. W. and De Barro, P. J., Eds.), pp. 485-496, CSIRO Publishing, Collingwood, Australia.

Suomalainen, E., Saura, A. and Lokki, J. (1987). 'Cytology and Evolution in Parthenogenesis,' CRC Press, Inc., Boca Raton, Florida.

Swofford, D. L. (2000). 'PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods),' Sinauer, Sunderland, MA.

Swofford, D. L., Olsen, G. J., Waddell, P. J. and Hillis, D. M. (1996). Phylogenetic inference. *In* 'Molecular Systematics' (Hillis, D. M., Moritz, C. and Mable, B. K., Eds.), pp. 407-514, Sinauer, Sunderland, MA.

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

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

Telford, M. J. and Holland, P. W. H. (1997). Evolution of 28S ribosomal DNA in chaetognaths: duplicate genes and molecular phylogeny. *J. Mol. Evol.* **44:** 135-144.

Thelwell, N. J., Huisman, R. A., Harbach, R. E. and Butlin, R. K. (2000). Evidence for mitochondrial introgression between *Anopheles bwambae* and *Anopheles gambiae*. *Insect Mol. Biol.* **9:** 203-210.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). Clustal-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673-4680.

Thorne, J. L. (2000). Models of protein sequence evolution and their applications. *Curr. Opin. Genet. Dev.* **10:** 602-605.

Toumi, L. and Lumaret, R. (1998). Allozyme variation in cork oak (*Quercus suber* L.): the role of phylogeography and genetic introgression by other Mediterranean oak species and human activities. *Theor. Appl. Genet.* **97:** 647-656.

Townsend, C. C. and Guest, E. R. Eds. (1980). 'Flora of Iraq. Volume 4, part 1: Cornaceae to Rubiaceae,' Ministry of Agriculture and Agrarian Research, Baghdad.

True, J. R., Edwards, K. A., Yamamoto, D. and Carroll, S. B. (1999). *Drosophila* wing melanin patterns form by vein-dependent elaboration of enzymatic prepatterns. *Curr. Biol.* **9:** 1382-1391.

Turelli, M. (1994). Evolution of incompatibility-inducing microbes and their hosts. *Evolution* **48:** 1500-1513.

Turelli, M., Hoffmann, A. A. and McKechnie, W. S. (1992). Dynamics of cytoplasmic incompatibility and mtDNA variation in natural *Drosophila simulans* populations. *Genetics* **132:** 713-723.

Turrill, W. B. (1929). 'The plant life of the Balkan peninsula; a phytogeographical study,' The Clarendon Press, Oxford.

van Meer, M. M. M., Witteveldt, J. and Stouthamer, R. (1999). Phylogeny of the arthropod endosymbiont *Wolbachia* based on the *wsp* gene. *Insect Mol. Biol.* **8:** 399-408.

van Valen, L. (1976). Ecological species, multispecies, and oaks. *Taxon* **25:** 233-239.

Vavre, F., Fleury, F., Lepetit, D., Fouillet, P. and Bouletreau, M. (1999). Phylogenetic evidence for horizontal transmission of *Wolbachia* in host-parasitoid associations. *Mol. Biol. Evol.* **16:** 1711-1723.

Venkatesh, B., Ning, Y. and Brenner, S. (1999). Late changes in spliceosomal introns define clades in vertebrate evolution. *Proc. Natl. Acad. Sci. U. S. A.* **96:** 10267-10271.

Verneau, O., Catzeflis, F. and Furano, A. V. (1998). Determining and dating recent rodent speciation events by using L1 (LINE-1) retrotransposons. *Proc. Natl. Acad. Sci. U. S. A.* **95:** 11284-11289.

von Dohlen, C. D. and Moran, N. A. (2000). Molecular data support a rapid radiation of aphids in the Cretaceous and multiple origins of host alternation. *Biol. J. Linn. Soc.* **71:** 689-717.

Wagner, A., Blackstone, N., Cartwright, J., Dick, M., Misof, B., Srow, P., Wagner, G. P., Bartels, J., Murtha, M. and Pendleton, J. (1994). Surveys of gene families using polymerase chain reaction: PCR selection and PCR drift. *Syst. Biol.* **43:** 250-261.

Wehrmaker, A. (1998). On the beginnings of cecidology in 19th century north America: Bassett's discovery of heterogony in oak gall wasps (Hymenoptera:Cynipidae). *In* 'The biology of gall-inducing arthropods' (Csóka, G., Mattson, W. J., Stone, G. N. and Price, P. W., Eds.), pp. 106-110, U.S. Department of Agriculture, Forest Service, North Central Research Station, St. Paul, Minnesota.

Weis, A. E. and Abrahamson, W. G. (1986). Evolution of host-plant manipulation by gall makers - ecological and genetic factors in the *Solidago - Eurosta* system. *Am. Nat.* **127:** 681-695.

Weis, A. E., Walton, R. and Crego, C. L. (1988). Reactive plant tissue sites and the population biology of gall makers. *Annu. Rev. Entomol.* **33:** 467-486.

Weld, L. H. (1957). 'Cynipid galls of the Pacific Slope,' Privately printed, Ann Arbor, Michigan.

Weld, L. H. (1959). 'Cynipid galls of the Eastern United States,' Privately printed, Ann Arbor, Michigan.

Weld, L. H. (1960). 'Cynipid galls of the Southwest,' Privately printed, Ann Arbor, Michigan.

Wenseleers, T. and Billen, J. (2000). No evidence for *Wolbachia*-induced parthenogenesis in the social Hymenoptera. *J. Evol. Biol.* **13:** 277-280.

Wenseleers, T., Ito, F., Van Borm, S., Huybrechts, R., Volckaert, F. and Billen, J. (1998). Widespread occurrence of the micro-organism *Wolbachia* in ants. *Proc. R. Soc. London B* **265:** 1447-1452.

Werren, J. H. (1997). Biology of *Wolbachia*. *Annu. Rev. Entomol.* **42:** 587-609.

Werren, J. H. (1998). *Wolbachia* and speciation. *In* 'Endless forms: species and speciation' (Howard, D. and Berlocher, S., Eds.), pp. 245-260, Oxford University Press, Oxford.

Werren, J. H. and Bartos, J. D. (2001). Recombination in *Wolbachia*. *Curr. Biol.* **11:** 431-435.

Werren, J. H., Wan, Z. and Lirong, G. (1995a). Evolution and phylogeny of *Wolbachia*: reproductive parasites of arthropods. *Proc. R. Soc. London B* **261:** 55-63.

Werren, J. H., Windsor, D. and Guo, L. (1995b). Distribution of *Wolbachia* among neotropical arthropods. *Proc. R. Soc. London B* **262:** 197-204.

Werren, J. H. and Windsor, D. M. (2000). *Wolbachia* infection frequencies in insects: evidence of a global equilibrium? *Proc. R. Soc. London B* **267:** 1277-1286.

West, S. A., Cook, J. M., Werren, J. H. and Godfray, C. J. (1998). *Wolbachia* in two insect host-parasitoid communities. *Mol. Ecol.* **7:** 1457-1465.

White, T. J., Bruns, T., Lee, S. and Taylor, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *In* 'PCR Protocols: A Guide to Methods and Applications' (Innis, M. A., Gelfand, D. H. and Sninsky, J. J., Eds.), pp. 315-322, Academic Press.

Whitfield, J. B. and Cameron, S. A. (1998). Hierarchical analysis of variation in the mitochondrial 16S rRNA gene among Hymenoptera. *Mol. Biol. Evol.* **15:** 1728-1743.

Whitham, T. G. (1978). Habitat selection by *Pemphigus* aphids in response to resource limitation and competition. *Ecology* **59:** 1164-1176.

Whiting, M. F., Carpenter, J. C., Wheeler, Q. D. and Wheeler, W. C. (1997). The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from the 18S and 28S ribosomal sequences and morphology. *Syst. Biol.* **46:** 1-68.

Whiting, M. F. and Wheeler, W. C. (1994). Insect homeotic transformation. *Nature* **368:** 696.

Wiebes-Rijks, A. A. and Shorthouse, J. D. (1992). Ecological relationships of insects inhabiting cynipid galls. *In* 'Biology of Insect-Induced Galls' (Shorthouse, J. D. and Rohfritsch, O., Eds.), pp. 238-257, Oxford University Press.

Wiegmann, B. M., Mitter, C., Regier, J. C., Friedlander, T. P., Wagner, D. M. and Nielsen, E. S. (2000). Nuclear genes resolve Mesozoic-aged divergences in the insect order Lepidoptera. *Mol. Phylog. Evol.* **15:** 242-259.

Wood, B. W. and Payne, J. A. (1988). Growth regulators in chestnut shoot galls infected with oriental chestnut gall wasp, *Dryocosmus kuriphilus* (Hymenoptera, Cynipidae). *Env. Entomol.* **17:** 915-920.

Wu, C. I. (1991). Inferences of species phylogeny in relation to segregation of ancient polymorphisms. *Genetics* **127:** 429-435.

Yaltirik, F. Ed. (1982). 'Fagaceae,' Edinburgh University Press, Edinburgh.

Yang, L., Wiegmann, B. M., Yeates, D. K. and Irwin, M. E. (2000). Higher-level phylogeny of the Therevidae (Diptera: Insecta) based on 28S ribosomal and elongation factor-1a gene sequences. *Mol. Phylog. Evol.* **15:** 440-451.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10:** 1396-1401.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39:** 306-314.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11:** 367-372.

Yang, Z., Goldman, N. and Friday, A. (1994). Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11:** 316-324.

Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol. Biol. Evol.* **14:** 717-724.

Yang, Z. and Yoder, A. D. (1999). Estimation of the transition/transversion rate bias and species sampling. *J. Mol. Evol.* **48:** 274-283.

Zantoko, L. and Shukle, R. H. (1997). Genetics of virulence in the hessian fly to resistance gene H13 in wheat. *J. Hered.* **88:** 120-123.

Zantoko, L. and Shukle, R. H. (1999). An STS linked to a hessian fly locus controlling virulence to resistance gene H13 in wheat. *J. Hered.* **90:** 242-247.

Zardoya, R. and Meyer, A. (1996). Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. *Mol. Biol. Evol.* **13:** 933-942.

Zhang, D. X. and Hewitt, G. M. (1996). Nuclear integrations: Challenges for mitochondrial DNA markers. *Trends Ecol. Evol.* **11:** 247-251.

Zhou, W., Rousset, F. and O'Neill, S. (1998). Phylogeny and PCR-based classification of *Wolbachia* strains using *wsp* gene sequences. *Proc. R. Soc. London B* **265:** 509-515.

Zohary, M. Ed. (1966). 'Flora Palaestina. Volume 1: Equisetaeae to Moringaceae,' The Israel Academy of Sciences and Humanities, Jerusalem.

APPENDIX I. Published works

This Appendix contains my published contributions that have appeared in scientific journals during the course of this Ph.D. Most contributions are not included in the preceding chapters, unless specifically stated [in brackets].

Contents (in order of appearance):

1. Rokas, A., J. Kathirithamby and P. W. H. Holland 1999. Intron insertion as a phylogenetic character: the *engrailed* homeobox of Strepsiptera does not indicate affinity with Diptera. *Insect Mol. Biol.* **8**: 527-530. [research article; the experimental work was conducted prior to the start of this Ph.D.]
2. Rokas, A. 2000. *Wolbachia* as a speciation agent. *Trends Ecol. Evol.* **15**: 44-45. [commentary]
3. Averof, M., A. Rokas, K. H. Wolfe and P. M. Sharp 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* **287**: 1283-1286. [research article; the experimental work was conducted prior to the start of this Ph.D.]
4. Cook, J. M. and A. Rokas 2000. Influential passengers come of age, 1st International *Wolbachia* Conference, Orthodox Academy, Kolymbari, Crete, Greece, 7–12 June 2000. *Trends Genet.* **16**: 378-379. [meeting report]
5. Rokas, A. and G. McVean 2000. A Bayesian guide to tree felling. *Trends Ecol. Evol.* **15**: 396. [journal club]
6. Rokas, A. and P. W. H. Holland 2000. Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.* **15**: 454-459. [review; Chapter 2]
7. Rokas, A. 2001. Evolutionary biology meets genomics. *Trends Ecol. Evol.* **16**: 15-16. [journal club]
8. Stone, G. N., R. J. Atkinson, A. Rokas, G. Csóka and J.- L. Nieves-Aldrey 2001. Differential success in northwards range expansion between ecotypes of the

marble gallwasp *Andricus kollari*: a tale of two refugia. *Mol. Ecol.* **10**: 761-778. [research article]

9. Rokas, A. 2001. Selfish element wars in a wasp battleground. *Trends Ecol. Evol.* **16**: 277. [journal club]
10. Rokas, A. 2001. What is a character? *The Character Concept in Evolutionary Biology* by G. P. Wagner (ed.), Academic Press, 2000. *Trends Ecol. Evol.* **16**: 590. [book review]
11. Rokas, A., R. J. Atkinson, G. S. Brown, S. A. West and G. N. Stone 2001. Understanding patterns of genetic variation in the oak gallwasp *Biorhiza pallida*: demographic history or a *Wolbachia* selective sweep? *Heredity* **87**: 294-304. [research article; Chapter 4]
12. Rokas, A. 2001. Getting it right for the wrong reason. *Trends Ecol. Evol.* **16**: 668. [journal club]

Three additional contributions have not been included in this Appendix:

13. Rokas, A. 2000. Bacteria altering reproduction in gall wasps or what happens if you don't take your antibiotics! *Cecidology* **15**: 2-7. [commentary]
14. Rokas, A. and D. Charlesworth 2001. *Molecular Evolution and Phylogenetics* by M. Nei and S. Kumar, Oxford University Press, 2000. *Genet. Res.* **77**: 117-119. [book review]
15. Rokas, A., J. A. A. Nylander, F. Ronquist and G. N. Stone 2002. A maximum likelihood analysis of eight phylogenetic markers in gallwasps (Hymenoptera: Cynipidae); implications for insect phylogenetic studies. *Mol. Phylog. Evol.*: in press. [research article; Chapter 3]

Insect Molecular Biology (1999) 8(4), 527–530

Intron insertion as a phylogenetic character: the *engrailed* homeobox of Strepsiptera does not indicate affinity with Diptera

A. Rokas,* J. Kathirithamby† and P.W. H. Holland

School of Animal and Microbial Sciences, The University of Reading, U.K. and †Department of Zoology, University of Oxford, U.K.

Abstract

The phylogenetic relationships of the order Strepsiptera are unclear. Affiliation to Coleoptera has been proposed, however this implies that dipteran halteres and strepsipteran haltere-like organs evolved convergently. An alternative is a sister group relationship with Diptera. In this case, halteres could be homologous but a radical homeotic mutation may have switched their position to the Strepsipteran mesothorax. Ribosomal DNA sequence analysis has been used to support Dipteran affiliation, although this is controversial. Here we investigate the potential of an intron insertion site as a phylogenetic character. We find that the *en* homeobox gene of the strepsipteran *Stichotrema dallatorreanum* lacks a derived intron insertion shared by representatives of Diptera and Lepidoptera. We argue against a close affiliation between Strepsiptera and Diptera.

Keywords: Strepsiptera, *Stichotrema*, phylogeny, intron, *engrailed*.

Introduction

Insects of the order Strepsiptera are entomophagous endoparasitoids with extreme sexual dimorphism. Many species are known only from the free-living winged males, because females are permanently endoparasitic (with the exception of the family Mengenillidae where females emerge to pupate externally, but are still wingless as adults). Ever since the first strepsipteran was described in 1793, their phylogenetic affinities have been a subject of debate and confusion. In recent years, four alternative phylogenetic placements have been seriously suggested: as a sister group to Coleoptera (e.g. Kathirithamby, 1989; Kukalova-Peck & Lawrence, 1993), within the Coleoptera (Crowson, 1960), outside the Holometabola (Kristensen, 1991) and as a sister group to the Diptera (Whiting *et al.*, 1997). A proposed affinity to Coleoptera is based primarily on the use of hindwings for flight in both orders, several associated morphological characters and the pattern of hindwing venation (Crowson, 1960; Kathirithamby, 1989; Kukalova-Peck & Lawrence, 1993). The major morphological feature suggesting affinity to Diptera is the fact that strepsipteran forewings are reduced, resembling the metathoracic halteres of dipterans. The reduced forewings of strepsipterans play the same role as dipteran halteres; both act as gyroscopic balancing organs during flight (Pix *et al.*, 1993). Possession of a similar structure on different segments of the body is often taken as evidence against homology. This is not necessarily the case, because mutations in developmental control genes are capable of transforming the positions of structures along the body axis (homeotic mutations; Lewis, 1978). Whiting & Wheeler (1994) made the intriguing suggestion that the different segmental position of halteres and haltere-like organs in Diptera and Strepsiptera may reflect a natural homeotic mutation that caused a reversal of T2 and T3 segment identities in the ancestral strepsipteran lineage. This example has been cited as a very rare case of homeotic mutation contributing to body plan evolution (Raff, 1996).

In the face of conflicting morphological evidence, molecular data can give an independent insight into phylogenetic relationships. Molecular phylogenetic analysis of 18S ribosomal DNA (rDNA) sequences has been interpreted as supporting a sister group relationship between Strepsiptera and Diptera (Chalwatzis *et al.*, 1996; Whiting *et al.*, 1997). However, a thorough analysis of 28S and 5.8S rDNA sequences concluded that strong support could only be given to placement of Strepsiptera within the Holometabola; affinity with either Diptera or Coleoptera

*Present address: Institute of Cell, Animal and Population Biology, University of Edinburgh, King's Building, West Mains Road, Edinburgh EH9 3JT, U.K.

Received 2 March 1999; accepted 10 June 1999. Correspondence: Professor Peter W. H. Holland, School of Animal and Microbial Sciences, The University of Reading, Whiteknights, PO Box 228, Reading RG6 6AJ, U.K. E-mail: p.w.h.holland@reading.ac.uk

could not be distinguished due to saturation of nucleotide substitutions (Hwang *et al.*, 1998). Furthermore, other authors have suggested caution in interpretation of the 18S rDNA data, noting the extremely fast rates of rDNA evolution in Diptera and Strepsiptera (Carmean & Crespi, 1995). It is known that fast evolving sequences can be artefactually grouped by molecular phylogenetic analyses (Felsenstein, 1978), and recent work suggests that strepsipteran and dipteran rDNA branch lengths are long enough to attract in this manner (Huelsenbeck, 1997; Huelsenbeck, 1998). Therefore, current molecular data do not definitively resolve the phylogenetic placement of Strepsiptera within the holometabolous insects.

When the resolving power of primary sequence data is reduced, rare large-scale molecular changes provide an independent source of phylogenetic information, largely immune from rate effects (Holland & Garcia-Fernàndez, 1996). The principal difficulty with this approach is identifying rare mutations unique to clades of interest. Examples of rare mutations that have been used for phylogenetic inference include mitochondrial gene order, retroposon integrations, gene duplication and protein domain evolution (Holland & Garcia-Fernàndez, 1996; Balavoine, 1997). Here we propose the use of intron insertions as shared derived characters, and examine the potential of a recently evolved intron in the *engrailed* (*en*) homeobox gene as a diagnostic character for resolving the phylogenetic affinities of Strepsiptera.

Results and Discussion

We first compared intron-exon organization of all *en* class homeobox genes currently available on GenBank. We note that an intron insertion within the homeobox is located at an identical site in the two *en* class genes (*en* and *inv*) of the dipteran *Drosophila melanogaster*. The *en* gene of *Anopheles gambiae* (Diptera), the *en* gene of *D. virilis* (Diptera) and the *en* and *inv* genes of *Bombyx mori* (Lepidoptera) all have an intron present at the same site within the homeobox. These represent all the *en* class genes characterized at the genomic level from Diptera and Lepidoptera. In contrast to these two orders, an intron at this position in the homeobox is absent from *en* class genes characterized from representatives of Coleoptera (*Tribolium castaneum*), Hymenoptera (*Apis mellifera*) and all outgroup taxa (Crustacea, Chelicerata, Onycophora, Annelida, Mollusca, Brachiopoda, Echinodermata, Chordata; see Experimental procedures). Clearly, absence of the intron is the primitive state, and presence of the intron in Diptera and Lepidoptera is the derived condition. Because intron insertion is rare, and intron position is likely to be selectively neutral, these types of mutation are effectively immune from convergent evolution. Hence, the intron considered here is a shared derived character of Diptera and Lepidoptera.

If Strepsiptera are indeed the sister group to Diptera, and evolved by homeosis, we predict that they are likely to possess this intron, at precisely the same insertion site as in flies and moth; absence is predicted if Strepsiptera are allied to Coleoptera or to any other order of insects. In an attempt to resolve between these alternatives, we used degenerate PCR to clone the *en* homeobox from the Papua New Guinea strepsipteran *Stichotrema dallatorreanum* Hofeneder, an endoparasite of a tettigonid grasshopper, *Segestidea novaeguineae* (Brancsik).

One set of PCR primers was designed to flank the diagnostic intron (JM36b and AROUT); a second set targets a region adjacent to the intron to control for potential amplification bias resulting from intron size (AR3 and MAQGLY; Fig. 1). After DNA amplification with each primer set, bands were cloned and multiple independent clones sequenced. Two distinct *en* class genes were isolated from the strepsipteran sample. One sequence has high identity to the *en* gene of the locust *Schistocerca gregaria*, suggesting it is host-derived; this was confirmed by cloning of this (and no other) *en* gene from an unparasitized specimen of the host. The second sequence is derived from the strepsipteran genome. This was verified by designing a gene-specific primer (STREN) which, when used with the 5′ degenerate primer (JM36b), amplified the strepsipteran gene from two additional strepsipteran specimens (*Stichotrema* sp.). One specimen yielded a sequence identical over seventy-eight nucleotides, whereas the other yielded clones with one synonymous mismatch (G to A) at position 72. Because all four clones sequenced from this animal shared this transition, it is not a PCR error; it is likely to be an allelic variant of the same gene (Fig. 2). These primers did not amplify from control host DNA, confirming strepsipteran origin.

Importantly, we found that the strepsipteran *en* gene does not possess an intron in the homeobox (nor does the grasshopper gene). This contrasts to all four previously reported dipteran *en* class genes (*Drosophila melanogaster*, *D. virilis*, *Anopheles en*; *D. melanogaster inv*) and both lepidopteran *en* class genes (*Bombyx en* and *inv*).

(A) *Stichotrema dallatorreanum en*

```
AGTAGTGCTCAACTGGCAAGACTGAAGCACGAGTTTGCGGAAAATCGTTACCTAACCGAA
 S  S  A  Q  L  A  R  L  K  H  E  F  A  E  N  R  Y  L  T  E
AAACGGAGACAGCAATTGAGCAATGAGTTGGGACTCAACGAGGCGCAGATCAAGATCTGG
 K  R  R  Q  Q  L  S  N  E  L  G  L  N  E  A  Q  I  K  I  W
TTTCAAAATAAAAGGGCAAAAATTAAAAAATCTTC
 F  Q  N  K  R  A  K  I  K  K  S  S
```

(B) *Segestidea novaeguineae en*

```
AGCGGAGAGCAACTGGCGAGGCTGAAGCACGAGTTCGCGGAGAACCGCTACCTGACGGAG
 S  G  E  Q  L  A  R  L  K  H  E  F  A  E  N  R  Y  L  T  E
CGAAGACGCCAAGAGTTGGCCCGGGAGCTGGGCCTCAACGAGGCGCAGATCAAGATCTGG
 R  R  R  Q  E  L  A  R  E  L  G  L  N  E  A  Q  I  K  I  W
TTCCAGAACAAACGTGCAAAGATCAAGAAGGCGAGTGGTCAGAAG
 F  Q  N  K  R  A  K  I  K  K  A  S  G  Q  K
```

Figure 1. Generalized structure of an *en* class homeobox showing position of the diagnostic intron and the PCR primers used for intron detection. The translation shown is that of the *D. melanogaster* engrailed homeodomain.

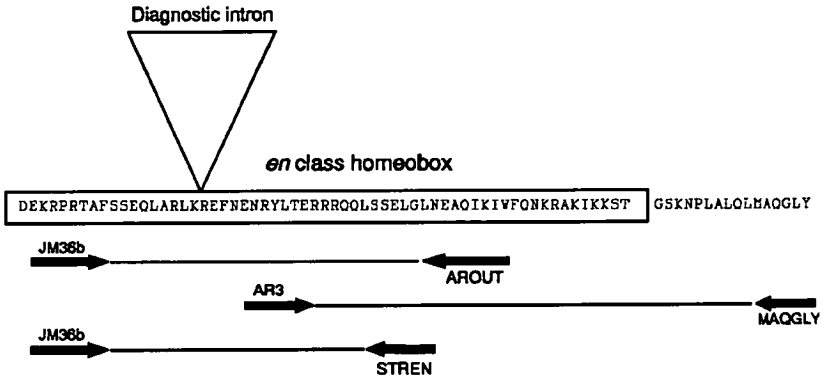


Figure 2. DNA and deduced protein sequences of the (A) strepsipteran and (B) orthopteran *en* class homeoboxes (excluding primer sequences). One of three strepsipteran specimens had a synonymous G to A substitution at position 72. GenBank accession numbers: AF130851–2.

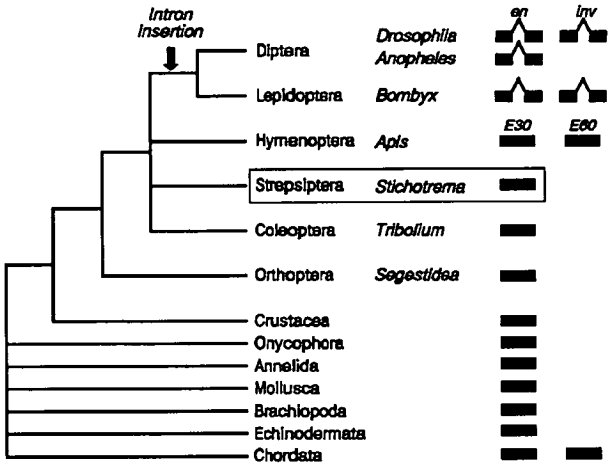


Figure 3. Phylogenetic tree indicating a shared intron insertion in *en* class homeoboxes of Diptera and Lepidoptera but not Strepsiptera, Coleoptera, Hymenoptera or outgroup taxa. Black boxes denote *en* class homeoboxes, with or without the diagnostic intron.

Lack of the intron is shared with Coleoptera, more basal insect orders and outgroups. The finding that the *en* homeobox of Strepsiptera lacks the diagnostic intron implies that the *en* gene adds no support to the suggestion that Strepsiptera are closely allied to Diptera. We suggest it is more likely that Strepsiptera occupy a more basal phylogenetic position within the Holometabola along with the Coleoptera, Hymenoptera and other non-Panorpoid insects (Fig. 3). If our interpretation is correct, then Strepsiptera are not homeotic Diptera. Instead, the common ancestor of Diptera and Strepsiptera did not possess halteres and these organs evolved from wings independently in the two orders.

The robustness of our interpretation depends on the reliability of intron insertion as a phylogenetic indicator. Notwithstanding the debate about the evolutionary origins of introns, it is clear that new introns have been inserted into genes during the course of animal evolution (O'Neill *et al.*, 1998; Tarrio *et al.*, 1998). Because insertion into a given gene is very rare, and intron position is likely to be selectively neutral, we argue that intron insertion events are effectively immune from convergent evolution. Identical intron positions must reflect common ancestry. The probability of intron loss is more difficult to estimate; intron loss has not yet been documented in the *en* class genes and we note it is rare in other gene families such as calmodulin (Côrte-Real *et al.*, 1994). However, possible loss of the *en* intron in the strepsipteran lineage cannot be totally excluded. Further sampling for the presence or absence of this intron within Strepsiptera and other insect orders will test our argument that the halteres of Strepsiptera and Diptera represent a remarkable case of convergent evolution and not a rare case of natural homeotic transformation.

Experimental procedures

Adult female specimens of the strepsipteran *Stichotrema dallatorreanum* Hofeneder and *Stichotrema* sp. (family Myrmecolacidae) were collected from Papua New Guinea, as were unparasitized specimens of their insect host, the long-horned grasshopper *Segestidea novaguineae* (Brancsik) (family Tettigonidae). All specimens were stored in 95% ethanol prior to DNA extraction. Total DNA was extracted using the QIAamp Blood Kit (QIAGEN cat. no. 29104; QIAGEN, Crawley, U.K.) following the supplier's protocol for insect DNA extraction. Oligonucleotide primers JM36b and AROUT were designed to amplify across the diagnostic intron site in *en* class homeobox genes:

JM36b (5′–3′) GAGAAGCGNCCACGNACNGCNTT
AROUT (5′–3′) GATCTTGATCTGCGCCTCGTTGAG

To control for potential amplification bias from intron insertion, primers AR3 and MAQGLY were also designed; these target a region adjacent to the diagnostic intron:

AR3 (5′–3′) GA(G,A)AA(C,T)CGNTATCTGACNGAG
MAQGLY (5′–3′) GTGGTTGTACAGNCC(C,T)TGNGCCAT

PCR reactions were performed using standard reaction concentrations (Holland, 1993) with the following cycling parameters: 94 °C for 2 min, followed by thirty-five cycles of 94 °C for 1 min, 55 °C for 1 min and 72 °C for 1 min, and a final extension stage

of 10 min at 72 °C. PCR fragments were purified from agarose gels, cloned into plasmid vectors, and multiple recombinant clones sequenced for each amplified band. After sequencing of the strepsipteran *en* class homeobox, a strepsipteran-specific primer (STREN) was designed to enable a test of clone authenticity.

STREN (5′–3′) TTCAGTCCCAACTCATT(G,A)CT

DNA sequences obtained in this study are available on GenBank (*Stichotrema dallatorreanum en* AF130851, *Segestidea novaguineae en* AF130852). Accession numbers for other *en* class genes to which comparison was made are: Diptera (*Drosophila melanogaster* K03055–8, *D. virilis* X04727, *Anopheles gambiae* U42214), Lepidoptera (*Bombyx mori* M64335–6), Hymenoptera (*Apis mellifera* M29489–90), Coleoptera (*Tribolium castaneum* S73225), Crustacea (U69098–105), Chelicerata (AF071404), Onychophora (AKY10771), Annelida (X58692, U26639), Mollusca (U23153–4, U23212–4, U23431–3, U21675, U21857), Brachiopoda (X62688), Echinodermata (U58775, M19709), Chordata (X59120–6, U82487).

Acknowledgements

We thank Rebecca Saff and Natalie Halbert for DNA from the two supplementary specimens of *Stichotrema* sp. Antonis Rokas was partially funded by the Greek Foundation of State Scholarships, through the SOCRATES link between the Universities of Crete and Reading. P.W.H.H. acknowledges a Royal Society grant. Specimens were collected by J.K. and staff of the Entomology Division, Papua New Guinea Oil Palm Research Association Inc., with special thanks to the Director, Mr Ian Orrell.

References

Balavoine, G. (1997) The early emergence of platyhelminths is contradicted by the agreement between 18S rRNA and Hox gene data. *Comptes Rendus de l'Academie des Sciences serie III de la Vie – Life Sci* **320**: 83–94.

Carmean, D. and Crespi, B.J. (1995) Do long branches attract flies? *Nature* **373**: 666.

Chalwatzis, N., Hauf, J., van de Peer, Y., Kinzelbach, R. and Zimmermann, R.K. (1996) 18S ribosomal RNA genes of insects: Primary structure of the genes and molecular phylogeny of the Holometabola. *Ann Entomol Soc Am* **89**: 788–903.

Côrte-Real, H.B.S.M., Dixon, D.R. and Holland, P.W.H. (1994) Intron-targeted PCR: a new approach to survey neutral DNA polymorphism in bivalve populations. *Mar Biol* **120**: 407–413.

Crowson, R.A. (1960) The phylogeny of Coleoptera. *Ann Rev Entomol* **5**: 111–134.

Felsenstein, J. (1978) Cases in which parsimony and compatibility methods will be positively misleading. *Syst Zool* **27**: 401–410.

Holland, P.W.H. and Garcia-Fernàndez, J. (1996) *Hox* genes and chordate evolution. *Dev Biol* **173**: 382–395.

Holland, P.W.H. (1993) Cloning genes using PCR. *Essential Developmental Biology: a Practical Approach* (Stern, C.D. and Holland, P.W.H., eds), pp. 243–255. IRL Press at Oxford University Press.

Huelsenbeck, J.P. (1997) Is the Felsenstein zone a fly trap? *Syst Biol* **46**: 69–74.

Huelsenbeck, J.P. (1998) Systematic bias in phylogenetic analysis: is the Strepsiptera problem solved? *Syst Biol* **47**: 519–537.

Hwang, U.W., Kim, W., Tautz, D. and Friedrich, M. (1998) Molecular phylogenetics at the Felsenstein zone: approaching the Strepsiptera problem using 5.8S and 28S rDNA sequences. *Mol Phylogen Evol* **9**: 470–480.

Kathirithamby, J. (1989) Review of the order Strepsiptera. *Syst Entomol* **14**: 41–92.

Kristensen, N.P. (1991) Phylogeny of extant hexapods. *Insects of Australia: a Textbook for Students and Research Workers* (Naumann, I.D., Carne, P.B., Lawrence, J.F., Nielsen, E.S., Spradberry, J.P., Taylor, R.W., Whitten, M.J. and Littlejohn, M.J., eds), pp. 125–140. 2nd edn. CSIRO and Melbourne University Press, Melbourne.

Kukalova-Peck, J. and Lawrence, J.F. (1993) Evolution of the hind wing in Coleoptera. *Can Entomol* **125**: 181–258.

Lewis, E.B. (1978) A gene complex controlling segmentation in *Drosophila*. *Nature* **276**: 565–570.

O'Neill, R.J.W., Brennan, F.E., Delbridge, M.L., Crozier, R.H. and Graves, J.A.M. (1998) *De novo* insertion of an intron into the mammalian sex determining gene, SRY. *Proc Natl Acad Sci USA* **95**: 1653–1657.

Pix, W., Nalbach, G. and Zeil, J. (1993) Strepsipteran forewings are haltere-like organs of equlibrium. *Naturwissenschaften* **80**: 371–374.

Raff, R.A.(1996). *The Shape of Life: Genes, Development and the Evolution of Animal Form.* University of Chicago Press, Chicago.

Tarrio, R., Rodriguez-Trelles, F. and Ayala, F.J. (1998) New *Drosophila* introns originate by duplication. *Proc Natl Acad Sci USA* **95**: 1658–1662.

Whiting, M.F. and Wheeler, W.C. (1994) Insect homeotic transformation. *Nature* **368**: 696.

Whiting, M.F., Carpenter, J.C., Wheeler, Q.D. and Wheeler, W.C. (1997) The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from the 18S and 28S ribosomal sequences and morphology. *Syst Biol* **46**: 1–68.

polished my prose. This is paper 99–114 of the Institut des Sciences de l'Evolution.

François Rousset

Institut des Sciences de l'Evolution, CC065 ISEM USTL, Pl. E. Bataillon, 34095 Montpellier Cedex, France (rousset@isem.univ-montp2.fr)

References

1 Nagylaki, T. (1989) Gustave Malécot and the transition from classical to modern population genetics. *Genetics* 122, 253–268

2 Malécot, G. (1951) Un traitement stochastique des problèmes linéaires (mutation, linkage, migration) en génétique de population. *Annales de l'Université de Lyon A* 14, 79–117

3 Malécot, G. (1999) Statistical methods and the subjective basis of scientific knowledge. *Gen. Sel. Evol.* 31, 269–298 (translation by D. Gianola)

4 Jacquard, A. (1975) Inbreeding: one word, several meanings. *Theor. Popul. Biol.* 7, 338–363

5 Crow, J.F. (1987) Population genetics history: a personal view. *Annu. Rev. Genet.* 21, 1–22

Wolbachia as a speciation agent

Bacteria of the genus *Wolbachia* are reproductive parasites of arthropods. They are cytoplasmically inherited (i.e. from mothers to daughters, like mitochondria) and their phenotypic effects on their hosts range from induction of parthenogenesis in certain hymenopteran groups, to feminization of genetic males in isopod crustaceans and to induction of cytoplasmic incompatibility in many insects[1] (Box 1). Researchers have envisioned *Wolbachia* playing a role in insect speciation by generating reproductive isolation[2–4]. However, theory predicts that *Wolbachia* might not have a major role in the process of speciation[5]. A recent paper by Shoemaker *et al.*[6] demonstrates experimentally how *Wolbachia* might act as a speciation enhancer, thus bridging the gap between theoretical predictions and experimental results.

But how can *Wolbachia* drive speciation? Three ways have been suggested[4]. The first suggestion is induction of parthenogenesis (assuming that the infection becomes fixed in the species). This process appears to occur in the hymenopteran *Encarsia formosa*[7]; males in this species can be obtained when mothers are treated with antibiotics, however, they are incapable of mating. The second way is offered by bidirectional incompatibility (Box 1). If a population is infected with two different strains of *Wolbachia* that are incompatible with each other, then bidirectional incompatibility will act as a post-zygotic reproductive barrier. A promising candidate is the species complex of the wasp genus *Nasonia*[2,3]. The third way is offered by unidirectional incompatibility (Box 1). In this case, *Wolbachia* is acting as just one of the reproductive barriers enhancing speciation between two taxa. It is in this area of *Wolbachia* research that Shoemaker *et al.*[6] offer a new perspective, providing the first experimental data.

The work focuses on two closely related species of *Drosophila*, *D. recens* and *D. subquinaria*, which belong to the *subquinaria* species group. Both are ecologically similar, live in the cooler regions of North America, and use mushrooms as mating and ovipositing sites. *D. recens* is distributed across the northeastern USA and southeastern Canada, whereas *D. subquinaria* spreads across the northwestern USA and southwestern Canada. They live in sympatry over a broad area in central Northern USA. The principal finding is that *Wolbachia*-induced unidirectional incompatibility, coupled with a behavioural isolation mechanism acting in the reciprocal direction, establishes a reproductive barrier between the two *Drosophila* species.

When Shoemaker *et al.*[6] crossed *D. recens* and *D. subquinaria*, they revealed an asymmetrical behavioural isolation mechanism. Although *D. recens* females will not mate with *D. subquinaria* males (Table 1), *D. recens* males will mate with *D. subquinaria* females. In fact, there was no significant difference between the latter interspecific cross and the pairing probabilities within either species.

In a screening for *Wolbachia* infection, *D. recens* was found to be positive, whereas *D. subquinaria* was not infected. To investigate the role of *Wolbachia* in crosses involving *D. recens*, the number of offspring obtained from matings between infected *D. recens* males and uninfected *D. subquinaria* females, was compared with the number of progeny obtained from matings between *D. recens* males that were treated with antibiotics (and thus *Wolbachia*-negative) and *D. subquinaria* females. The number of progeny in the first cross was significantly lower (Table 1), suggesting a *Wolbachia*-mediated barrier to this cross. No statistically significant difference in the number of offspring was observed for all the other control crosses.

However, a possible involvement of other post-zygotic factors restricting gene flow has not yet been discarded. Two post-zygotic factors are most important as reproductive barriers, namely hybrid sterility and hybrid breakdown. Hybrid sterility was tested by crossing hybrid males or females with either of the parental species. The results suggested that hybrid males were invariably sterile, whereas hybrid females were fertile (Table 1). This is in agreement with Haldane's rule[8] (i.e. if a partial reproductive barrier exists this should affect the heterogametic sex most severely) and suggests the presence of an incomplete post-zygotic isolation barrier.

Hybrid breakdown is the inviability or sterility of progeny resulting from a backcross of hybrid progeny with either of the

Box 1. Cytoplasmic incompatibility: definitions and theoretical predictions

Cytoplasmic incompatibility: can be defined as the failure of a cross to produce any offspring (or only male offspring in haplodiploid species) because of cytoplasmic factors. It can either be unidirectional or bidirectional[3].

Unidirectional incompatibility: the cross between a *Wolbachia*-infected male and an uninfected female fails to produce any progeny whereas the reciprocal cross (uninfected male X infected female) is normal. Theory suggests that natural selection (on the parasites) acts on increasing the fecundity of the infected females[5,9], rather than increasing the incompatibility levels between infected versus uninfected. The consequence is that strains with lower incompatibility levels, but with a lower cost in the fecundity of the host, might be selected for.

Bidirectional incompatibility: the cross between two individuals infected with different strains of *Wolbachia* fails to produce any progeny. Theory predicts that natural selection (on the parasites) will act towards establishing a single infection in populations that are infected with two different *Wolbachia* strains, if sufficient gene flow exists between the two populations[5]. In the example discussed in the text (*Nasonia* species complex), it is unclear whether infection with different *Wolbachia* strains preceded, and thus drove, speciation or followed species formation.

 0169-5347/00/$ – see front matter PII: S0169-5347(99)01783-8

parental species. Hybrid progeny have a set of chromosomes from each parental species. Progeny resulting from a cross of one of these hybrids with either of the parental species do not, and thus a 'hybrid breakdown' might occur leading to sterility and/or inviability. Backcross hybrids were mated with either of the parental species to test whether there was any hybrid breakdown effect. Most of the male backcross hybrids were sterile – such as the F_1 male hybrids – whereas the female backcross hybrids did not show any significant reduction in the number of progeny they produced, thus suggesting an absence of 'breakdown' effects (Table 1).

Table 1. Synopsis of the most important crosses performed

Cross no.	Male	Female	Observation tested	Explanation of observation
1	*subquinaria*	*recens*	Mating frequency	Behavioural isolation
2	*recens*^{W}	*subquinaria*^{O}	Offspring production	Cytoplasmic incompatibility
3	F_1 hybrid	Either parental species	Offspring production	Hybrid male sterility (Haldane's rule)
4	Either parental species	F_1 hybrid	Offspring production	No female sterility effect
5	Either parental species	Backcross or F_2 hybrid (offspring of cross no. 4)	Offspring production	No hybrid breakdown effect

Key: *recens*^{W} = *Wolbachia*-infected *D. recens*; *subquinaria*^{O} = uninfected *D. subquinaria*. Details from Ref. 6.

Shoemaker *et al.*[6] suggest that this might be one of the ways *Wolbachia* acts as a speciation agent. Coupled with one pre-zygotic barrier (unidirectional behavioural isolation) and another post-zygotic barrier (hybrid male sterility), *Wolbachia* acts as a reproductive barrier (and, thus, a speciation inducer) between the two species. This is the first explicit demonstration of how a *Wolbachia* infection, causing unidirectional incompatibility, contributes to the establishment of a reproductive barrier between the two species.

Although theory suggests that *Wolbachia* parasites themselves might not be important as speciation agents, these data suggest a possible role for *Wolbachia* as one of the many reproductive barriers between two incipient species. Therefore, rather than envisioning *Wolbachia* as the sole driver of the whole speciation process, Shoemaker *et al.*[6] suggest that *Wolbachia* can act as one of the speciation agents. One particular role for *Wolbachia* might be in reducing gene flow between geographically separated, genetically divergent populations, which subsequently overlap before reproductive barriers are complete. The data offered by Shoemaker *et al.*[6] suggest that *Wolbachia* might be a reinforcement mechanism in such situations.

Some comments should be made on the sampling strategy used by Shoemaker *et al.*[6], which perhaps reduces the force of the paper's arguments. Species samples were not taken from places where the two species occur in sympatry, but from sites on opposite sides of the USA. Therefore, we do not actually know whether behavioural isolation and hybrid sterility occur in the zone of sympatry with the same intensity as in the tests described for (highly) allopatric populations. Given that the behavioural isolation mechanism is not perfect (some matings do occur), it is at least possible that the frequency of matings, in the cross in which behavioural isolation is observed, will decrease as we move from the allopatric zone to the sympatric zone of the two species. More careful sampling would have provided a stronger case for what the authors suggest.

Notwithstanding this argument, the first step has already been taken towards elucidating *Wolbachia*'s role (when the effect is unidirectional incompatibility) in speciation events. The right questions have been asked and the first evidence is certainly promising. It should not take long before new, and more complete, studies are published. *Wolbachia*'s share of the explanation of what Darwin called 'the mystery of the mysteries' has slowly started to emerge. We should certainly expect more wonders from this amazing reproductive parasite.

Acknowledgements

I thank Graham Stone for valuable comments on this article.

Antonis Rokas

Institute of Cell, Animal and Population Biology, King's Buildings, University of Edinburgh, Edinburgh, UK EH9 3JT (a.rokas@ed.ac.uk)

References

1 O'Neill, S.L. *et al.*, eds (1997) *Influential Passengers: Inherited Microorganisms and Arthropod Reproduction*, Oxford University Press

2 Breeuwer, J.A.J. and Werren, J.H. (1990) Microorganisms associated with chromosome destruction and reproductive isolation between two insect species. *Nature* 346, 558–560

3 Werren, J.H. (1997) Biology of *Wolbachia*. *Annu. Rev. Entomol.* 42, 587–609

4 Werren, J.H. (1997) *Wolbachia* and speciation. In *Endless Forms: Species and Speciation* (Howard, D. and Berlocher, S., eds), pp. 245–260, Oxford University Press

5 Turelli, M. (1994) Evolution of incompatibility-inducing microbes and their hosts. *Evolution* 48, 1500–1513

6 Shoemaker, D.D. *et al.* (1999) *Wolbachia* and the evolution of reproductive isolation between *Drosophila recens* and *Drosophila subquinaria*. *Evolution* 53, 1157–1164

7 Zchori-Fein, E. *et al.* (1992) Male production by antibiotic treatment in *Encarsia formosa* (Hymenoptera: Aphelinidae), an asexual species. *Experientia* 48, 102–105

8 Turelli, M. (1998) The causes of Haldane's rule. *Science* 282, 889–891

9 Prout, T. (1994) Some evolutionary possibilities for a microbe that causes incompatibility in its host. *Evolution* 48, 909–911

Letters to *TREE*

Correspondence in *TREE* may address topics raised in very recent issues of *TREE*, or (occasionally) other matters of general current interest to ecologists and evolutionary biologists. Letters should be **no more than 500 words long with a maximum of 12 references and one small figure**; original results, new data or new models are not allowed. Letters should be sent by e-mail to TREE@current-trends.com. The decision to publish rests with the Editor, and the author(s) of any *TREE* article criticized in a Letter will normally be invited to reply. Full-length manuscripts in response to previous *TREE* articles will not be considered.

manner similar to the method described in (5). Briefly, Z-domain was immobilized on microtiter wells at a concentration of 5 μg/ml, blocked, and washed as described. A matrix of mixtures of biotin-IgG-Fc (312 to 0.3 nM) and peptide (215 μM to 0.8 nM) were prepared. These mixtures were incubated with immobilized Z-domain for 1 hour. Plates were then washed and developed as described with avidin–horseradish peroxidase conjugate. Inhibition curves were then computed for each concentration of biotin-IgG-Fc, and the curve of half-maximal inhibition was extrapolated to zero biotin-IgG-Fc concentration to obtain a K_i.

11. The DNA sequence of the peptide was moved to a monovalent phage display format by cassette mutagenesis to give a construct with the STII signal sequence, the peptide KEASCSYWLGELVWCVAGVE, a GGGPGGG linker, and the M13 gene III protein starting at residue 253.
12. A series of second-generation monovalent phage display libraries were constructed based on the sequence KEASCSYWLGELVWCVAGVE, in which five sequential residues were randomized by using NNS codons in each library starting at positions 1, 4, 7,10, 12, and 16, excluding the two cysteines. Each library had a diversity of $\sim 1 \times 10^8$. These libraries were independently screened for binding to IgG-Fc for six rounds and then sequenced.
13. Three additional libraries were constructed by using the degeneracy of the genetic code to recombine the preferred amino acids at each position into one peptide. The DNA sequences for these libraries contained the following mixtures of bases (IUPAC codes): DRG GWA GMA RRC TGC KCT TRS CAC MTG GGC GAG CTG GTC TGG TGC RVC RVM BKC GAS KDW, DRS VWG SVG RRC TGC KCC TRS YRS MTG GGC GAG CTG GTC TGG TGC RNC VVS NBS GWS KDM, and DNS NNS NNS VNS TGC BVG TDS HRS MDS GGC GAG STC KKG WRG TGC RNM NNS NNS NNS NNM. These libraries were also sorted against IgG-Fc for six rounds and then sequenced.
14. Inhibition assays were performed as described (10) at pH 7.2 and at pH 6.0. The peptide was found to inhibit fourfold more tightly at the lower pH. Kinetic and steady-state binding to immobilized IgG_1 was also measured directly by BIAcore (Pharmacia), giving $K_{on} = 1.6 \times 10^6\ M^{-1}\ s^{-1}$, $K_{off} = 2.5 \times 10^{-2}\ s^{-1}$, and K_d = 16 nM in 25 mM MES (pH 6.0), 0.05% Tween-20.
15. S. R. Fahnestock *et al.*, in *Bacterial Immunoglobulin-Binding Proteins* (Academic Press, New York, 1990), vol. 1, chap. 11; R. Karlsson, L. Jendeberg, B. Nilsson, J. Nilsson, P. Nygren, *J. Immunol. Methods* **183**, 43 (1995).
16. Crystals were grown in 100 mM NaOAc (pH 6.0), 20% polyethylene glycol 4000, and 20% isopropanol by vapor diffusion from 4-μl drops containing 100 μM IgG-Fc, up to 150 μM peptide, and a 50% dilution of reservoir solution. Data were collected to 2.6 Å at the Stanford Synchrotron Radiation Laboratory (SSRL) and were reduced with DENZO [W. Minor and Z. Otwinowski, *Methods Enzymol.* **176**, 307 (1997)]. Phasing was accomplished by molecular replacement with AmoRE [J. Navaza, *Acta Crystallogr.* **A50**, 157 (1994)], with an IgG-Fc subunit derived from Deisenhofer *et al.* (3) as a search model. The crystal contained one Fc dimer and two peptide molecules per asymmetric unit. The structure was refined with X-PLOR 3.1 [A. T. Brünger *et al.*, *Science* **245**, 458 (1987)], with noncrystallographic restraints on the Fc dimer over regions >10 Å away from nonequivalent crystal contacts. The final dimeric Fc model consisted of IgG_1 residues 237 to 443 with eight sugars per monomer.
17. Surface area and geometric measurements were made with the Crystallography and NMR System (CNS) [A. T. Brünger *et al.*, *Acta Crystallogr. D.* **54**, 905 (1998)]. A solvent probe radius of 1.4 Å was used, and surface area changes were computed by subtracting complexed from uncomplexed solvent-accessible surface areas. Contact regions were defined as the set of atoms that lie within 5.0 Å of any nonhydrogen atom on the opposing molecule.
18. Protein A domain numbering is from H. Gouda *et al.*, *Biochemistry* **31**, 9665 (1992).
19. The computer program SITEFINDER (WLD) was used to generate 2.5 million patches of contiguous surface atoms having solvent-accessible surface areas of 525 Å^2. Patches were randomly distributed across all of the available structures (PDB codes: 1FC1, 1FC2, 1FCC, 1ADQ, and 1DN2) and were of a random globular shape. To ensure even sampling, probabilities were weighted so that each solvent-exposed atom was included in an equal number of surface patches (~10,000 patches per atom) The properties of each site were computed and then compared with those of the consensus binding patch on Fc.
20. L. Young, R. L. Jernigan, D. G. Covell, *Protein Sci.* **3**, 717 (1994); S. Jones and J. M. Thornton, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13 (1996); C. Tsai, S. L. Lin, H. J. Wolfson, R. Nussinov, *Protein Sci.* **6**, 53 (1997); L. L. Conte, C. Chothia, J. Janin, *J. Mol. Biol.* **285**, 2177 (1999).
21. Patches from (19) were ranked separately by polarity and solvent-accessible surface fraction. For each atom in the Fc dimer, the average rank of all patches involving the atom was then computed. The average atomic ranks for polarity and accessibility were then combined linearly (⟨accessibility⟩ − ⟨polarity⟩) to give a composite score incorporating both properties.
22. C. Medesan, D. Matesoi, C. Radu, V. Ghetie, E. S. Ward, *J. Immunol.* **158**, 2211 (1997).
23. W. L. DeLano, in preparation.
24. Single alanine mutants of the peptide and Protein A Z-domain were displayed monovalently on gene III, assayed by enzyme-linked immunosorbent assay, and directly compared with unmutated control peptide as described [B. C. Cunningham, D. G. Lowe, B. Li, B. D. Bennett, J. A. Wells, *EMBO J.* **13**, 2508 (1994)]. Assays were performed under conditions where EC_{50}(wt)/EC_{50}(mut) will approximate K_d(wt)/K_d(mut) (EC_{50}, median effective concentration). Additional mutagenesis data are available in L. Jendeberg *et al.*, *J. Mol. Recog.* **8**, 270 (1995).
25. B. K. Kay, A. V. Kurakin, R. Hyde-DeRuysher, *Drug Discovery Today* **8**, 370 (1998).
26. We thank B. C. Cunningham, J. K. Tong, and M. Dennis for assistance in the initial selection experiments against Fc; A. Braisted for training in solid phase peptide synthesis; C. Wiesmann for help with crystallographic refinement; and the SSRL for use of their facility in data collection

19 August 1999; accepted 27 December 2000

Evidence for a High Frequency of Simultaneous Double-Nucleotide Substitutions

Michalis Averof,[1]* Antonis Rokas,[2] Kenneth H. Wolfe,[3] Paul M. Sharp[4]*

Point mutations are generally assumed to involve changes of single nucleotides. Nevertheless, the nature and known mechanisms of mutation do not exclude the possibility that several adjacent nucleotides may change simultaneously in a single mutational event. Two independent approaches are used here to estimate the frequency of simultaneous double-nucleotide substitutions. The first examines switches between TCN and AGY (where N is any nucleotide and Y is a pyrimidine) codons encoding absolutely conserved serine residues in a number of proteins from diverse organisms. The second reveals double-nucleotide substitutions in primate noncoding sequences. These two complementary approaches provide similar high estimates for the rate of doublet substitutions, on the order of 0.1 per site per billion years.

Mutational events can be studied either by direct observation of mutations in the laboratory or by comparing sequences that have been accumulating mutations naturally, during evolution. Studies of the first kind have suggested that some mutations can involve multiple nucleotide changes (*1*, *2*), and indeed, mechanisms that affect neighboring nucleotides are known. Examples include template-directed mutations occurring during DNA repair and replication (*1*) or dipyrimidine lesions induced by ultraviolet light (*2*, *3*). Some evolutionary comparisons have also suggested that simultaneous double-nucleotide substitutions occur at neighboring sites (*4*), but the significance and generality of these observations have been questioned (*5*). Thus, changes in neighboring nucleotides are usually attributed to coincidence of independent mutations.

We used two independent and complementary approaches based on sequence comparisons to study double-nucleotide substitutions and to obtain estimates of their frequency. The first approach examined changes that have occurred over long evolutionary time scales, between two particular dinucleotides, TC and AG. Serine is unique among amino acids in that it is encoded by two groups of codons, TCN and AGY, which cannot be interconverted by a single-nucleotide mutation. Switches between these groups of codons could occur indirectly, by two separate single-nucleotide mutations

[1]Institute of Molecular Biology and Biotechnology (IMBB)–FORTH, Vassilika Vouton, 711 10 Iraklio, Crete, Greece. [2]Institute of Cell, Animal and Population Biology, University of Edinburgh, King's Buildings, West Mains Road, Edinburgh, EH9 3JT, UK. [3]Department of Genetics, Trinity College, University of Dublin, Dublin 2, Ireland. [4]Institute of Genetics, University of Nottingham, Queens Medical Centre, Nottingham NG7 2UH, UK.

*To whom correspondence should be addressed. E-mail: averof@imbb.forth.gr (M.A.) or paul@evol.nott.ac.uk (P.M.S.)

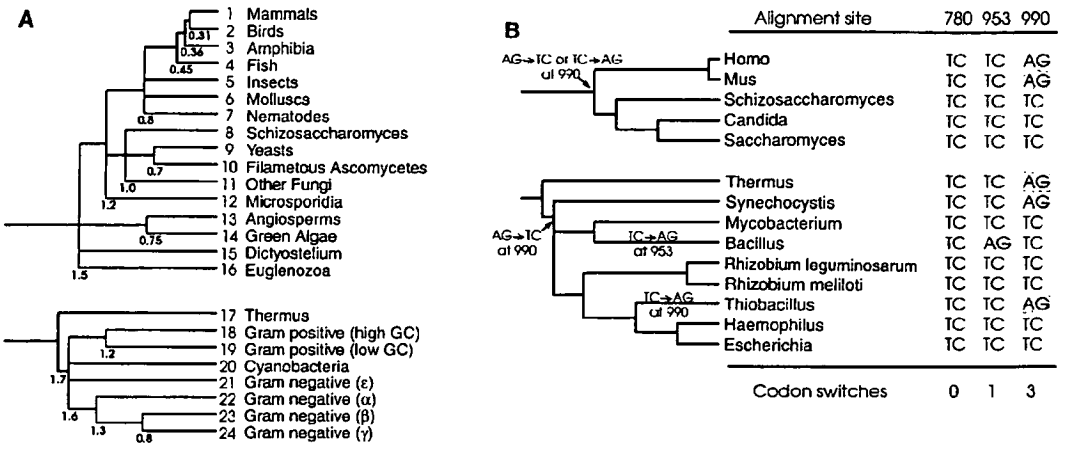


Fig. 1. (**A**) Overview of the phylogeny and divergence times used for the analysis of serine codon switches. The phylogeny is based on a number of recent phylogenetic analyses (*20, 24*), with points of uncertainty shown as unresolved polychotomies. Times of common ancestors are indicated in Gyr before present. (**B**) Determination of serine codon switches. The data set of glutamine fructose-6-phosphate transaminase is shown as an example. There are three sites where serine is absolutely conserved in the protein sequence (alignment sites 780, 953, and 990). At least four codon switches can be observed. The time sampled by this data set (sum of branch lengths) is 3 × 14.71 Gyr.

(TC↔AC↔AG or TC↔TG↔AG), or perhaps directly by simultaneous double-nucleotide mutation (TC↔AG). In the former case, the switch would involve an intermediate step whereby the triplet would encode either threonine (ACN) or cysteine (TGY), residues that are ionically and sterically different from serine (*6*), so such changes are unlikely to be tolerated in critical functional or structural sites of a protein. Nevertheless, TCN↔AGY switches have been observed at sites encoding extremely conserved serine residues, for example in ubiquitin (*7*) and in the active site of serine proteases (*8*). Switches at these sites seem most likely to result from simultaneous double-nucleotide mutations, which in this context are synonymous and most likely selectively neutral.

To investigate the generality and frequency of such switches, we studied 23 data sets of homologous proteins containing serine residues absolutely conserved over a wide range of eukaryotes and/or prokaryotes (Fig. 1A). We analyzed the distribution of TCN and AGY codon types in these conserved serines, inferring the position and frequency of codon switches during evolution (illustrated in Fig. 1B) (*9*). Our analysis reveals a widespread occurrence of codon switches at such sites (Table 1), with an estimated frequency of about 0.1 per site per billion years (94/774 = 0.12 per site per Gyr). This rate appears to be consistent among different phylogenetic lineages and different genes (Fig. 2). Rate estimates from bacteria and eukaryotes are very similar, 0.11 and 0.12 per site per billion years (Gyr), respectively.

Of the 70 switches where the direction of change could be inferred (by parsimony and with reference to outgroups), 60 were in the TC→AG rather than the AG→TC direction. However, independent rate estimates for each direction are very similar, 0.10 and 0.11 per site per Gyr, respectively. The bias therefore reflects a preponderance of TCN-type codons as potential targets, rather than a bias in the direction of mutation [this points to a strong codon bias in the ancestral representation of serines (*10*)].

Most codon switches at such highly conserved serines appear to result from simultaneous double-nucleotide mutations. However,

Table 1. Rates of serine codon switches in 23 data sets of highly conserved proteins. The phylogenetic assemblages (species) represented in each data set are indicated by numbers as specified in Fig. 1A. The inferred number of codon switches and estimated time sampled by each data set (in Gyr) are indicated.

Protein	Species	Switches	Time
Ribosomal protein S7	1,3,4,5,8,9	1	5.51
Ribosomal protein S11	1,3,5,8,9,13,14	2	22.62
Ribosomal protein S12	1,3,5,7,9,16	1	6.36
Ribosomal protein S17	1,2,5,9,10,13,15	0	15.49
Arginosuccinate synthetase	1,9,18,24	3	18.12
Glycine dehydrogenase	1,2,8,9,13	3	28.20
Glutamine fructose-6-phosphate transaminase	1,8,9,17,18,19,20,22,23,24	4	44.13
2-oxaloglutarate dehydrogenase	1,9,23,24	5	32.35
Asparagine synthase	1,9,13	2	38.08
Adenylosuccinate synthase	1,7,8,9,13,15,19,20,22,23,24	2	17.11
dUTP pyrophosphatase	1,9,13,24	0	5.71
Uridine-5-diphosphate glucose-4-epimerase	1,9,18,19,22,23,24	5	32.32
Phosphoenol pyruvate carboxykinase	1,2,5,7,11	7	55.40
Arginosuccinate lyase	1,2,3,8,9,14,18,21,24	9	63.05
1,4-α-glycan branching enzyme	1,9,13,18,19,20,24	6	28.92
Histidine tRNA synthetase	1,4,7,9	2	16.08
Tryptophanyl tRNA synthetase	1,8,9	2	9.38
Ribonucleotide reductase (large subunit)	1,5,7,8,9,12	16	123.80
Fumarate hydratase	1,9,11	3	21.06
Aspartate ammonia lyase	18,19,23,24	3	46.40
DNA topoisomerase 2	1,5,8,9,13,16	5	51.60
Dimethylallyl transferase	1,9,10,13	1	18.21
Ribonucleotide reductase (small subunit)	1,5,6,7,8,9,12,13,15	12	74.10
Total		94	774.00

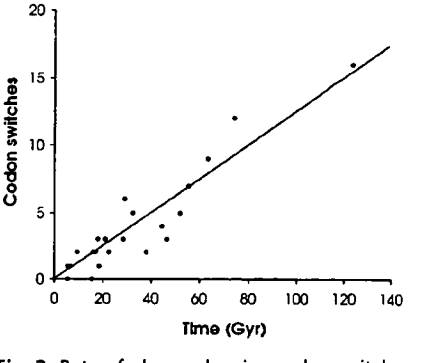


Fig. 2. Rate of observed serine codon switches for 23 proteins. Data is from Table 1. The line has a slope of 0.12 switches per site per Gyr.

Table 2. Analysis of single- and double-nucleotide substitutions in the pseudo eta globin locus on each branch of the catarrhine primate phylogeny (Fig. 3). Positions of substitutions were inferred by parsimony. L, number of aligned nucleotides; ObsS, ObsD, numbers of changes observed as single- or double-nucleotide substitutions, respectively; ExpD, number of doublet substitutions expected by coincidence of two separate single-nucleotide substitutions; RealD, number of excess double changes, inferred to have occurred as simultaneous double-nucleotide substitutions.

Tree branch	L	ObsS	ObsD	ExpD	RealD
Node 1–rhesus monkey	6617	284	19	10.83	8.17
Node 1–gibbon	6996	129	4	2.25	1.75
Node 1–node 2	7187	39	1	0.19	0.81
Node 2–orangutan	6974	92	2	1.17	0.83
Node 2–node 3	7187	46	1	0.28	0.72
Node 3–gorilla	7055	41	0	0.24	−0.24
Node 3–node 4	7187	7	0	0.01	−0.01
Node 4–human	6997	26	0	0.10	−0.10
Node 4–node 5	7187	34	0	0.16	−0.16
Node 5–chimp	7024	23	2	0.05	1.95
Node 5–pigmy chimp	7003	11	1	0.01	0.99
Total		732	30	15.30	14.70

it is conceivable that these switches could occur by two separate single-nucleotide mutations, through intermediates that encode threonine or cysteine. Kimura suggested that slightly deleterious intermediates may sometimes survive to be rescued by rapidly selected compensatory mutations (*11*), but there are a number of observations that argue against this possibility in this case. First, Kimura's model applies to situations where compensatory mutations are relatively frequent (e.g., when many different mutations can have a compensatory effect) or when the selective coefficient against the intermediates is rather low, which seem very unlikely. Second, if deleterious alleles were involved, we would expect these to survive much more frequently in the presence of additional copies of the gene, but we observe very similar rates of codon switches in haploid and diploid genomes, as well as in proteins that belong to multigene families (*12*). Moreover, we have also noticed TCN↔AGY switches among codons encoding highly conserved serines in closely related sequences, with no evidence of a transition through nonserine intermediates (*13*).

Other mechanisms have also been proposed that could explain switches in serine codons through nondeleterious intermediates (*8, 14–16*). For example, a transient substitution of serine by another amino acid could be complemented by the presence of a neighboring serine residue (*16*), an alternative genetic code may have allowed TGN to encode serine (*15*), or the two types of serine codon may reflect independent origins from a different ancestral amino acid (*8*). These explanations may apply in special cases and could contribute to a small proportion of codon switches. However, they are unlikely to account for the widespread distribution of codon switches, as observed in diverse phylogenetic lineages, in different proteins, and in serine residues whose position and identity has been absolutely conserved.

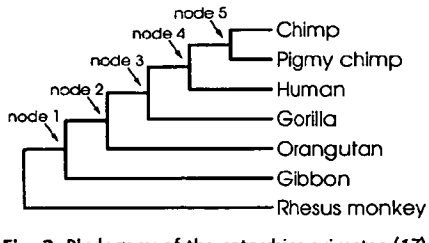


Fig. 3. Phylogeny of the catarrhine primates (*17*) used in the analysis of pseudo eta globin sequences (Table 2). Branch lengths are not to scale.

In our second approach, we examined double-nucleotide substitutions among noncoding sequences of closely related species. In these sequences, substitutions are expected to accumulate in a manner that is unbiased by selection, and so directly reflect mutational processes. We compared a long (about 7 kb) noncoding sequence from the pseudo eta globin locus of seven closely related catarrhine primates (Fig. 3) (*17*) to determine whether mutations in that region involve a significant fraction of clustered nucleotide changes (*18*). Using parsimony analysis, we determined the number of single- and double-nucleotide changes that have occurred during the evolution of these species and found a significant excess of double-nucleotide substitutions relative to what would be expected by coincidence of single-nucleotide changes alone (Table 2). The excess, apparently simultaneous, dinucleotide mutations are estimated to have occurred at a rate of 0.1 per site per Gyr (*19*), on average, at any nucleotide doublet.

These two analyses are complementary: they examine double-nucleotide substitutions in different contexts and over very different time scales. Any concerns that the serine codon switches might have involved compensatory changes via nonserine intermediates are offset by the observation of similarly high levels of doublet changes in closely related noncoding sequences. Equally, although the rates for all dinucleotide changes were estimated from just one particular region of the primate genome, the rates of TC↔AG changes estimated from serine switches apply to a wide range of loci from diverse organisms. Both approaches point to the conclusion that the rate of double-nucleotide substitutions is high compared to expectations based on the coincidence of individual neutral nucleotide substitutions, which typically occur at a rate of around 1 to 10 per site per Gyr (*20, 21*).

We expect that the rates of different doublet mutations will vary considerably depending on a cell's exposure to different mutational mechanisms. For example, we would expect to see a much higher incidence of dipyrimidine lesions in cells that are exposed to ultraviolet light (e.g., exposed unicellular organisms, skin cells) than in cells that are not (e.g., the germ line of large multicellular animals). Such differences might explain why the estimated frequency of specific TC→AG and AG→TC substitutions in serine codons, which may involve dipyrimidines (TC in the coding strand or CT in the noncoding strand, respectively), is higher than would be predicted by the average frequency of double-nucleotide substitution estimated from the eta globin pseudogene. The sequence-specificity of mutational mechanisms could result in different rates of substitution among various doublets in different cell types. These observations may be important in the context of models of molecular evolution and phylogenetic reconstruction, as well as mutational mechanisms of human disease.

References and Notes

1. G. B. Golding and B. W. Glickman, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 8577 (1985).
2. D. M. Hampsey, J. F. Ernst, J. W. Stewart, F. J. Sherman, *J. Mol. Biol.* **201**, 471 (1988).
3. H. Nakazawa *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 360 (1994).
4. K. H. Wolfe and P. M. Sharp, *J. Mol. Evol.* **37**, 441 (1993).
5. D. Mouchiroud, C. Gautier, G. Bernardi, *J. Mol. Evol.* **40**, 107 (1995); J. M. Comeron and M. Kreitman, *Genetics* **150**, 767 (1998); N. G. C. Smith and L. D. Hurst, *Genetics* **153**, 1395 (1999).
6. R. Grantham, *Science* **185**, 862 (1974).
7. P. M. Sharp and W.-H. Li, *J. Mol. Evol.* **25**, 58 (1987).
8. S. Brenner, *Nature* **334**, 528 (1988).
9. We collected data sets of homologous proteins showing wide phylogenetic conservation (representing diverse eukaryotic and eubacterial lineages) and unambiguous relationships (excluding multigene families, cases of horizontal transfer, concerted evolution). For each data set, conserved protein sequences were obtained from the SWISS-PROT database (release 8/98) using BLAST (*22*). Protein sequences were aligned using CLUSTALW (*23*) and were searched for unambiguously aligned sites where serine is absolutely conserved (i.e., present in all available sequences). The corresponding codons were determined from the respective nucleotide sequences, obtained from the GenBank/EMBL (European Molecular Biology Laboratory) database. Changes in serine codon type were determined in the most parsimonious way on the basis of phylogenies (illustrated Fig. 1B), and rates of change were estimated as the number of inferred changes over the time sampled at each site (the sum of all branch lengths). Phylogenetic relationships and times of divergence were based on published data for the respective species (*20, 24*) (Fig. 1A). Trees were also constructed from the protein sequences themselves [using the Neighbor-Joining method (*25*)], and sequences showing an inconsistent phylogenetic placement were eliminated. Because of difficulties in determining

their times of divergence, eubacterial and eukaryotic sequences were treated separately and archaebacterial sequences were excluded. In estimating the rates of codon switches, we tried to be conservative, for example, by overestimating times of divergence in cases of uncertainty. The analysis was also carried out on the basis of alternative published phylogenies and the results were always robust (M. Averof *et al.*, data not shown).

10. Y. Diaz-Lazcoz, A. Henaut, P. Vigier, J. L. Risler, *J. Mol. Biol.* **250**, 123 (1995).
11. M. Kimura, in *Population Genetics and Molecular Evolution*, T. Ohta and K. Aoki, Eds. (Springer-Verlag, Berlin, 1985), pp. 19–39.
12. M. Averof and P. M. Sharp, data not shown.
13. Ubiquitin, one of the most highly conserved proteins known, is encoded by repeats sharing recent common ancestry due to concerted evolution (*7*); serine switches among repeats have been reported [P. M. Sharp and W.-H. Li, *Trends Ecol. Evol.* **2**, 328 (1987); P. M. Sharp, M. Averof, A. T. Lloyd, G. Matassi, J. F. Peden, *Philos. Trans. R. Soc. London Ser. B* **349**, 241 (1995)]. We have also found serine switches in orthologous genes from the two strains of *Helicobacter pylori* for which complete genome sequences have been determined [J.-F. Tomb *et al.*, *Nature* **388**, 539 (1997); R. A. Alm *et al.*, *Nature* **397**, 176 (1999)]. Switches were seen at sites where serine is conserved in a range of other species.
14. D. M. Irwin, *Nature* **336**, 429 (1988).
15. P. S. Goldfarb, *Nature* **336**, 429 (1988).
16. E. V. Koonin and A. E. Gorbalenya, *Nature* **338**, 467 (1989).
17. M. M. Miyamoto, B. F. Koop, J. L. Slightom, M. Goodman, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 7627 (1988); W. J. Bailey *et al.*, *Mol. Phylogenet. Evol.* **1**, 97 (1992).
18. Sequences from the pseudo eta globin region of seven catarrhine primates were collected from the EMBL database (accession numbers U01317, K02542, X79724, K02543, M18038, M54985, J03818). The sequences were aligned using CLUSTALW (*23*) and were inspected manually to eliminate ambiguously aligned positions. Single- and double-nucleotide substitutions were then determined by parsimony analysis using PAUP (*26*). The number of double-nucleotide substitutions expected to have occurred by coincidence of single-nucleotide substitutions at adjacent sites was determined by resolution of the following equations: ObsS = RealS + 2(RealD), ObsD = RealD + ExpD, and ExpD = (RealS/L)^{2}L, where L is the total number of aligned nucleotides (sample size), ObsS and ObsD are the numbers of changes observed as single- and double-nucleotide substitutions, respectively, RealS and RealD are the numbers of changes that have occurred as single- and simultaneous double–nucleotide substitutions, respectively, and ExpD is the number of doublet substitutions (observed at adjacent nucleotides) that have occurred by coincidence of two separate single-nucleotide changes. The difference of ObsD and ExpD was evaluated by a chi-squared test. In mammalian genomes, CG dinucleotides are methylated, which renders them susceptible to mutation (by deamination), yielding TG and CA dinucleotides at relatively high frequencies [A. P. Bird, *Nucleic Acids Res.* **8**, 1499 (1980)]. To prevent sequential CN→CG→TG and NG→CG→CA changes (or parallel CG→TG and CG→CA mutations in independent lineages) from being counted as doublet mutations, CN→TG and NG→CA substitutions were excluded from the analysis. We also examined the rates of substitution of each of the four nucleotides separately; the biases in these rates are not sufficient to affect our results.
19. The rate of double-nucleotide substitution was estimated as 2% of the overall rate of single-nucleotide substitution (RealD/ObsS in Table 2). Average rates of neutral nucleotide substitutions are around 5 per site per Gyr (*20*, *21*), yielding an estimated doublet rate of 0.1 per site per Gyr. The excess of double-nucleotide substitutions (the difference between ExpD and ObsD) is highly significant, with a chi-squared value of 14.2 (df = 1, $P \ll 0.005$).
20. H. Ochman and A. C. Wilson, *J. Mol. Evol.* **26**, 74 (1987).
21. W.-H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997); B. S. Gaut, *Evolutionary Biology*, M. K. Hecht *et al.*, Eds. (Plenum, New York, 1998), pp. 93–120; W. Makalowski and M. S. Boguski, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9407 (1998).
22. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
23. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).
24. G. J. Olsen, C. R. Woese, R. Overbeek, *J. Bacteriol.* **176**, 1 (1994); D. F. Feng, G. Cho, R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 13028 (1997); T. M. Embley, R. P. Hirt, *Curr. Opin. Genet. Dev.* **8**, 624 (1998); J. D. Palmer, R. K. Jansen, H. J. Michaels, M. W. Chase, J. R. Manhart, *Ann. Mo. Bot. Gard.* **75**, 1180 (1988); K. H. Wolfe, M. Gouy, Y. W. Yang, P. M. Sharp, W.-H. Li, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 6201 (1989); F. J. Ayala, A. Rzhetsky, F. J. Ayala, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 606 (1998); A. M. A. Aguinaldo *et al.*, *Nature* **387**, 489 (1997); S. Kumar and S. B. Hedges, *Nature* **392**, 917 (1998).
25. N. Saitou and M. Nei, *Mol. Biol. Evol.* **4**, 406 (1987).
26. D. Swofford, *PAUP (Phylogenetic Analysis Using Parsimony)*, version 3.0q (Illinois Natural History Survey, Champaign, IL, 1990).
27. This work was initiated in the Genetics Department of Trinity College, University of Dublin, and was continued with the support of C. Delidakis at the IMBB, Crete. It was supported in part by the EPET II programme of the General Secretariat for Research and Technology, Greece, and a UK Biotechnology and Biological Sciences Research Council grant G04905.

12 August 1999; accepted 23 December 1999

the total skew (the right-hand end of the plot) in that strain is further away from zero than in strain J99.

Moreover, GC skew across the contiguous block l–m–n in strain J99 is practically uniformly positive. Assuming that was the original order of these fragments in the common ancestor, the cumulative GC skew diagram of *H. pylori* appears to be very similar to diagrams of other bacteria. Despite the above rearrangements near the putative origin at the global minimum, the immediate surroundings of *oriC* (green tip, Fig. 1) are conserved between the two strains. The minimum of the green tip resides over the putative ORF JHP1416 in strain J99 (HP1527 in strain *26695*), next to the *dnaA* gene, which encodes the chromosome replication initiation protein. This is an additional indication that the minimum is close to *oriC* of *H. pylori*, because *dnaA* is found near the global minimum of cumulative GC skew (and the assigned origin) in many other bacterial genomes, including *B. burgdorferi*, *Bacillus subtilis*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *M. tuberculosis* and *Treponema pallidum*. Taken together, these facts point to a putative *oriC* location, which can be tested experimentally. The presence of *dnaA* in this region has also been noted by Salzberg *et al.*[11], although the midpoint of their predicted *oriC*-containing interval (with 18 oligomers) is some 20 kbp to the left of this gene.

Interestingly, the *cag* island and the plasticity zone are practically devoid of the ubiquitous and most skewed octamer RRTAGGGG (not shown). This implies that the plasticity zone is not a result of differential retention of ancestral DNA in these strains but is a product of horizontal transfer, in agreement with Alm *et al.*[2], who suggested this region might represent pathogenicity islands.

Restoring the order l–m–n increases the number of genes transcribed in the direction of replication (given the *oriC* location next to *dnaA*). Such 'co-directional' organization is seen in many bacterial genomes and has been interpreted in terms of minimizing 'head-on' collisions between DNA and RNA polymerases[17]. However, despite these differences in gene orientation, *26695* is a viable strain and it remains to be seen if the rearrangements so close to *oriC* have decreased its fitness.

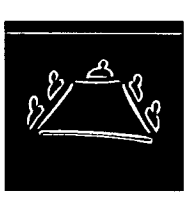
In conclusion, the described method of compact representation of large-scale genome differences can be applied for the graphical comparison of other genomes and strains to help in linking their organizational features (such as *oriC*), rearrangements and base composition. Perl scripts for calculating cumulative diagrams are available on request from the author.

Acknowledgements

I thank the anonymous referees, and also H. Loferer and I. Ivanov for their comments on earlier versions of the manuscript.

References

1 Tomb, J-F. *et al.* (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388, 1020–1022
2 Alm, R.A. *et al.* (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* 397, 176–180
3 Marczynski, G.T. and Shapiro, L. (1993) Bacterial chromosome origins of replication. *Curr. Opin. Genet. Dev.* 3, 775–782
4 Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* 13, 660–665
5 Freeman, J.M. *et al.* (1998) Patterns of genome organization in bacteria. *Science* 279, 1827a
6 Mrazek, J. and Karlin, S. (1998) Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci. U. S. A.* 95, 3720–3725
7 Grigoriev, A. (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.* 26, 2286–2290
8 McLean, M.J. *et al.* (1998) Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J. Mol. Evol.* 47, 691–696
9 Picardeau, M. *et al.* (1999) Physical mapping of an origin of bidirectional replication at the centre of the *Borrelia burgdorferi* linear chromosome. *Mol. Microbiol.* 32, 437–445
10 Qin, M.H. *et al.* (1999) Characterization of the functional replication origin of *Mycobacterium tuberculosis*. *Gene* 233, 121–130
11 Salzberg, S.L. *et al.* (1998) Skewed oligomers and origins of replication. *Gene* 217, 57–67
12 Lopez, P. *et al.* (1999) Identification of putative chromosomal origins of replication in Archaea. *Mol. Microbiol.* 32, 883–886
13 Edgell, D.R. and Doolittle, W.F. (1997) Archaea and the origin(s) of DNA replication proteins. *Cell* 89, 995–998
14 Myllykallio, H. *et al.* (2000) Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science* 288, 2212–2215
15 Censini, S. *et al.* (1996) *cag*, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proc. Natl. Acad. Sci. U. S. A.* 93, 14648–14653
16 Akopyants, N.S. *et al.* (1998) Analyses of the *cag* pathogenicity island of *Helicobacter pylori*. *Mol. Microbiol.* 28, 37–53
17 Brewer, B.J. (1990) Replication and the transcriptional organization of the *Escherichia coli* chromosome. In *The Bacterial Chromosome* (Drlica, K. and Riley, M., eds), pp. 61–83, ASM



Influential passengers come of age

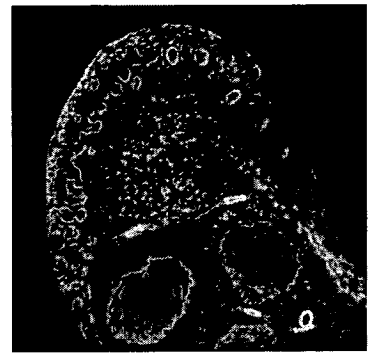
1st International *Wolbachia* Conference, Orthodox Academy, Kolymbari, Crete, Greece, 7–12 June 2000

To a good approximation, in 1990 nobody worked on *Wolbachia*. By contrast, a mere ten years later, 110 biologists from 21 countries assembled for the first meeting dedicated to these master manipulators of arthropod reproduction[1,2]. *Wolbachia*, which infect a quarter of all insects[1,2], are not-so-distant relatives of *Rickettsia*, and belong to the bacterial division from which mitochondria are descended. Perhaps the ability to 'cross-talk' with mitochondrial pathways underlies the success of these obligate intracellular parasites in altering host biology to enhance their own transmission[1,2] (Fig. 1).

Why do they do what they do?

Wolbachia can induce parthenogenesis (PI), cytoplasmic incompatibility (CI), male-killing (MK) or feminization (F) in their hosts; but why the variation? It is clear that host and parasite factors are both involved, and bacterial density was much discussed at this conference. For example, Greg Hurst (University College London, UK) found that the degree of MK in *Drosophila bifasciata* was correlated with *Wolbachia* density in eggs. *Wolbachia* can modify sperm, preventing normal decondensation of the paternal chromosomes. However, if control depends on bacterial density, low *Wolbachia* numbers may allow partial decondensation, resulting in some aneuploid offspring. Selection on hosts to depress bacterial density was a feature of several formal and informal coevolution models, and Fabrice Vavre (Université Claude Bernard, France) argued that it might underlie changes in the form of CI.

FIGURE 1. *Wolbachia* assure their germline transmission to the next generation



The bacteria (small red dots) can be seen around polyploid nurse cell nuclei (large central red bodies) in a developing *Drosophila* egg chamber and also 'swarming' in the area of the presumptive oocyte (right). The green features are ring canals, through which it is thought that *Wolbachia* move during development. Photograph kindly supplied by Stephen Hadfield (University of Oxford, UK).

In haplodiploids, fertilized eggs produce diploid females while unfertilized eggs produce haploid males. However, PI-*Wolbachia* cause gamete duplication, resulting in unfertilized diploid eggs, which become parthenogenetic females. Until recently, *Wolbachia* parthenogenesis was known only in various wasps (order Hymenoptera) but the haplodiploidy link is strengthened by discovery of PI in other haplodiploid taxa, namely certain mites (subclass Acari) and thrips (order Thysanoptera).

Wolbachia in humans

Perhaps the most exciting recent discovery is that all major human filarial nematodes harbour *Wolbachia*. Nematodes (unlike arthropods) seem to have a mutualistic relationship with *Wolbachia*, whose elimination leads to worm sterility, slow growth or death. Perhaps this symbiosis underlies the apparent congruence of nematode and *Wolbachia* phylogenies. *Wolbachia* lipopolysaccharide is the major inflammatory molecule in soluble extracts of filariae[3] and may be responsible for the inflammatory pathology seen in filariasis infections, including following chemotherapy, which often causes severe reactions in patients. The release of numerous *Wolbachia* into the blood from dead worms may have a direct involvement, since bacterial DNA is detected in the blood of individuals who suffered severe adverse reactions after chemotherapy (Helen Cross and Mark Taylor, Liverpool School of Tropical Medicine, UK).

Wolbachia on the move

Since *Wolbachia* are transmitted vertically in eggs (Fig. 1), their evolutionary trees should track their insect hosts, but they don't[1,2,4]. This suggests that *Wolbachia* jump between host species, but how? Recent screening of a community of *Drosophila* species and their parasitoid wasps[5] suggests that the much-touted host–parasitoid wasp interaction[1,2,4] is indeed a significant conduit. But how might transfer actually occur? Richard Stouthamer (Wageningen University, The Netherlands) described elegant experiments in which female wasps from infected and uninfected *Trichogramma kaykai* strains oviposited in the same host egg (quite common in nature). The *Wolbachia* somehow infected the previously uninfected larvae, which showed the trademark parthenogenesis phenotype when adult[6]. The physical mechanism is unknown but, in another *Trichogramma* species, competing larvae in the same egg puncture each other with their mandibles (David Merrett, University of Queensland, Australia), which has to help!

Two's company, three's a crowd

An emerging pattern is that many hosts harbour multiple *Wolbachia* strains. This is challenging since theoretical models find it far easier to eliminate competing strains than to allow their coexistence. However, the models described by Masakazu Shimada (University of Tokyo, Japan) suggest that triple infections are more likely in metapopulations, with frequent extinction and recolonization of patches. Mapping spatial and temporal patterns of infection, and probing the dynamics of competing strains within individuals and populations is clearly an important area, and Japanese researchers are establishing the bean weevil, *Callosobruchus chinensis*, with three *Wolbachia* strains, as a model.

The future

When our gracious organizer, Kostas Bourtzis (University of Ioannina and IMBB-FORTH, Greece), proposed the same venue for *Wolbachia* 2002, he received the loudest applause of the week! A highlight will surely be comparisons of five *Wolbachia* genomes differing in size, host and phenotype that are being sequenced by two initiatives. One is being headed by Scott O'Neill (Yale University, USA) and funded by the National Institutes of Health and New England Biolabs, and the other is a European (EU-funded) project coordinated by Kostas Bourtzis, involving seven institutes and a bioinformatics company. Recombination, common in other bacteria, but now raised in earnest for *Wolbachia* by Francis Jiggins (Cambridge University, UK), also seems likely to feature. Finally, a poster suggesting that more-sensitive PCR procedures may detect *Wolbachia* in 75% (not 25%) of insects, left us all with plenty to think about!

James M. Cook
j.cook@ic.au.uk

Antonis Rokas*
a.rokas@ed.ac.uk

Dept of Biology, Imperial College, Silwood Park, Ascot, Berks, UK.
*Institute of Cell, Animal and Population Biology, West Mains Rd, Edinburgh, UK

Further reading

1 O'Neill, S.L. *et al.*, eds (1997) *Influential Passengers, Inherited Micro-organisms and Arthropod Reproduction*, Oxford University Press
2 Werren, J.H. (1997) Biology of *Wolbachia*. *Annu. Rev. Entomol.* 42, 587–609
3 Taylor, M.J. *et al.* (2000) Inflammatory responses induced by the filarial nematode *Brugia malayi* are mediated by lipopolysaccharide-like activity from endosymbiotic *Wolbachia* bacteria. *J. Exp. Med.* 191, 1429–1435
4 Cook, J.M. and Butcher, R.D.J. (1999) The transmission and effects of *Wolbachia* bacteria in parasitoids. *Res. Popul. Ecol.* 41, 15–28
5 Vavre, F. *et al.* (1999) Phylogenetic evidence for horizontal transmission of *Wolbachia* in host–parasitoid associations. *Mol. Biol. Evol.* 16, 1711–1723
6 Huigens, M.E. *et al.* (2000) Infectious parthenogenesis. *Nature* 405, 178–179

A Bayesian guide to tree felling

To most evolutionary biologists, building phylogenetic trees is a bore. Comparative approaches to understanding patterns as diverse as the correlation of body size and brain size in primates, and the evolution of extended phenotypes in gallwasps, require a robust phylogeny on which to place changes in character state. But, for every question you are really interested in there are a thousand petty arguments concerning tree building. Which method should you use? Which parameter values should you plug in? How much can you trust the results? As Joe Felsenstein wrote, 'Comparative biologists may understandably feel frustrated upon being told that they need to know the phylogenies of their groups in great detail, when this is not something they had much interest in knowing.' However, he also goes on to say 'Nevertheless phylogenies are fundamental to comparative biology; there is no doing it without taking them into account.' Phylogenetic trees might be a nuisance, but until now you could not do without them. And, if your 'favourite' tree was wrong, so were your conclusions. However, a recent paper by Huelsenbeck *et al.*[1] has raised an exciting new possibility: evolution without trees.

The authors do not suggest that you don't do your phylogenetic tree-building homework, but that you can eliminate the uncertainty involved in tree building by considering all possible phylogenies and parameter values, weighting according to the probability that each is correct. The idea is just one more application of a field that is becoming increasingly common in evolutionary analysis: Bayesian statistics. In practice, this means that you replace parameters, such as tree topologies, biases in DNA mutation and rates of change of characters, by distributions representing your certainty, or otherwise, about them. These can be based on previous analyses, current data or even biological intuition. The probability of a given hypothesis being correct given the observed data is then summed over all possible realizations, weighting each by their relative probability. As Huelsenbeck *et al.*[1] demonstrate, this can lead to considerably different conclusions from what popular methods of tree reconstruction, such as parsimony, would suggest. More importantly, it gives you a good feel for how much faith to have in your conclusions.

Performing this task requires some serious computational effort, but there is new software appearing that can actually do the number crunching in reasonable time. Perhaps, more importantly, Bayesian analysis relies exclusively on the likelihood framework for tree building, and die-hard opponents of such model-based methods could argue that unless you are sure that the model of sequence evolution you are using is correct, the likelihood value is meaningless. Of course, you can always include more parameters, but this is a path that can spiral rapidly into absurdity. Nevertheless, tree felling the Bayesian way is going to make a large difference to everyone working with the comparative method because it eliminates the potential nightmare of using the wrong tree. As the lumberjacks say, if you can't see the wood for the trees, chop them down.

Antonis Rokas
a.rokas@ed.ac.uk

Gilean McVean
g.mcvean@ed.ac.uk

Reference

1 Huelsenbeck, J.P. *et al.* (2000) Accommodating phylogenetic uncertainty in evolutionary studies. *Science* 288, 2349–2350

Species loss after habitat fragmentation

Currently, we are losing large areas of natural and seminatural habitats. The big conservation question now is: how far does this habitat loss translate into species loss? Species–area relationships (SAR) describe the increase in the number of species S with increasing area A of a habitat. Generally, SAR obey a power law with an exponent <1; that is, the number of species increases faster with increasing area at smaller areas than at larger areas. If the SAR remains the same before and after the loss or fragmentation of habitats, the corresponding species loss is entirely determined by the amount of lost habitat, and can simply be looked up on the SAR. However, this only holds true if the geographical ranges of all species are uniform and if species extinction is determined by a certain abundance threshold. Although these assumptions are realistic for some systems, they need not be for others. This is from where Muriel Ney-Nifle and Marc Mangel[1] start out. In their model, they consider a distribution of geographical species ranges characterized by few species with large ranges and many species with small ranges. Moreover, they consider extinction to depend on the realized range of a species, not its abundance.

Ney-Nifle and Mangel find that habitat loss shifts the SAR, and that this change depends on the shape of the lost area. This implies that predictions of species loss from SAR have to be considered with caution, if they are based on the SAR that was valid before the habitat loss. This is illustrated by their modeling example of habitat fragmentation, for which Ney-Nifle and Mangel find a much higher species loss (30% of species lost after a 50% reduction of total area) than the one predicted from prefragmentation SAR (10% of species lost).

The work by Ney-Nifle and Mangel suggests that, in many cases, consequences of habitat fragmentation are more severe than previously thought. This is especially relevant for communities with a large fraction of species with aggregated distributions, such as tropical forests. Now, what are the goals for the future? Of course, in the real world, both abundance and realized range might determine extinction, and both might be confounded. Therefore, the goal for the near future is to integrate both abundance requirements and range requirements into predictions of species loss. At the same time, this might constitute a large step towards the ultimate goal: the integration of the biology of species into SAR and predictions of species loss.

Markus Fischer
fischerm@uwinst.unizh.ch

Reference

1 Ney-Nifle, M. and Mangel, M. (2000) Habitat loss and changes in the species–area relationship. *Conserv. Biol.* 14, 893–898

Journal Club – a new section for *TREE*

Each month a panel of contributors will provide timely commentaries of significant new papers from ecology and evolution journals for our new **Journal Club**. The invited panel consists of established researchers from a broad range of backgrounds, reflecting the diversity of subjects covered in *TREE*.

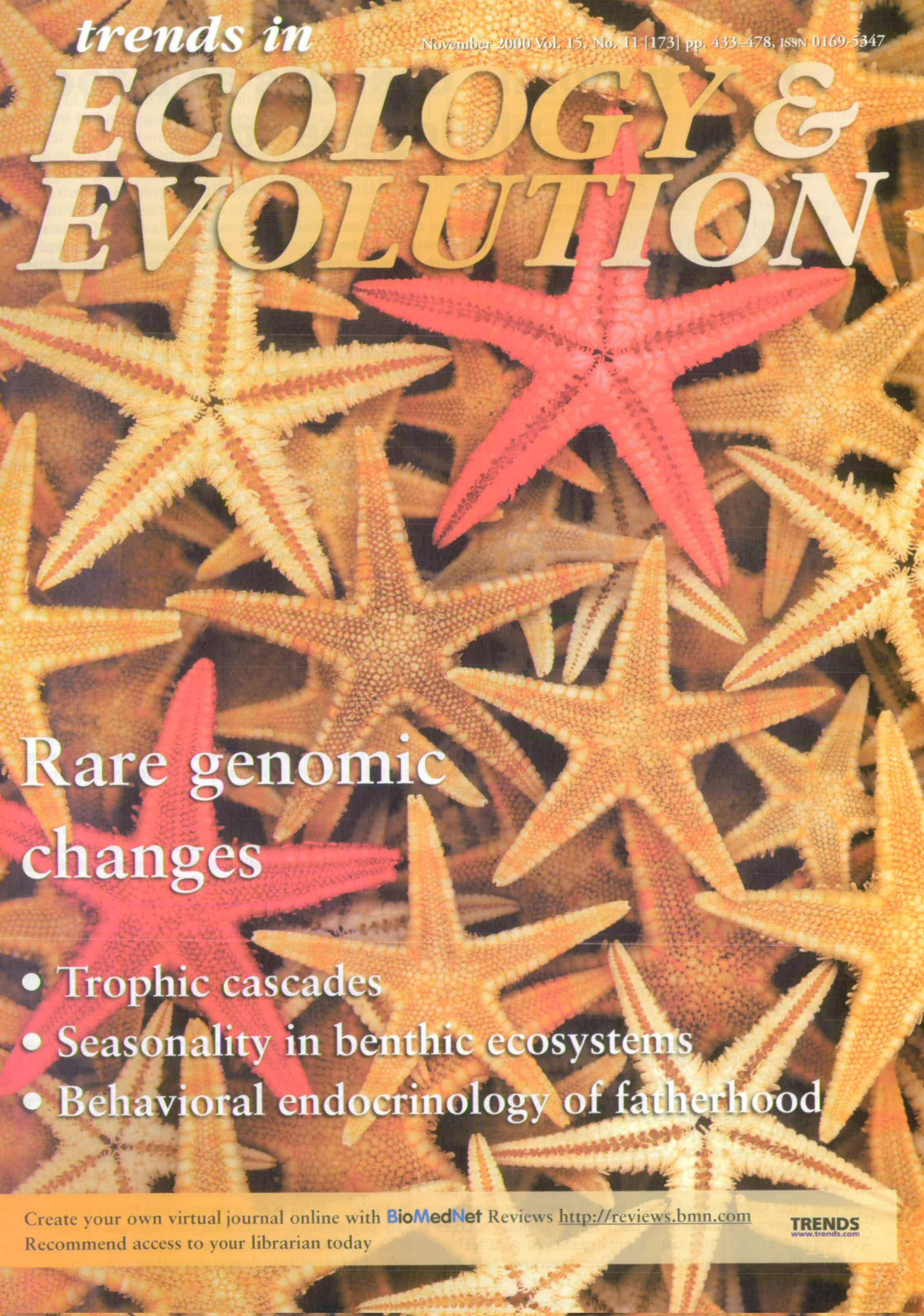
Many of the commentaries are also published in advance through the Commentaries section of BioMedNet. Access to Commentaries is FREE. All you need do to find the Commentaries section of BioMedNet is go to http://www.bmn.com, click on the 'News and Comment' link and then press the Commentaries button (you might need to register as a BioMedNet member if you haven't already).

If you have any comments regarding this new section please contact the Editor at tree@current-trends.com

 0169-5347/00/$ – see front matter

trends in
November 2000 Vol. 15, No. 11 [173] pp. 433–478, ISSN 0169-5347

ECOLOGY & EVOLUTION



Rare genomic changes

- Trophic cascades
- Seasonality in benthic ecosystems
- Behavioral endocrinology of fatherhood

Rare genomic changes as a ool for phylogenetics

Antonis Rokas and Peter W.H. Holland

DNA sequence data have offered valuable insights into the relationships between living organisms. However, most phylogenetic analyses of DNA sequences rely primarily on single nucleotide substitutions, which might not be perfect phylogenetic markers. Rare genomic changes (RGCs), such as intron indels, retroposon integrations, signature sequences, mitochondrial and chloroplast gene order changes, gene duplications and genetic code changes, provide a suite of complementary markers with enormous potential for molecular systematics. Recent exploitation of RGCs has already started to yield exciting phylogenetic information.

Antonis Rokas is at the Institute of Cell, Animal and Population Biology, Ashworth Laboratories, King's Buildings, West Mains Road, University of Edinburgh, Edinburgh, UK EH9 3JT (a.rokas@ed.ac.uk); Peter Holland is at the School of Animal and Microbial Sciences, University of Reading, Whiteknights, PO Box 228, Reading, UK RG6 6AJ (p.w.h.holland@reading.ac.uk).

In recent years, considerable progress has been made in the field of molecular phylogenetics. A significant driving force has been the increasing technical ease of DNA sequencing, which has led to the dominance of primary sequence data as indicators of the historical relationships between taxa. Important advances have also occurred in the computational analysis of DNA sequence data[1], such as improved methods for modelling patterns of nucleotide substitutions[2]. However, the task of phylogenetic reconstruction using molecular sequences is not without problems. To a large extent, these stem from the fact that the dominant methods for molecular phylogeny reconstruction exploit nucleotide substitutions (plus, in some cases, single-site insertions or deletions) as indicators of divergence or common descent. Convergent evolution of nucleotide bases, differing substitution rates among sites and lineages[3], saturation of mutations at variable sites[4], nonindependent substitutions among sites[5] and functional constraints at the molecular level[6] are just a small sample of the potential caveats that apply when using these types of data. As a result, phylogenetic hypotheses based on primary sequence data can sometimes be equivocal[7,8], whereas others can simply be incorrect[9,10]. The advent of the genomic era has brought the opportunity to consider other types of information embedded in DNA sequences. Here, we consider the phylogenetic use of large-scale mutations – rare genomic changes (RGCs; Box 1), which occur relatively infrequently. Researchers have already started using RGCs for inferring relationships between living organisms.

Rare genomic changes

We define RGCs as large-scale mutational changes that have occurred in the genomes of particular clades. Examples of RGCs (Table 1) include intron indels, retroposon integrations, signature sequences, changes in organelle gene order, gene duplications and genetic code variants. Most RGCs represent changes caused by single (or a few) mutational events; in our discussion of RGCs we do not include genomic characteristics that are, most probably, the end result of multiple processes (e.g. genomic compositional contrasts[11]). Until recently, many studies mapped RGCs onto existing phylogenies to gain insight into their mode of evolution. The consensus that has emerged is that RGCs are often evolutionarily conserved and phylogenetically informative. We believe the time has come to turn the question around: what can RGCs tell us about phylogenies themselves?

RGCs provide an independent source of phylogenetic information, largely immune from some of the problems that affect primary sequence data. A major difficulty with this approach is the identification of these rare mutations in the clades of interest. However, the increasing automation of molecular techniques has brought us to the dawn of the genomic era where tremendous amounts of information, freely available in the primary literature and public databases, are generated. Additionally, protocols have been developed for the targeted identification of many, if not all, RGCs. Here, we argue that the application of RGCs to phylogenetics can offer new insights into evolutionary history. Furthermore, in cases where primary sequence data generate conflicting or equivocal results, RGCs offer an independent way of evaluating alternative phylogenetic scenarios.

RGCs as 'Hennigian' markers

The field of phylogenetics has been strongly influenced by the founder of the cladistic methodology, the German entomologist Willi Hennig. Hennig argued that only shared derived characters (synapomorphies; Box 1) should be used as indicators of common descent. Plotting the distribution of synapomorphies is the essence of cladistic reconstruction. The principal hindrance to this task is homoplasy (see Box 1 for definition). In general, character states that arise rarely will not be prone to extensive convergent or parallel evolution, which should contribute to a low level of homoplasy. Although the precise frequency of occurrence of most RGCs has not been robustly estimated, large-scale mutations are generally rare. Additionally, precise secondary loss of the character (homoplasy because of reversion) is likely to be extremely rare for most large-scale mutations and has been demonstrated to be so in some cases [e.g. short interspersed element insertions (SINEs)[12]; Box 1]. Therefore, with respect to homoplasy, RGCs might constitute good markers of common descent. In Table 1, we provide a summary of the characteristics and phylogenetic applicability of various categories of RGC, and in the next section we expand on a few examples published recently to demonstrate their potential use.

 0169-5347/00/$ – see front matter PII: S0169-5347(00)01967-4

Of fish and flies: intron indels as clade markers

The power and robustness of RGCs is well demonstrated by the study of Venkatesh *et al.*[13], in which intron indels (Box 1) were used to investigate fish phylogeny. Venkatesh *et al.* identified seven intron positions (in five genes) that are present in the pufferfish *Takifugu rubripes* but not in the homologous genes of mammals. Four introns were also found in the rhodopsin gene that were present in the ancestral chordate rhodopsin gene (as inferred by their presence in basal chordates, such as lampreys and skates, and in the more apical lineage of mammals) but were absent in the pufferfish. Several ray-finned fish species (class Actinopterygii) were screened for the presence or absence of these eleven intron indels, and these data were used to reconstruct evolutionary relationships[13]. Only one indel showed considerable homoplasy and an unclear phylogenetic signal; all the others were unique synapomorphies able to resolve phylogenetic relationships. It is noteworthy that some of the relationships resolved, such as the basal position of bichirs (*Polypterus* spp.) within the Actinopterygii, have proved contentious using primary sequence comparisons.

Another recent use of an intron indel as a phylogenetic character deals with the placement of the insect order Strepsiptera within holometabolous insects[14]. Strepsipteran forewings resemble the hindwing balancing organs of flies (order Diptera), which are known as halteres. Among other phylogenetic scenarios, an affinity of Strepsiptera to Coleoptera has been widely discussed, based primarily on the use of hindwings for flight in both orders. An alternative proposal is a sister group relationship with Diptera[15,16]. In this case, halteres could be homologous, but a radical homeotic mutation might have reversed their position in Strepsiptera[15]. Evidence from morphology is equivocal[17] and 18S rDNA sequence data have generated a lively debate between researchers favouring different phylogenetic reconstruction methods[8,16]. Rokas *et al.* noted a unique intron insertion in the homeobox of the *engrailed* gene of Diptera and Lepidoptera, which is absent from other insects and all outgroups[14]. Possession of the intron in Strepsiptera would support a sister group relationship with Diptera, whereas its absence would argue against this affinity. Cloning of the Strepsipteran homologue of *engrailed* showed that the intron is absent in Strepsiptera, thus suggesting that the halteres of Strepsiptera and Diptera might not represent a rare case of natural homeotic transformation but a remarkable case of convergent evolution[14] (Fig. 1).

Box 1. Glossary

Bilateria: the bilaterally symmetrical animals.
Homoplasy: a general term denoting that the acquisition of the same character state in two taxa is not because of common descent. This can arise by parallel evolution (independent acquisition from the same ancestral state), convergent evolution (independent acquisition from different ancestral conditions) or secondary loss (reversion from the derived to the ancestral condition).
Indel: an insertion or deletion event.
LINEs (long interspersed elements): a class of retroposons that are capable of self-transposition.
Orthology: the relationship between two homologous loci derived from a speciation event.
Paralogy: the relationship between two homologous loci derived from a duplication event.
Polyphyly: when a group does not include the most recent common ancestor of all its members.
Protein domain: a well defined region within the protein. It can be distinguished on the basis of function or structure. For example, the homeodomain is a 60-amino acid domain shared by proteins encoded by homeobox genes.
Protein motif: any stretch of contiguous sequence within a protein that has been evolutionarily conserved.
Rare genomic change (RGC): a large-scale mutational change that has occurred in the genome of a particular clade.
Retroposons: the class of transposable elements that relocate in the genome via an RNA intermediate using the enzyme reverse transcriptase.
Signature sequences: shared conserved insertions or deletions in proteins or RNAs.
SINEs (short interspersed elements): a particular class of retroposons that have lost the ability to transpose themselves (to transpose they use another class of mobile elements, LINEs).
Synapomorphy: a shared derived character state that suggests a monophyletic grouping.

Of SINEs and LINEs

Retroposons (Box 1) belong to the group of transposable elements that use an RNA-mediated mode of transposition[12]. Retroposon integrations, especially from the class of SINEs (retroposons that lack the ability for self-amplification), have been used successfully as phylogenetic markers; an application pioneered by Okada and colleagues in the 1990s (Refs 18,19). It has been argued that SINE integrations come close to being 'perfect' markers of common descent because integration is apparently random and irreversible, and because most eukaryotic genomes have an abundance of SINE elements[12]. Their presence or primitive absence can also be readily detected by PCR amplification across integration sites. Successful applications of SINEs include the generation of convincing

Table 1. Summary of useful characteristics of rare genomic change (RGC) markers for phylogenetic purposes[a]

Marker	Taxonomic resolution	Homoplasy	Taxa in which RGCs are applicable
Intron indels	Wide ranging	Low	Eukaryotes
Retroposons (SINEs and LINEs)[b]	Within orders	Zero to very low	Animals
Signature sequences	Wide ranging	Unknown but recognizable	All life
mtDNA genetic code variants	Phyla to classes	Low to moderate	Eukaryotes
Nuclear DNA genetic code variants	Phyla	Low to moderate	All life
mtDNA gene order	Wide ranging (phyla to families)	Low to moderate in animals. High in plants, fungi and protists	Eukaryotes
cpDNA gene order	Families	Low	Plants
Gene duplications	Wide ranging	Unknown	All life
Comparative cytogenetics	Within phyla	Unknown	All life (lateral gene transfer is prevalent in prokaryotes)

[a]For more detailed information, see references cited in the text.
[b]Abbreviations: SINEs, short interspersed elements; LINEs, long interspersed elements.

encoded by many Hox genes possess specific sequence motifs near the homeodomain, which is a domain common to all Hox genes. These sequence motifs have helped distinguish orthologous and paralogous (Box 1) Hox genes. Each of the three major clades has its own unique Hox genes that do not have identifiable orthologues in the others. In other words, gene duplications have yielded distinct genes in each lineage and these have acquired unique signature protein motifs (Box 1). For example, the lophotrochozoans share *Lox2*, *Lox4*, *Lox5*, *Post-1* and *Post-2*, whereas the ecdysozoans share *Ubx* and *Abd-B* (Ref. 25).

As well as providing independent support for the controversial Lophotrochozoa and Ecdysozoa clades, this approach has been used to investigate the affinities of a particularly enigmatic animal phylum: the dicyemid mesozoa[26]. These are microscopic parasites of squid and octopus, with an amazingly simple body plan consisting of a solitary axial cell surrounded by a single layer of 10–40 ciliated outer cells. Morphology and 18S rDNA sequence data have previously failed to adequately resolve their phylogenetic position. Recently, Kobayashi *et al.*[26] cloned the *Lox5* gene from a dicyemid, including the diagnostic *Lox5* peptide, thus demonstrating that these animals are almost certainly highly degenerate members of the Lophotrochozoa clade (Fig. 1). Indeed, dicyemids represent one of the most extreme cases of secondary simplification of morphology known in the animal kingdom.

The Hox gene data represent a special case of 'signature sequences' (Box 1). In Hox genes, the existence of distinct signature sequences in different genes, and in different clades, suggests that these motifs (such as the *Lox5* peptide) have biochemical functions. In other examples, insertions might have little functional significance; nonetheless, they can be used as RGCs for phylogenetic reconstruction; for example, there is an ongoing debate in prokaryote phylogenetics about whether archaebacteria are monophyletic. Recently, the paraphyly hypothesis of archaebacteria has been supported by several signature sequences[27,28]. However, it should be noted that historical associations within prokaryotes are still incompletely resolved owing to extensive subsequent gene transfer. No molecular marker is immune from this all-pervading complication[29]; for example, a signature sequence in the gene *hsp70* used to support the paraphyly of archaebacteria[28] is also present in one of the three copies of the gene in *Escherichia coli*, suggesting a possible recent transfer[4]. Other important studies using the signature-sequence approach include a confirmation that the archezoa are true eukaryotes that have lost mitochondria[30], and an investigation of branchiopod crustacean phylogeny[31]. In branchiopod crustaceans, three unique helices in 18S rDNA were used to distinguish cladocerans from other branchiopods, demonstrating that useful sequences can be found in RNA, as well as in protein sequences.

Deviant codes and shuffled genes

Several organisms use genetic codes that deviate from the standard 'universal' code. These 'deviant' codes can be useful markers for higher level phylogenetics. Keeling and Doolittle[32] showed that a genetic code in which TAA and TAG codons encode glutamine, rather than termination, is used by almost all diplomonads, with the exception of the genus *Giardia*, which employs the standard genetic code. This argues for an early divergence of *Giardia* in the evolution of diplomonads and is in agreement with phylogenies from primary sequence data[32]. The diplomonad deviant code has also been found in certain green algae and in ciliates, showing that homoplasious changes can occur. Mitochondrial genomes have the widest variety of deviant codes, whereas plastids show no deviation from the universal (so far)[32,33]. Variant mtDNA codes in animals have been studied in some detail, aided by complete sequences of mtDNA from a wide range of animals. For example, a sister group relationship between echinoderms and hemichordates is supported by the assignment of the ATA codon to the amino acid isoleucine[34], as well as by sequence analyses[35] (Fig. 1), although the same reassignment has occurred independently in Cnidaria.

Gene order changes, particularly in circular genomes such as mitochondria and chloroplasts, comprise another type of RGC that has already proved useful in phylogenetics[36]. These arrangements, effected by inversions, translocations and duplications, generally affect several adjacent genes. They are unlikely to be reversed precisely because of their complexity; therefore, they satisfy one of the principal criteria demanded of the perfect phylogenetic marker. The second criterion, low levels of homoplasy, is also predicted to be true because convergence or parallelism would imply bias towards particular gene rearrangements or gene orders. Isolated cases of convergence have been detected[37], suggesting bias in some taxa; however, this does not seem to be a widespread problem. Some key phylogenetic problems have been tackled using mtDNA order as a marker, with definitive results. For example, until recently it was widely accepted that insects and myriapods were close relatives within the arthropods; indeed, these two primarily terrestrial taxa share many derived morphological characters. Several lines of evidence, including developmental data and primary sequence comparisons, have challenged this relationship, raising the alternative possibility of a crustacean–insect clade[38]. This suggestion is effectively confirmed by the shared presence of a rare tRNA translocation within insects and crustacean mtDNA, which is not seen in myriapods, chelicerates, tardigrades, onychophorans or outgroups[39] (Fig. 1). In most animal taxa, changes to mtDNA gene order are rare, making these markers useful for higher level phylogenetics[39,40]; although one exception might be the gastropod molluscs, where mtDNA gene order is extremely variable[41]. Similarly, plant, fungi and protist mtDNAs display rapid genome reshaping, making gene order a more appropriate marker for lower-level phylogenetics[42].

Chloroplast DNA (cpDNA) gene order has been exploited in a similar way to mtDNA gene order. For example, in 1987 Jansen and Palmer used a cpDNA inversion within the sunflower family to propose the basal position of the Barnadesiinae, with implications for biogeography and morphological evolution in this group[43]. More recently, Doyle *et al.*[44] surveyed 132 legume genera for the occurrence of a 50 kb inversion, finding evidence that at least two tribes within the legumes were polyphyletic. A qualitatively different sort of rearrangement from those discussed above is deletion. For example, monophyly of the conifers is supported by loss of one copy of an inverted repeat found in cpDNA (Ref. 45).

Other potential RGCs

The list of RGCs we have described so far is not exhaustive; several other categories of large-scale mutation exist, some of which have potential for phylogenetics. For example, gene duplications have not yet been widely exploited. One difficulty is technical: unless a family of genes is arranged in a tandem array, discerning whether a duplicated copy of a gene exists is difficult because absence of

14 Rokas, A. *et al.* (1999) Intron insertion as a phylogenetic character: the *engrailed* homeobox of Strepsiptera does not indicate affinity with Diptera. *Insect Mol. Biol.* 8, 527–530

15 Whiting, M.F. and Wheeler, W.C. (1994) Insect homeotic transformation. *Nature* 368, 696

16 Whiting, M.F. *et al.* (1997) The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from the 18S and 28S ribosomal sequences and morphology. *Syst. Biol.* 46, 1–68

17 Kristensen, N.P. (1999) Phylogeny of endopterygote insects, the most successful lineage of living organisms. *Eur. J. Entomol.* 96, 237–253

18 Kido, Y. *et al.* (1991) Shaping and reshaping of salmonid genomes by amplification of transfer RNA-derived retroposons during evolution. *Proc. Natl. Acad. Sci. U. S. A.* 88, 2326–2330

19 Murata, S. *et al.* (1993) Determination of the phylogenetic relationships among Pacific salmonids by using short interspersed elements (SINEs) as temporal landmarks of evolution. *Proc. Natl. Acad. Sci. U. S. A.* 90, 6995–6999

20 Nikaido, M. *et al.* (1999) Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: hippopotamuses are the closest extant relatives of whales. *Proc. Natl. Acad. Sci. U. S. A.* 96, 10261–10266

21 Miyamoto, M.M. (1999) Perfect SINEs of evolutionary history? *Curr. Biol.* 9, R816–R819

22 Hillis, D.M. (1999) SINEs of the perfect character. *Proc. Natl. Acad. Sci. U. S. A.* 96, 9979–9981

23 Verneau, O. *et al.* (1998) Determining and dating recent rodent speciation events by using L1 (LINE-1) retrotransposons. *Proc. Natl. Acad. Sci. U. S. A.* 95, 11284–11289

24 Balavoine, G. (1998) Are platyhelminthes coelomates without a coelom? An argument based on the evolution of Hox genes. *Am. Zool.* 38, 843–858

25 de Rosa, R. *et al.* (1999) Hox genes in brachiopods and priapulids and protostome evolution. *Nature* 399, 772–776

26 Kobayashi, M. *et al.* (1999) Dicyemids are higher animals. *Nature* 401, 762

27 Rivera, M.C. and Lake, J.A. (1992) Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* 257, 74–76

28 Gupta, R.S. (1998) What are archaebacteria: life's third domain or monoderm prokaryotes related to gram-positive bacteria? A new proposal for the classification of prokaryotic organisms. *Mol. Microbiol.* 29, 695–707

29 Doolittle, W.F. (1999) Lateral genomics. *Trends Biochem. Sci.* 24, M5–M8

30 Hashimoto, T. *et al.* (1998) Secondary absence of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny. *Proc. Natl. Acad. Sci. U. S. A.* 95, 6860–6865

31 Crease, T.J. and Taylor, D.J. (1998) The origin and evolution of variable-region helices in V4 and V7 of the small-subunit ribosomal RNA of branchiopod crustaceans. *Mol. Biol. Evol.* 15, 1430–1446

32 Keeling, P.J. and Doolittle, W.M. (1997) Widespread and ancient distribution of a noncanonical genetic code in diplomonads. *Mol. Biol. Evol.* 14, 895–901

33 Palmer, J.D. and Delwiche, C.F. (1998) The origin and evolution of plastids and their genomes. In *Molecular Systematics of Plants II: DNA Sequencing* (Soltis, D.E. *et al.*, eds), pp. 375–409, Kluwer

34 Castresana, J. *et al.* (1998) The mitochondrial genome of the hemichordate *Balanoglossus carnosus* and the evolution of deuterostome mitochondria. *Genetics* 150, 1115–1123

35 Bromham, L.D. and Degnan, B.M. (1999) Hemichordates and deuterostome evolution: robust molecular phylogenetic support for a hemichordate + echinoderm clade. *Evol. Dev.* 1, 166–171

36 Boore, J.L. (1999) Animal mitochondrial genomes. *Nucleic Acids Res.* 27, 1767–1780

37 Mindell, D.P. *et al.* (1998) Multiple independent origins of mitochondrial gene order in birds. *Proc. Natl. Acad. Sci. U. S. A.* 95, 10693–10697

38 Averof, M. and Akam, M. (1995) Insect–crustacean relationships: insights from comparative developmental and molecular studies. *Philos. Trans. R. Soc. London Ser. B* 347, 293–303

39 Boore, J.L. *et al.* (1998) Gene translocation links insects and crustaceans. *Nature* 393, 667–668

40 Boore, J.L. *et al.* (1999) Complete sequence, gene arrangement, and genetic code of mitochondrial DNA of the cephalochordate *Branchiostoma floridae* (amphioxus). *Mol. Biol. Evol.* 16, 410–418

41 Kurabayashi, A. and Ueshima, R. (2000) Complete sequence of the mitochondrial DNA of the primitive opisthobranch gastropod *Pupa strigosa*: systematic implication of the genome organization. *Mol. Biol. Evol.* 17, 266–277

42 Lang, B.F. *et al.* (1999) Mitochondrial genome evolution and the origin of eukaryotes. *Annu. Rev. Genet.* 33, 351–397

43 Jansen, R.K. and Palmer, J.D. (1987) A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). *Proc. Natl. Acad. Sci. U. S. A.* 84, 5818–5822

44 Doyle, J.J. *et al.* (1996) The distribution and phylogenetic significance of a 50-kb chloroplast DNA inversion in the flowering plant family Leguminosae. *Mol. Phylog. Evol.* 5, 429–438

45 Raubeson, L.A. and Jansen, J.K. (1992) A rare chloroplast-DNA structural mutation is shared by all conifers. *Biochem. Syst. Ecol.* 20, 17–24

46 Garcia-Fernàndez, J. and Holland, P.W.H. (1994) Archetypal organization of the amphioxus Hox gene cluster. *Nature* 370, 563–566

47 Stellwag, E.J. (1999) Hox gene duplication in fish. *Semin. Cell Dev. Biol.* 10, 531–540

48 O'Brien, S.J. and Stanyon, R. (1999) Ancestral primate viewed. *Nature* 402, 365–366

49 O'Brien, S.J. *et al.* (1999) The promise of comparative genomics in mammals. *Science* 286, 458–481

50 Darwin, C. (1859) *On the Origin of Species* (2nd edn), John Murray

Journal Club

Treat 'em mean, keep 'em keen: are females drawn to aggressive males?

The evolution of extreme male courtship displays poses a challenge to evolutionary biologists, who have responded with the development and testing of numerous theories. Most of these theories assume that female preference and male trait co-evolved, or that male traits evolved in response to already evolved female preferences. Less attention has been paid to the pre-existing trait hypothesis, which suggests that females evolved preferences for existing male display traits that are subsequently co-opted for use in courtship. Under this hypothesis, the display traits are considered to have been co-opted from displays of male–male aggression. Such signals are suggested to be useful to females in allowing them to reliably evaluate a potential mate's ability to provide genetic and/or proximate benefits. Lack of supporting evidence is the probable reason why this hypothesis has received limited attention. Although similarity of courtship and aggressive calls have been noted in a broad range of species, this provides only weak support, because it does not rule out other scenarios, such as courtship calls being co-opted for aggression. Now, much stronger support is presented by the work of Gerald Borgia and Seth William Coleman on the 'skraa calls' of bowerbirds[1].

By analysing sonograms, these authors first demonstrate that skraa calls used in courtship are homologous with those used in aggression. Then, using phylogenetic reconstruction, they provide strong evidence that skraa calls were used in aggression before their use in courtship. This analysis also suggests that aggressive-like courtship calls were present near the predicted time of co-option. The authors are also able to demonstrate a female preference for very intense, aggressive-like courtship displays. They suggest that bower designs predicted to be in use at the time of co-option feature aspects that provide protection for the female from any adverse effects of such aggressive-like courtship. This argument is supported by experimentation. Male spotted bowerbirds court females through a see-through wall that provides a barrier between the two. Females prefer males with more intense displays. However, if this wall is experimentally removed, then courtship still occurs but males reduce their display intensity, suggesting that an intense display is effective only when females are offered protection by the bower wall. More ingeniously, new biogeographical comparisons indicate that these ideas can explain much of the diversity that we see in bower design between species. In species where terrain limits the ability of the birds to build bowers that modulate negative aspects of display intensity, male courtship calls themselves are reduced in intensity.

There are two key messages from this work. First, the pre-existing trait hypothesis for the evolution of male display signals has been neglected for too long. The empirical support given by this study should prompt empiricists and theoreticians into giving it further study. Second, we feel that bowerbirds provide an excellent model system for testing hypotheses about sexual selection. We already have a large body of knowledge on the bower characteristics that females find attractive, and experimental manipulation of these nonmorphological traits can be a great deal easier than for signals such as plumage coloration.

1 Borgia, G. and William Coleman, S. (2000) Co-option of male signals from aggressive display in bowerbirds. *Proc. R. Soc. London B Biol. Sci.* 267, 1735–1740

Stuart Humphries
s.humphries@bio.gla.ac.uk

Graeme D. Ruxton
g.ruxton@bio.gla.ac.uk

PII: S0169-5347(00)02028-0

Evolutionary biology meets genomics

Four years ago, a *TREE* article warned readers to prepare for the flood of data that was starting to accumulate from advances in genomics, hinting at the potential benefits for evolutionary research. Four years later, genome projects on living creatures as diverse as bacteria, flies, plants and humans are at or near completion. Furthermore, databases are ever increasing in size and a rising number of projects are dependent on them. So far so good, some might say. But, until now, evolutionary biologists have mostly benefited from genomics indirectly, by using data originally generated for other purposes. What about genomics *and* evolutionary biology?

Now, David Pollock *et al.*[1] are addressing exactly this issue, namely the application of technological advances in genomics for molecular evolutionary studies. The classic

approach to comparative studies has been the separate isolation of markers for each species, which can be very time-consuming. Pollock *et al.* argue that the time and cost of these projects could be reduced by combining the DNA samples from the different taxa and shotgun sequencing of the mixture. Shotgun sequencing simply involves the random breaking of DNA into pieces, with subsequent cloning of these fragments into vectors. The final step is the random sequencing of the fragments. Given sufficient sequence coverage, the assembly of the individual genomes, from the mix of sequences, will be easy because overlap between the sequence fragments will allow their identification. To illustrate the applicability of this technique, Pollock *et al.* used published mitochondrial genome sequences from ten species and simulated the 'shotgunning' and assembly of the fragments under assumptions expected to occur in a such a project. The results are promising, with six out of the ten genomes completely and correctly reconstructed, and the other four correctly reconstructed but with short gaps.

So what? Should we start shotgun sequencing mixtures of living creatures? The authors have already started doing this in a pilot project, and propose rapid expansion to sequence 2000 mitochondrial genomes from vertebrates! Given the numbers of described vertebrate species (4000 mammals, 7000 birds, 20 000 fish and thousands of reptiles and amphibians), it is the first time that such extensive data will be accumulated for comparative analysis, with tremendous benefits for molecular evolutionary studies. The authors believe that the same principle will eventually be applied to whole nuclear genomes and for a range of taxonomic ranks (including intraspecific questions), but nothing beats good ideas like data. Only they will tell if there is grandeur in this view of life...

1 Pollock, D.D. *et al.* (2000) A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. *Mol. Biol. Evol.* 17, 1776–1788

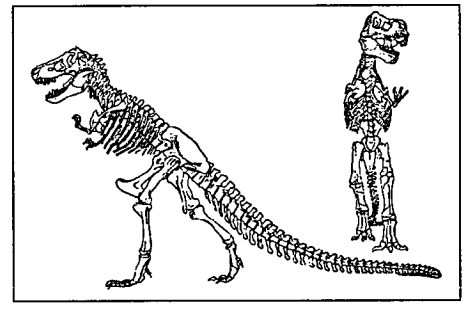
Antonis Rokas
a.rokas@ed.ac.uk

PII: S0169-5347(00)02051-6

Digit identity and digit number: indirect support for the descent of birds from theropod dinosaurs

The evolutionary origin of birds is the subject of heated debate. As more data are collected, a consensus is emerging in favour of a descent from theropod dinosaurs. However, in 1997, important evidence refuting this hypothesis was put forward by Burke and Feduccia. They provided new data that confirmed old claims that the digit identity of the three fingers of the hand in theropods differs from that of birds (digits I, II and III versus II, III and IV), which implies that there is no homology between the first, second and third digits of the hand of these groups.

In 1999, Wagner and Gauthier subsequently proposed a hypothesis that reconciled the differences in digit homology with an avian descent from theropods. They suggested that changes in digit identity are the cause of homeotic changes (i.e. changes of the identity of one structure into the identity of another structure). Homeotic changes of identity in serial structures are a common occurrence in the vertebral column (e.g. 5% of humans have a thoracic vertebra with a rib instead of the first ribless lumbar one). Also, shifts in identity for an entire group of serial structures, such as that proposed by Wagner and Gauthier, have been amply documented for vertebrae in transgenic mice. There are several different mutations that shift the identity of one vertebra along the entire vertebral column in either the anterior or the posterior direction (e.g. mutations of the polycomb-group and trithorax-group genes). However, no firm evidence for homeotic changes in digits had been found, although it is plausible that the presence of two opposable digits instead of one in koalas and many tree-dwelling birds represents such a homeotic change (here, a change is the identity of the second digit into that of the first).



In a new paper, Drossopoulou *et al.*[1] now present evidence for the developmental independence of the determination of digit number and digit identity. In an elegant experiment, they demonstrate that the important gene Sonic hedgehog (Shh) is initially involved in the determination of the number of digits and later on in the specification of digit identity via the induction of Bmp genes. This complies with existing knowledge that ectopic expression of Shh induces supernumerary digits (polydactyly). By carefully manipulating the gene activity of Shh and Bmp2, they produced phenotypes with the same number of (supernumerary) digits but with differences in digit identity (digit 2, 3 or 4), thus showing that the identity of digits can be changed without a change in the number of digits. Evolutionarily, it is important to realize that Shh is not unique in this respect. At present, at least one other gene is known that has an effect comparable to Shh on the development of supernumerary digits – the related Indian hedgehog gene.

Although the evidence is far from complete, these results indirectly support the hypothesis of Wagner and Gauthier. They prove the underlying assumption that homeotic changes in the identity of digits are possible and can occur without a change in digit number. Apart from the relevance to the interpretation of the evolution of birds, Drossopoulou *et al.* have also provided an important piece in the intriguing puzzle of limb and digit development.

1 Drossopoulou, G. *et al.* (2000) A model for anteroposterior patterning of the vertebrate limb based on sequential long- and short-range Shh signalling and Bmp signalling. *Development* 127, 1337–1348

Frietson Galis
galis@rulsfb.leidenuniv.nl

PII: S0169-5347(00)02048-6

Molecular Ecology (2001) 10, 761–778

Differential success in northwards range expansion between ecotypes of the marble gallwasp *Andricus kollari*: a tale of two lifecycles

GRAHAM STONE,* RACHEL ATKINSON,* ANTONIS ROKAS,* GYÖRGY CSÓKA† and JOSÉ-LUIS NIEVES-ALDREY‡

*Edinburgh University, Institute of Cell, Animal and Population Biology, The Ashworth Laboratories, The King's Buildings, West Mains Road, Edinburgh EH9 3JT, UK, †Forest Research Institute, Department of Forest Protection, 3232 Mátrafüred, P.O. Box 2, Hungary, ‡Museo Nacional de Ciencias Naturales, José Gutiérrez Abascal 228006 Madrid, Spain

Abstract

The Marble gallwasp *Andricus kollari* has a native range divided into two geographically separated lifecycles. In Eastern Europe and Turkey, the lifecycle involves a sexual generation on Turkey oak, *Quercus cerris*, while in Iberia and North Africa the sexual generation host is cork oak, *Q. suber*. Over the last 500 years, *A. kollari* has expanded its range into northern Europe, following human planting of *Q. cerris* from Italy and the Balkans. We ask: (i) what is the genetic relationship between eastern and western distributions of *Andricus kollari*? Can we determine which lifecycle is ancestral, and how long ago they diverged? (ii) To what extent have eastern and western native ranges contributed to northwards range expansion? (iii) Is there any evidence for hybridization between the two life cycle types? We present analyses of allozyme data for 13 polymorphic loci and of sequence variation for a 433 bp fragment of the mitochondrial cytochrome *b* gene. These show: (i) that four haplotype lineages (one in Spain, two in Hungary/Italy and one in Turkey) diverged more or less simultaneously between 1 and 2 million years ago, suggesting the existence of at least four refuges through recent ice age cycles. Our data cannot resolve which lifecycle type is ancestral. (ii) Populations north of putative refuges are divided into two sets. Populations in south-west France are allied to Spain, while all remaining populations in northern Europe have been colonized from Italy and the Balkans. (iii) The transition from one race to another in south-west France is marked by abrupt transitions in the frequency of refuge-specific private alleles and corresponds closely to the northern limit of the distribution of cork oak. Although hybrids were detected in north-west France, none were detected where the two lifecycles meet in south-western France. The biology of oak gallwasps predicts that any hybrid zone will be narrow, and limited to regions where *Q. cerris* and *Q. suber* meet. Our data suggest that eastern and western *A. kollari* are effectively separate species.

Keywords: *Andricus*, gallwasp, glacial refugia, host race, invasion, range expansion

Received 4 June 2000; revision received 26 August 2000; accepted 26 August 2000

Introduction

Patterns of genetic diversity in many European plants and animals are dominated by histories of postglacial range expansion. During the last glacial, from 115 000 to ≈ 15 000 years ago, many terrestrial organisms were confined to refuges in southern Europe (Huntley 1990; Hewitt 1996).

Correspondence: Dr Graham N. Stone. Fax: 0131 667 3210; E-mail: graham.stone@ed.ac.uk

For many animals and plants, ice age refuges in southern Iberia, Italy, the Balkans and Turkey remain centres of genetic diversity (reviewed by Hewitt 1996, 1999). Refugial populations of many organisms were separated for long enough for refuge-specific alleles and haplotypes to evolve (reviewed by Hewitt 1996, 1999). Screening of such polymorphism in postglacial populations allows assessment of the genetic contribution of particular refuges to northward migration following retreat of the ice sheets. Northern European populations of some species show genetic input

from several refuges, while in others, large areas are dominated genetically by migrants from a single refuge (reviewed by Taberlet *et al.* 1998; Hewitt 1999). Differential genetic contributions from alternative refuges can sometimes be explained by physical features that restrict gene flow, such as the orientation of watersheds for aquatic organisms (Bernatchez & Wilson 1998). In other cases, biotic interactions may be important. For parasites or predators, escape from refuges can only follow range expansion by specific hosts or prey. Where hosts and parasites exist together in subdivided refuge populations, genetic divergence in host populations may lead to associated genetic divergence in parasite populations, the evolution of host-specific ecotypes, and ultimately speciation. If alternative parasite gene pools are available to contribute to northwards range expansion, which one succeeds may then be determined by which refuge or refuges contribute genotypes to the expanding host population.

Most analyses of postglacial range expansion concern events that occurred thousands of years ago, and it is rarely possible to track differential contributions from available refuges as they happen. In these cases, patterns generated by the range expansion process may have been modified by subsequent mutation, drift and selection. This study minimizes the impact of such modification by analysing an example of recent range expansion over a far shorter timescale. The Marble Gallwasp *Andricus kollari* (Hartig, 1843) (Hymenoptera: Cynipidae, Cynipini) is an obligate parasite of oak trees, and native to southern Europe, where it has two discrete lifecycles (described below). This species has invaded northern Europe over the last 500 years following human introduction of an oak host, Turkey oak (*Quercus cerris*). We examine the relationship between the two lifecycle types, and ask which has contributed more colonists to the invasion of northern Europe, and why.

Geographic variation in the lifecycle of A. kollari

As for many oak gallwasps, the lifecycle of *A. kollari* involves obligate cyclical parthenogenesis between a sexual generation in the spring, and a parthenogenetic generation in the summer/autumn (Askew 1984). In most oak cynipids, the galls of the two generations develop on the same oak host. *A. kollari* belongs to a group of gallwasps, found only in the genera *Andricus* and *Callirhytis*, whose lifecycles involve *host alternation* (or heteroecy; Folliot 1964), in which the sexual and parthenogenetic generations develop on different oak taxa (Nieves-Aldrey 1982, 1992; Stone & Sunnucks 1993; Cook *et al.* 1998). The natural distributions of such gallwasps are restricted to regions in which both host oak taxa occur.

A. kollari has a natural distribution divided geographically into two lifecycle types. From Italy east through the Balkans and Turkey the sexual generation develops in tiny bud galls (originally known as *A. circulans* Mayr, 1870) on Turkey oak (*Q. cerris*). The parthenogenetic generation develops in larger 'marble galls' on oaks in the section Quercus (*Q. robur, Q. petraea, Q. pubescens*, and others) (Beijerinck 1902; Marsden-Jones 1953; Folliot 1964). The natural distribution of this lifecycle in Europe is restricted to the area south and east of the Alps where *Q. cerris* and suitable oaks in the section Quercus naturally occur together (Fig. 1).

A. kollari is also found in Spain, Portugal and north-western Africa (Houard 1912; Nieves-Aldrey 1987; Ros-Farré & Pujade 1998). As in the eastern part of its range, the parthenogenetic generation develops in marble galls on oaks in the section Quercus. There is no *Q. cerris* in Iberia, and the sexual generation host is cork oak *Q. suber*. This oak is closely related to *Q. cerris* (Manos *et al.* 1999) and occurs throughout the western distribution of *A. kollari* (Fig. 1). The sexual generation of *A. kollari* in Spain has only been recently identified by careful rearing experiments (Pujade-Villar 1991, 1992; J. Pujade-Villar & R. Folliot, personal communication), and here we use analytical methods applied to other cyclical parthenogens (Sunnucks *et al.* 1997; Simon *et al.* 1999; Gómez & Carvalho 2000) to confirm the presence of sex across our sampled western distribution sites.

Human planting of Q. cerris *and gallwasp range expansion*

Over the last 500 years, *Q. cerris* has been planted extensively north of its native range (Fig. 1), precipitating rapid colonization of northern Europe by four host-alternating *Andricus* species (Docters van Leuwen 1959; Hails & Crawley 1991; Stone & Sunnucks 1993; Stone *et al.* 1995; Schönrogge *et al.* 1998). Three of these species — *A. ambiguus* (formerly identified as *A. corruptrix*), *A. lignicola* and *A. quercuscalicis* — have natural distributions restricted to the native range of *Q. cerris* (Fig. 1), and had a single eastern origin for range expansion into Europe. The fourth species is *A. kollari*, which differs from the other three in that either or both of eastern and western lifecycles could have contributed to northwards range expansion.

The ability of eastern and western lifecycles of *A. kollari* to contribute migrants to range expansion will depend on the ability of each generation in the lifecycle to locate and to induce galls on available oak hosts. The sexual generation must lay eggs, and the asexual generation induce galls, on available oaks in the section Quercus (*Q. petraea, Q. pubescens, Q. pyrenaica* and *Q. robur*). Extensive surveys of chloroplast DNA (cpDNA) (Ferris *et al.* 1993, 1995; Dumolin-Lapègue *et al.* 1997; Petit *et al.* 1997) show section Quercus species across northern continental Europe to contain haplotypes derived from both Iberian and central European refuges, suggesting that both eastern and western lifecycles of *A. kollari* should encounter suitable host genotypes. Evidence from other range expanding *Andricus* also suggests that the refugial origin of section Quercus oaks has little impact on their susceptibility as hosts; gallwasps

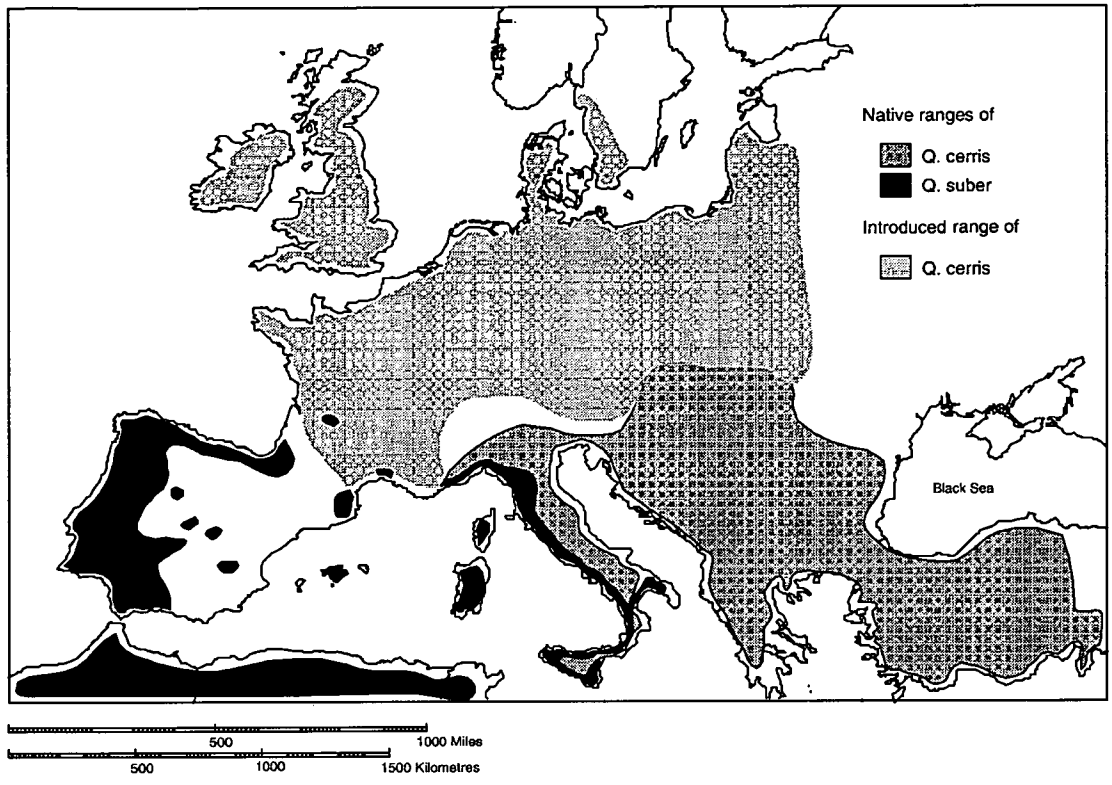


Fig. 1 The natural distributions of *Quercus cerris* and *Q. suber*, and the introduced distribution of *Q. cerris*. Suitable section Quercus oaks are found throughout the distributions of both these oak species. The natural distribution of the eastern lifecycle of *Andricus kollari* corresponds to the native range of *Q. cerris*. The lifecycle involving *Q. suber* is known to occur in Spain, but it is not known if this host is also used in Italy. Oak distributions in Europe are based on maps in Jalas & Suominen (1987) and Toumi & Lumaret (1998), and in Turkey on maps in Yaltirik (1984). *Q. suber* is probably planted in some of the unshaded areas in Iberia, but we know of no published study of its introduced distribution.

originating in the Balkans are able to exploit native British oaks, even though these are derived from the Iberian refuge (Stone & Sunnucks 1993; Dumolin-Lapègue *et al.* 1997; Csóka *et al.* 1998).

The asexual generation must lay eggs, and the sexual generation induce galls, on *Q. cerris* or *Q. suber*. Because cork oak is not found naturally north of south-western France (Fig. 1), and is extremely rare as an introduced tree outside its native range, *Q. cerris* is effectively the only available sexual generation host in northern Europe. Colonization from the eastern native distribution would be by immigrants for which the introduced Turkey oak is the natural sexual generation host. In contrast, range expansion from Iberia would necessitate a host shift from *Q. suber* to *Q. cerris*.

We address the following questions arising from the divided distribution of *A. kollari* and its colonization of northern Europe.

1 What is the genetic relationship between eastern and western lifecycles? Do they represent lineages independently derived from a larger ancient distribution, or has one been founded recently from the other? If so, can we infer which distribution (and which lifecycle) is ancestral, and how long the parts of the native range of *A. kollari* have been divided? If the two populations represent long-separated refugia, we expect refuge-specific genetic differences to have evolved. If one refuge has been recently colonized from the other, genetic variation in the recently colonized region is predicted to be a subset of that found in the source (Nei *et al.* 1975; Nichols & Hewitt 1994). This pattern is seen clearly in comparisons between native and invaded ranges for another invading oak gallwasp (Stone & Sunnucks 1993; Sunnucks & Stone 1996).
2 To what extent have eastern and western native ranges contributed migrants to northwards range expansion following human dispersal of *Q. cerris*?
3 Is there any evidence for gene flow between the two lifecycle types? Is there a hybrid zone where the two lifecycles meet?

We answer these questions through analyses of allele frequency data for 13 polymorphic allozyme loci for 46 populations (1457 individuals) distributed through Spain, France, Holland, Germany, Italy, Hungary and Turkey, and

of sequence data from a subset of 27 individuals for a 433 bp fragment of the mitochondrial cytochrome *b* gene.

Materials and methods

Sampling sites and sample preparation

Sampling sites are shown in Fig. 2, and the number of individuals screened for allozymes for each population is shown in Appendix I (because of its large size, the dataset is available from the Molecular Ecology website, URL http://www.blackwell-science.com/products/journals/suppmat/mec/mec1211/mec1211sm.htm). The western distribution is represented by 284 individuals from eight sites in Spain, while eastern populations are represented by 447 individuals from 12 sites in Hungary, 80 individuals from two sites in Italy and 42 individuals from one site in Turkey. Populations north of putative refuges are represented by 539 individuals from 18 sites in France, and 68 individuals from four sites in Germany and Holland. Sequence data were generated for 27 of the individuals screened for allozymes. Twelve were selected from the eastern and western native distributions — three from Hungary (Gödöllõ), three from Italy (two from Ruffeno, one from Greve), two from Turkey (Antalya) and four from Spain (two from Potes, one from Salamanca and one from Quintanilla). The remaining 15 individuals were from French populations, selected to traverse the apparent genetic divide in France (see below). Ten individuals were drawn from sites north of Bordeaux (two from St. Malo, two from Mortain, five from Forêt de Crécy, one from Nantes) and five individuals from south of Bordeaux (two from Amoux, two from Mugron, one from Bayonne).

All populations were sampled through rearing of parthenogenetic females (body mass 10–15 mg) from marble galls. Individual wasps were stored at −80 °C until required, homogenized in extraction buffer (Peakall & Beattie 1991), and screened immediately as described below. Samples were subsequently stored at −80 °C and used for DNA extraction.

Allozyme screening

An initial screening of commonly used allozyme systems was carried out using cellulose acetate electrophoresis (Zip-zone, Helena Laboratories) on three buffer systems across a range of pH, as described by Stone & Sunnucks (1993). The buffers were 40 mM sodium phosphate (pH 6.3; Stone & Sunnucks 1993), 0.1 M Tris-EDTA-maleate-$MgCl_2$ pH 7.6 (Richardson *et al.* 1986; buffer F), and 25 mM

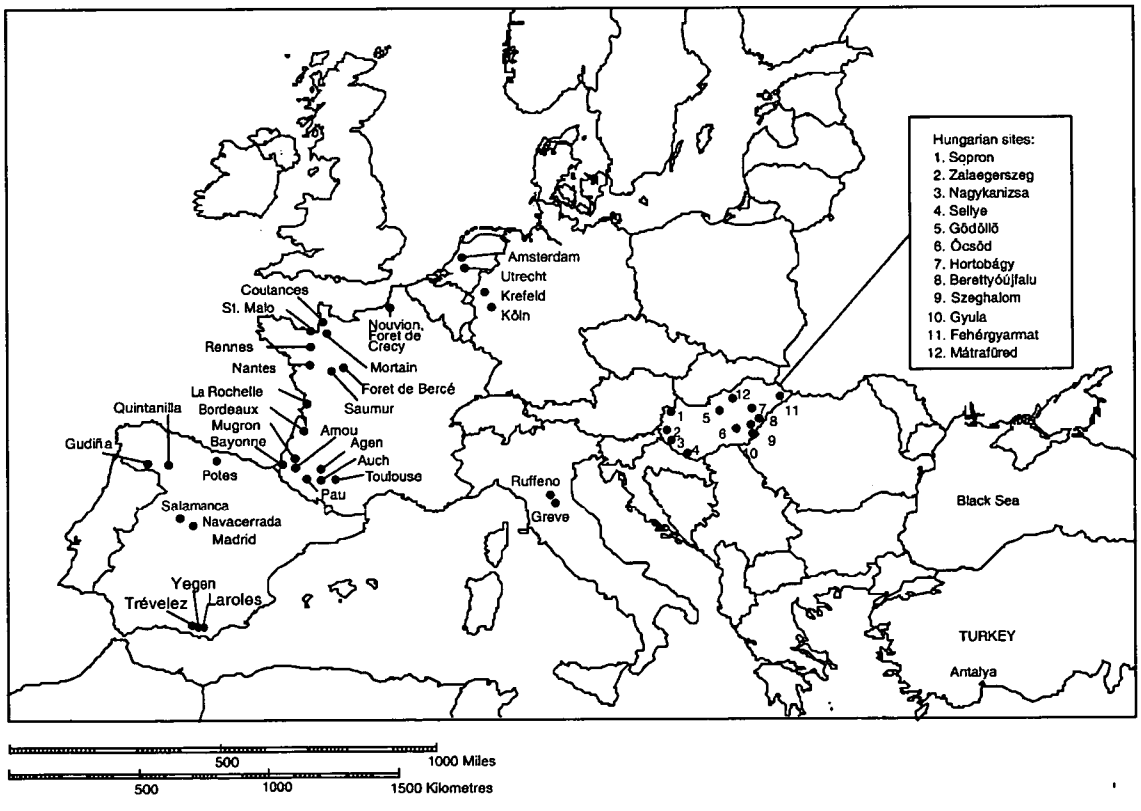


Fig. 2 Sample sites for *Andricus kollari* populations used in this study.

Tris-Glycine pH 8.5 (Richardson *et al.* 1986; buffer I). Substrate staining was carried out using protocols described by Richardson *et al.* 1986. The following 13 polymorphic loci were screened for all individuals, grouped by the running buffer used: AK, αGPD1 and αGPD2, PEP-b (sodium phosphate pH 6.3); GOT-s, GOT-m, GPI, MDH-s, MDH-m, ME, 6PGD (Tris-EDTA-maleate-$MgCl_2$ pH 7.6); HK and PGM (Tris-Glycine pH 8.5).

Analyses of allele frequency data

Summary analyses of allele frequency data were carried out using GENEPOP version 3.1 (Raymond & Rousset 1995) and GENETIX version 4.0 (Belkhir 1999). The generality of a sexual generation in the lifecycle of spanish *Andricus kollari* was checked by analysing the extent of departure from Hardy–Weinberg (HW) and linkage equilibrium. Persistent asexual reproduction, whether apomictic or automictic, is predicted to result in substantial departure from HW equilibrium across all loci. Furthermore, apomixis results in complete linkage across all loci, and automixis commonly results in complete homozygosity (Suomalainen *et al.* 1987; Sunnucks *et al.* 1997; Simon *et al.* 1999; Gómez & Carvalho 2000). Departures from HW equilibrium were tested using exact tests (Weir 1991) incorporated in GENEPOP, with *P* values estimated using a Markov chain method (Guo & Thompson 1992). Significance levels were adjusted for multiple tests using a Bonferroni correction (corrected threshold *P* value = $1 - (1 - \alpha)^{1/k}$ where k is the number of tests and α is the desired threshold value of 5%). Linkage disequilibrium was tested using the linkdos algorithm (Garnier-Géré & Dillman 1992) incorporated into GENEPOP and GENETIX. Significance in the first case is assessed by Fisher's exact tests, and in the second by permutation. We apply these two approaches to minimize the impact on our conclusions of assuming a given underlying distribution of a test statistic. Again, significance levels for multiple tests were adjusted by applying a Bonferroni correction. FSTAT (Goudet 1995) was used to compute *F*-statistics using the method of Weir & Cockerham (1984).

To assess the relationships between populations, and the dependence of such relationships on methods used, dendrograms linking populations were generated for two commonly used genetic distances — Nei's standard genetic distance and Cavalli-Sforza and Edward's chord distance. In the absence of a priori models, the two genetic distances were selected to represent different sets of assumptions concerning the underlying causes of genetic differentiation in *A. kollari*. Nei's unbiased genetic distance (Nei 1978) is formulated for an infinite isoalleles model of mutation, in which there is a rate of neutral mutation and each mutant is to a completely new allele. It is assumed that all loci have the same rate of neutral mutation, and that the genetic variability initially in the population is at mutation-drift equilibrium, with the effective population size of each population remaining constant. Cavalli-Sforza and Edward's chord distance assumes that there is no mutation, and that all gene frequency changes are by genetic drift alone (Cavalli-Sforza & Edwards 1967). This measure does not assume that population sizes have remained constant and equal in all populations. We used three tree building algorithms — maximum parsimony (MP), neighbour-joining (NJ, Saitou & Nei 1987) and maximum likelihood (ML). MP analysis was carried out using the Wagner routine of BIOSYS-1 version 1.2 (Swofford & Selander 1981). NJ was carried out using the NJBP program (Jean-Marie Cornuet, INRA Laboratoire de Modélization et Biologie Évolutive, Montpellier), and ML estimation was carried out using PHYLIP (Felsenstein 1993). For NJ trees, bootstrapping was carried out over both populations and loci, with 1000 bootstrap replicates. Because of the dramatic increase in computational complexity associated with ML estimation for a large number of populations, we used 100 bootstrap replicates for this method.

Cytochrome b *sequencing and phylogenetic analysis*

DNA was extracted from single adults using Proteinase K/SDS digestion followed by salting out of protein and precipitation of DNA in ethanol. Each dried DNA pellet was then resuspended in 50 μL of pure water. A 433 bp sequence of the cytochrome *b* gene was amplified using the primers CB1 (5′–3′ TATGTACTACCATGAGGACAAATATC) and CB2 (5′–3′ ATTACACCTCCTAATTTATTAGGAAT) (Jermiin & Crozier 1994; Stone & Cook 1998) in polymerase chain reaction (PCR) (35 cycles of denaturation at 92 °C for 60 s, annealing at 50 °C for 60 s, and extension at 72 °C for 120 s). PCR reactions had a total volume of 25 μL, with 1.0 μL DNA extract, 2.5 μL of PARR buffer (BIORAD), 1 μL 50 mM $MgCl_2$ (final concentration 2 mM), 0.5 μL 10 mM dNTP's, 0.35 μL of the primers CB1 and CB2 (each at 20 mM), 0.25 μL of *Taq* (Promega), and the balance made up with pure autoclaved water. All PCR and sequencing reactions were carried out on a PTC-200 DNA engine (MJ Research). The PCR product was gel-purified by electrophoresis through a 1% agarose gel stained with ethidium bromide. Gel fragments were excised from lanes showing a band at the correct size, and DNA extracted using the QIAquick gel extraction kit (QIAgen cat. 28704). Sequencing was direct from the purified PCR product. Mitochondrial genes are known to have nuclear pseudogene copies in insects (Sunnucks & Hales 1996; Zhang & Hewitt 1996; Bensasson *et al.* 2000), and if sequencing from an apparent single band suggested multiple copies of the cyt *b* gene, specific DNA sequences from particular individuals were amplified by cloning. When necessary, 2–4 μL of a PCR were used in standard ligation and transformation reactions using the TOPO TA cloning kit (Invitrogen cat. K4500-01). Plasmid DNA was

purified using the QIAprep spin miniprep kit (QIAgen cat. 27104). Sequencing was carried out using Perkin-Elmer Big Dye Terminator chemistry and an ABI 377 sequencer.

Sequences were all 433 bp long, and were aligned using CLUSTAL-W (Thompson *et al.* 1994). All sequences were found to possess the same reading frame, and contained no stop codons. Sequences for 27 individual *A. kollari* are deposited in GenBank, with acquisition numbers AF242739–AF242762 and AF242764–AF242766. Phylogenies were generated by MP and ML using test version 4.0b3 of PAUP* (Swofford 1998). We do not present NJ analyses of the sequence data for two reasons. First, MP and ML have been demonstrated to recover more closely a single true topology in simulation studies (Huelsenbeck & Hillis 1993). Second, ML methods allow the testing of specific models of substitution, and the identification of a model most appropriate to a given dataset (for a discussion of the impact of substitution models on phylogeny construction see Swofford *et al.* 1996 and Yang 1996). We used the program Modeltest 3.0 (Posada & Crandall 1998) to identify the most appropriate substitution model for our data. The model supported both by ML ratio tests and on the basis of Akaike information criterion was the Hasegawa–Kishino–Yano (HKY) model, which allows unequal transition and transversion ratios and unequal base frequencies (Hasegawa *et al.* 1985) with corrections for among-site variation (HKY + G + I). The following parameters returned by Modeltest were used in PAUP*: Transition and transversion (Ti/Tv) ratio = 18.23. Base frequencies A = 0.3492, C = 0.1183, G = 0.1102, T = 0.4223. Proportion of invariable sites = 0.733. Changes are gamma distributed, with a shape parameter 0.5123. Trees were rooted by using sequence for two other members of the *A. kollari* species group — *A. polycerus* (GenBank accession number AJ228457) and *A. conglomeratus* (GenBank accession number AJ228568) — known to be close relatives in a sister-clade to the *A. kollari* species group based on an earlier sequence-based phylogenetic analysis of the genus (Stone & Cook 1998). All trees were generated using 100 random additions in a heuristic search, using the tree bisection-reconnection (TBR) algorithm of PAUP. Ten trees were held at each step, with positions weighted equally. For both tree-building methods, we generated 100% consensus trees and subjected them to bootstrap analysis using full heuristic searches for 100 replicates. With the exception of minor rearrangements of terminal taxa, the tree topologies recovered by MP and ML were identical, with significant bootstrap values for the same nodes. Genetic distances given are for the Kimura 2-parameter model, which allows comparison with a large number of existing studies.

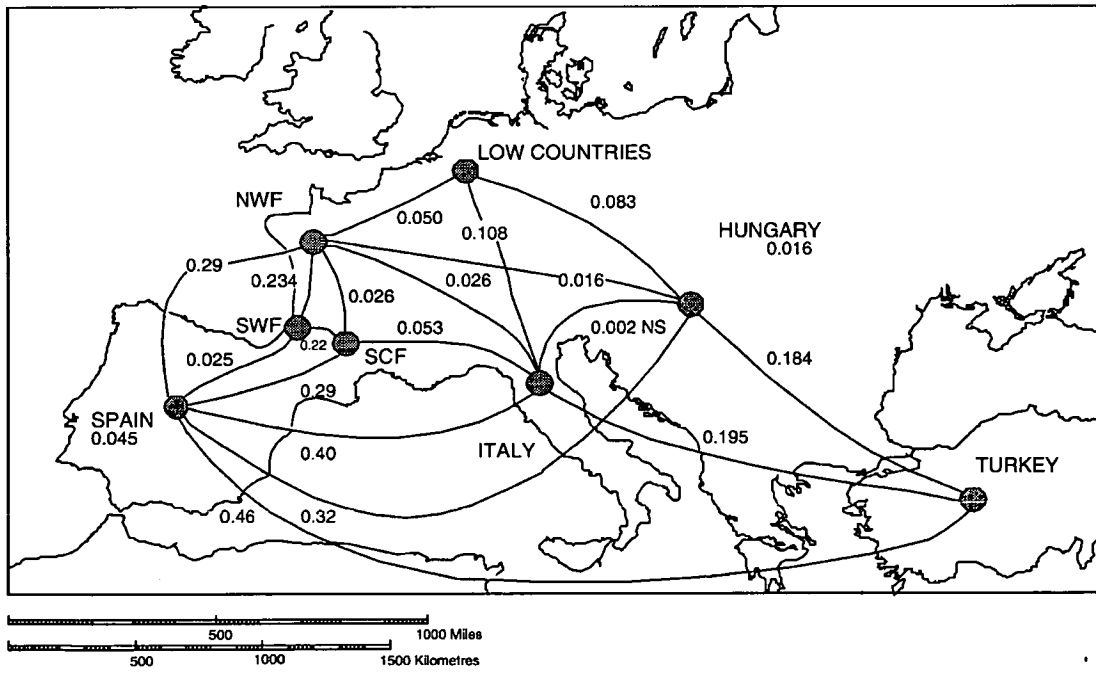


Fig. 3 Genetic subdivision between subsets of *Andricus kollari* populations, as indicated by pairwise F_{ST} calculations. All values except that between Italy and Hungary are highly significant by the permutation testing procedure in GENETIX. Each shaded circle represents a set of populations pooled together for pairwise F_{ST} between regions. Abbreviations for summed populations are as follows: NWF = north-west France (Coutances, Forêt de Bercé, Forêt de Crécy, La Rochelle, Nantes, Nouvion, Rennes, Saumur, St. Malo); SCF = south central France (Agen, Auch, Toulouse); SWF = south-west France (Amoux, Bayonne, Mugron, Pau); Low Countries = Utrecht, Amsterdam, Krefeld, Köln and data for 18 individuals from several sites in Belgium. Mean pairwise F_{ST} values are also included among populations within each of Hungary and Spain.

Results

Analyses based on allozyme allele frequency data

Relationships between the eastern and western native distributions. The 13 polymorphic loci possessed a total of 62 alleles over all 46 populations, with frequencies summarized for each population in Appendix I (URL http://www.blackwell-science.com/mec/). Hungary, Italy and Spain all possess regionally private alleles, and neither the eastern nor the western distributions is a simple genetic subset of the other. The only area not to have private alleles is Turkey, whose single sampled population contains a subset of the alleles found in Hungary. Four alleles were found only in Hungary (GOT-s allele 5, GPI allele 5, MDH-s allele 4 and PGM allele 5), one allele was found only in Italy (AK allele 5), and two alleles were found only in Spain (MDH-m allele 3 and AK allele 4). A further three alleles were found almost entirely in Spain, with only a single copy of each allele found in central and eastern European populations; αGPD1 allele 1 had a single copy in Sopron (north-west Hungary), αGPD1 allele 2 had a single copy in Gyula (south-east Hungary) and αGPD1 allele 3 had a single copy in Ruffeno (Appenines, Italy).

Several analytical methods suggest that Italian and Hungarian populations are far more similar to each other than either are to Spain or to Turkey. First, Hungarian and Italian sites share 17 alleles absent from Spain, while Spain and Italy share only one allele, αGPD1 allele 3, absent from Hungary (furthermore, as noted above only one copy of this allele is found in Italy). Second, F_{ST} between pooled Italian and pooled Hungarian sites (0.002, nonsignificant by permutation testing) is far lower than that between either and Spain or Turkey (Fig. 3). Third, a close relationship between Italy and Hungary, and their genetic divergence from Spain, is supported by analyses based on allele frequencies (Fig. 4) for all three tree-building methods and genetic distances used. Bootstrap support for the divide is more than 95% by ML or NJ, and for the latter the same result is obtained whether bootstrapped over loci or individuals.

Population genetic support for a sexual generation in western distribution Andricus kollari. All Spanish populations were in HW equilibrium, confirming the generality of a sexual generation in the *A. kollari* lifecycle. The small number of significant departures from HW equilibrium occurred elsewhere in the sampled range (Bordeaux for aGPD1, Bayonne for GOT-s, Fehergyarmat for GOT-m, and Mátrafüred, Agen, Forêt de Bercé and Mortain for PEP-b).

Significant linkage disequilibria, both across sets of populations (global tests) and for individual populations, are shown in Table 1. Both eastern and western populations show some significant linkage, although for no pair of loci

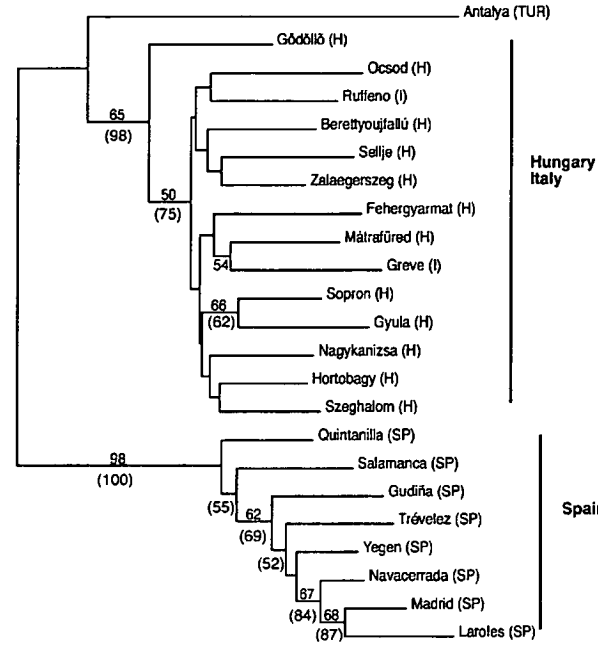


Fig. 4 The genetic divide between eastern and western native distributions in *Andricus kollari*, as demonstrated by allozyme allele frequency data. The phylogram shown is generated by neighbour-joining, using Cavalli-Sforza and Edward's chord distance. Numbers at nodes are bootstraps, over individuals, expressed as percentages for 1000 replicates. Values in parentheses are bootstraps for the same node obtained by maximum likelihood. Nodes without bootstraps are supported by less than 50% of replicates.

is linkage supported by both permutation and exact tests across all populations in a set. Some linkage relationships are supported for both Spanish and Hungarian sets of populations (those linking GOT-m with PEP-b, PGM and 6PGD, and PGM with 6PGD), while others are only supported for one set of populations (Table 1).

Postglacial range expansion by A. kollari. Sites north of the glacial refuges for *A. kollari* cluster into two clear groups on the basis of allozyme allele frequency data (Fig. 5), supported by bootstrapping for ML and NJ trees, and across both loci and individuals for NJ trees. Populations in the south-west of France (Bordeaux, Toulouse, Pau, Mugron, Amoux, Bayonne) cluster with Spain, while all other sites in northern Europe are more closely associated with Hungarian and Italian sites. The same result is shown by pairwise F_{ST} calculations among groups of populations (Fig. 3). All populations north of the Pyrenees and Alps, bar those in south-western France, show only slight (but nevertheless significant) genetic differentiation from Hungary/Italy, but major differentiation from Spain (Fig. 3). The only populations in northern Europe to show more substantial differentiation from Italy and Hungary are those in Germany, north-central France, and the Low Countries.

Locus 1	Locus 2	Global permutation	Global exact test	Individual populations
Hungarian populations ($n = 12$)				
GOT-s	GOT-m	No	No	Mátrafüred, Nagykanizsa
GOT-s	PGM	No	No	Mátrafüred
GOT-m	HK	No	No	Hortobagy
GOT-m	PEP-b	No	$P < 0.01$	Mátrafüred(*), Ocsod
GOT-m	6PGD	No	$P < 0.01$	Hortobagy
GOT-m	PGM	No	$P < 0.05$	
GPI	MDH-s	No	No	Szeghalom
GPI	6PGD	No	No	Ocsod, Szeghalom
HK	PGM	No	No	Mátrafüred
ME	PGM	$P < 0.05$	No	Fehergyarmat
PEP-b	6PGD	No	$P < 0.05$	Hortobagy
PGM	6PGD	No	No	Mátrafüred
Spanish populations ($n = 8$)				
aGPD1	PGM	$P < 0.05$	No	
aGPD1	6PGD	$P < 0.05$	No	
GOT-m	PEPb	No	No	Madrid
GOT-m	PGM	$P < 0.05$	No	Salamanca
GOT-m	6PGD	$P < 0.05$	No	Salamanca
PGM	6PGD	$P < 0.05$	No	Quintanilla

Table 1 Tests of linkage disequilibrium for each pair of loci over all populations (global tests) and within each population. Results are given for permutation tests in GENETIX and for Fisher's exact tests in GENEPOP. Named populations show significant results in permutation tests, and if also significant in exact tests are followed by an asterisk (*)

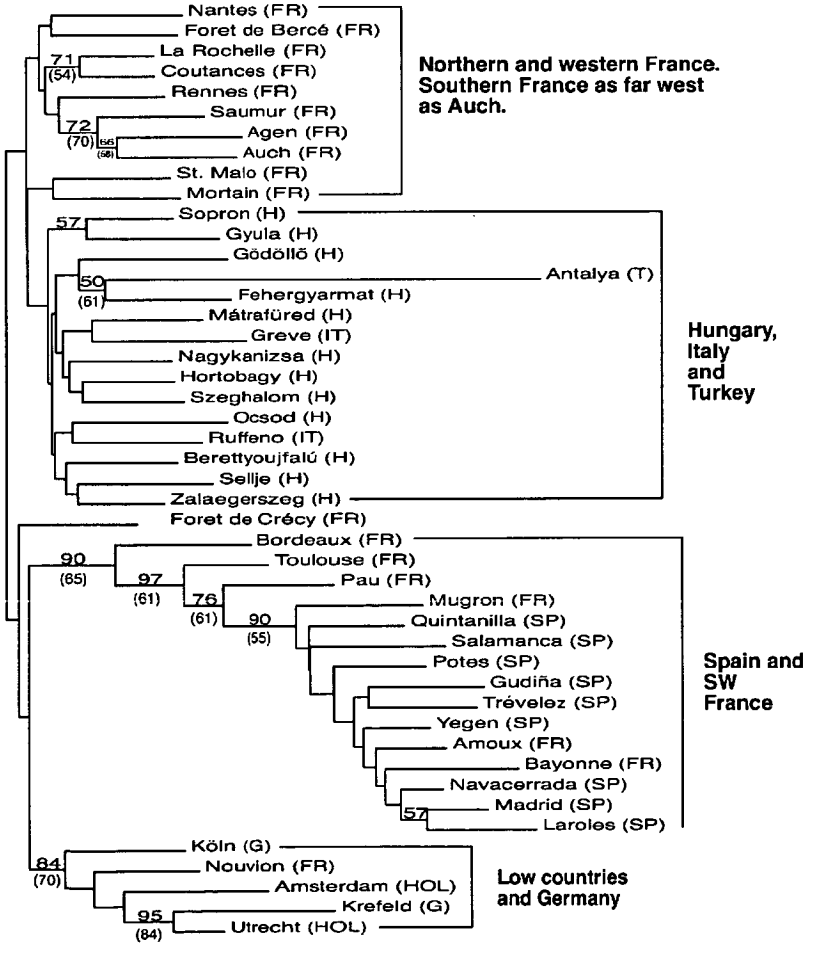


Fig. 5 A phylogram of relationships, based on allele frequency data, among *Andricus kollari* populations from both native and recently invaded regions. Relationships are generated by neighbour joining of Cavalli-Sforza and Edward's chord distance. Numbers at nodes are bootstraps, over individuals, expressed as a percentage of 1000 replicates. Values in parentheses are bootstraps for the same node obtained by maximum likelihood. Nodes without bootstraps are supported by less than 50% of replicates.

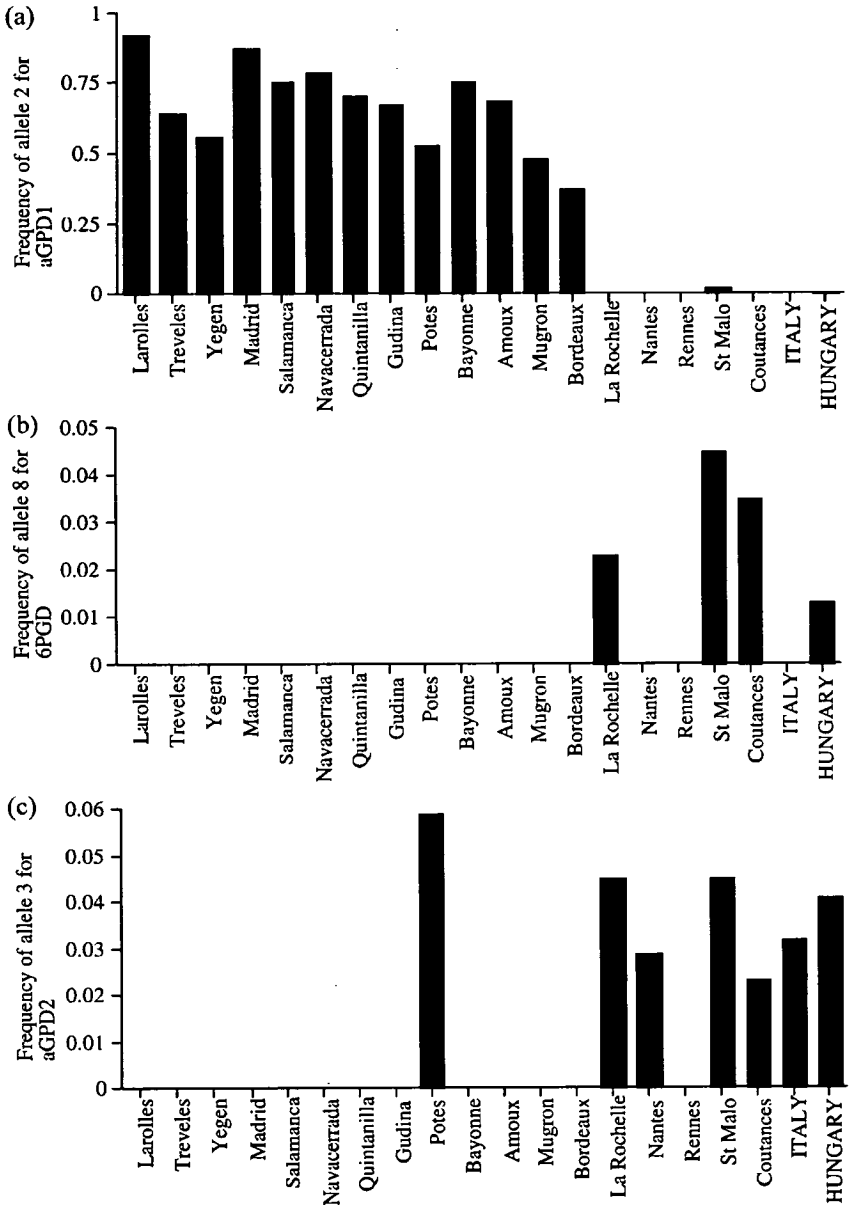


Fig. 6 Changes in the frequencies of three regionally private alleles indicating the genetic division between western and eastern gene pools in *Andricus kollari*. Sites are organized from south to north in a transect running from the extreme south of Spain to the channel coast of France (see Fig. 2 for locations).

These form a clade with high bootstrap support (Fig 5), and experience lower estimated gene flow with Italy and Hungary (pairwise F_{ST} 0.108 and 0.083, respectively) than populations in western or southern central France (Fig. 3).

The sharpness of the spatial transition from populations with Iberian affinities to those closer to central and eastern Europe is striking. Populations in north-western France, and north of the central Pyrenees (Auch, Agen and Toulouse) are all substantially differentiated from sites at the west of the Pyrenees (F_{ST} = 0.22–0.23) (Fig. 3). While French populations north of the central Pyrenees show little apparent gene flow with Spain (F_{ST} = 0.29), those at the west of the Pyrenees are closely allied to Spain (F_{ST} = 0.025). The differentiation is due mainly to abrupt changes in the frequency of several regionally private alleles, illustrated along a south-to-north series of sites in Fig. 6. In particular, allele 2 at αGPD1 falls from a high frequency (60–80%) in Spain and south-west France to zero at sites less than 100 km to the north.

Analyses of mitochondrial DNA (mtDNA) sequence data

Of the 433 bp in the amplified fragment of cytochrome *b*, 375 were constant across all sequences. Of the 58 variable

Table 2 Haplotypes for the 433 bp fragment of cytochrome *b* amplified in *Andricus kollari*. Parsimony-informative sites are indicated by an asterisk in the first row of the table

					11	111	111	111	122	222	222	222	222	223	333	333	333	333	333	4
	13	445	667	778	811	222	344	789	900	012	334	447	788	990	000	222	233	477	899	2
	702	695	790	392	625	147	025	213	625	840	281	454	706	282	458	235	967	069	279	1
Parsimony informative?	**	*	*	**	**		**	***	*	***	**	**	*		**	**	***	*	*	*	*	*
Haplotype 1	AAA	AAG	ATT	AAA	TTA	CTA	TAT	TAA	ATT	AAT	CTT	TCA	ATT	TTT	ATT	CCT	AGA	CTT	TAG	A
Haplotype 2	...	.G.	...	...	...	...	...	...	G..	...	...	...	...	...	...	...	...	...	...	.
Haplotype 3	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	T..	...	.
Haplotype 4	...	...	...	...	...	...	...	...	.C.	.G.	...	...	...	...	...	...	?..	T..	...	.
Haplotype 5	.G.	..A	...	...	...	...	...	.G.	...	.G.	...	...	...	...	...	..C	...	.?.	...	.
Haplotype 6	.G.	G.A	...	...	...	TC.	...	..T	G..	...	...	...	...	...	...	T.C	...	T..	CG.	G
Haplotype 7	.G.	G.A	...	...	...	TC.	...	..T	G..	...	...	...	...	...	...	T.C	...	T..	CG.	.
Haplotype 8	.G.	..A	.C.	..G	...	...	...	...	GCC	.G.	...	...	...	C..	...	TTC	G..	T..	...	.
Haplotype 9	.G.	..A	.C.	..G	...	...	...	...	GCC	.G.	...	...	...	C..	...	TTC	G..	T..	...	.
Haplotype 10	.G.	..A	.C.	..G	C..	...	...	...	GCC	.G.	...	...	...	C..	...	TTC	G..	T..	...	.
Haplotype 11	.GG	..A	.C.	..G	...	...	...	...	GCC	.G.	...	...	...	C..	...	T.C	G..	T..	...	.
Haplotype 12	.G.	..A	.C.	..G	...	...	...	...	GCC	.G.	...	...	...	C..	...	T.C	G..	T..	...	.
Haplotype 13	.G.	..A	.C.	...	...	...	...	...	GCC	.G.	...	...	...	C..	...	T.C	G..	T..	...	.
Haplotype 14	.G.	..A	.C.	...	...	...	...	...	.CC	.G.	...	...	...	C..	...	T.C	G..	T..	...	.
Haplotype 15	...	..A	.C.	...	.C.	..T	...	...	.CC	.G.	...	.T.	...	CC.	.CC	T.C	GA.	T..	..A	.
Haplotype 16	GG.	..A	...	.G.	...	...	C..	...	..C	.G.	.CC	C..	...	...	..C	T..	...	T..	...	.
Haplotype 17	.G.	..A	...	.G.	...	...	C..	...	..C	.G.	.CC	C..	...	...	..C	T..	...	T..	...	.
Haplotype 18	GG.	..A	...	.G.	...	...	C..	...	..C	.G.	.CC	...	...	...	..C	T..	...	...	...	.
Haplotype 19	GG.	..A	...	.GG	...	...	C..	...	..C	.G.	.CC	...	...	...	..C	T..	...	...	...	.
Haplotype 20	GG.	..A	...	.G.	...	...	C..	...	..C	.G.	.CC	...	...	...	..C	T..	...	T..	...	.
Haplotype 21	GG.	..A	...	.G.	...	...	C..	...	G.C	.G.	.CC	...	...	...	..C	T..	...	T..	...	.

Haplotype locations: haplotypes 1–15 are from the eastern distribution and associated sites in north-west France, haplotypes 16–21 are from Spain and associated sites in south-west France. 1 – Gödöllõ (H); 2 – Greve (I); 3 – Gödöllõ (H), St. Malo 37 (NW Fr); 4 – Foret de Crécy (NW Fr); 5 – Gödöllõ (H); 6 – Antalya (T); 7 – Antalya (T); 8 – Foret de Crécy (NW Fr); 9 – Ruffeno (I); 10 – Ruffeno (I); 11 – Foret de Crécy (NW Fr); 12 – Nantes (NW Fr); 13 – Foret de Crécy, Mortain (NW Fr); 14 – Mortain, St. Malo (NW Fr); 15 – Nantes (NW Fr); 16 – Potes (SP), Mugron, Amoux, Bayonne (SW Fr); 17 – Mugron (SW Fr); 18 – Amoux (SW Fr); 19 – Quintanilla (SP); 20 – Salamanca (SP); 21 – Potes (SP).

sites, 35 were informative in parsimony reconstruction. The 27 individuals sequenced yielded 20 haplotypes (Table 2).

Relationships between the eastern and western native distributions. Both tree-building approaches generate best fit trees with the same topology (Fig. 7), with the four Spanish sequences placed as a monophyletic clade within a paraphyletic group of eastern distribution sequences. This nesting of western within eastern distribution sequences is not, however, supported by bootstrap analysis (Fig. 7). While four lineages retain high bootstrap support (one contains the Spanish sequences, one contains the two Turkish sequences, and the remaining two contain sequences from Italy and Hungary), it is not possible to say whether eastern or western lineages are closer to the outgroup. If the HKY + G + I model is assumed, the level of sequence divergence between eastern and western distribution sequences ranges between 4.0 and 10.4%, close to that observed among the eastern lineages of 4.1–7.4% between any two of Ruffeno (Italy), Antalya (Turkey) and Gödöllõ (Hungary). If we use an estimate of the rate of mitochondrial sequence divergence, based on insects and other invertebrates, of 2.3% per million years (Brower 1994), these four lineages diverged from each other between 1.75 and 4.5 million years ago. Use of Kimura 2-parameter distances (a simpler model rejected for our data by the Modeltest program) gives a sequence divergence between the four lineages of 2.8–3.8%, and an estimated time since divergence of between 1.2 and 1.6 million years, still long before the end of the Pleistocene.

Postglacial range expansion by A. kollari. Sequences from sites north of putative glacial refuges fall into three of the four lineages established for the native range sequences. For both ML and MP methods, all sequences from sites in northern France lie within one of the two lineages containing sequences from Italy and Hungary (left-hand figure, Fig. 8), while sequences from south-west France cluster with those from Spain. Thus, the DNA results agree with site groupings

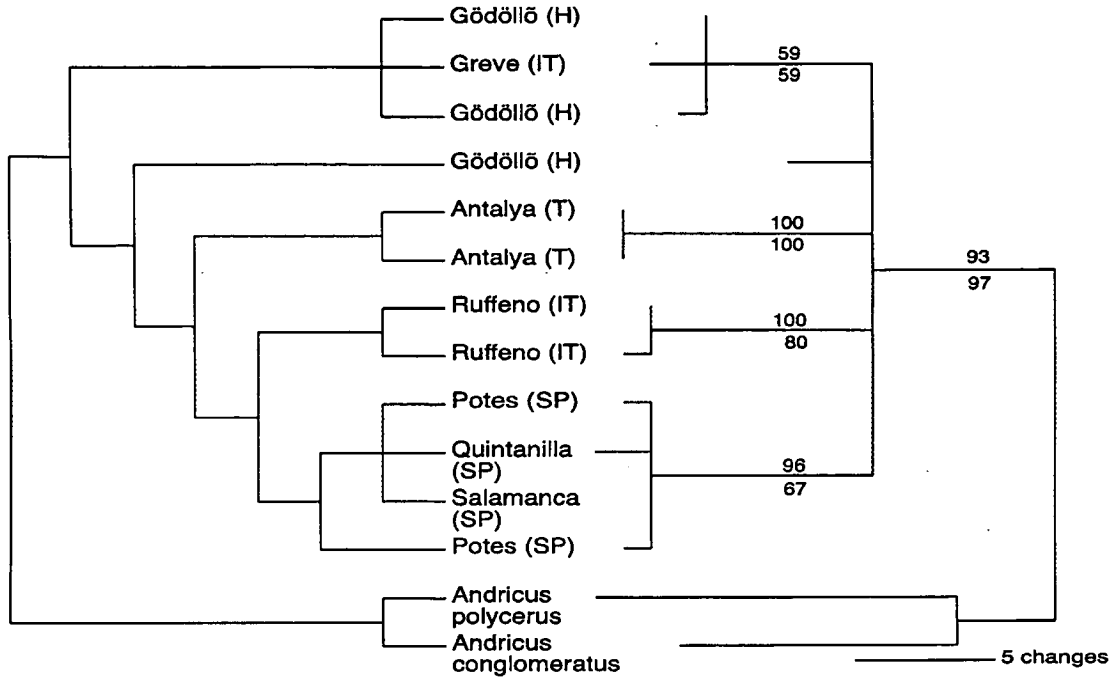


Fig. 7 Relationships, based on cytochrome *b* sequence data, among eastern and western distribution *Andricus kollari*. The left-hand cladogram represents the strict consensus of the two shortest trees generated using maximum parsimony in PAUP (length 69 steps, CI = 0.797, RI = 0.816). The best tree generated using maximum likelihood has almost exactly the same topology (–log likelihood = 902.4), with reversal of the relative positions of two clades (one containing Godollo and Greve, the second containing the two sequences from Ruffeno). The right-hand phylogram shows the majority rule consensus tree for 100 bootstrap replicates generated by maximum likelihood (–log likelihood = 903.93). Maximum parsimony returns an identical bootstrap consensus tree (length 75 steps, CI = 0.733, RI = 0.737). Numbers on the cladogram are bootstrap percentages, for maximum parsimony above the branch and maximum likelihood below.

on the basis of allele frequency data. Although this larger dataset again suggests that eastern sequences may be basal to western sequences (left-hand phylogeny in Fig. 8), bootstrap analysis shows membership of the main lineages to be only weakly supported for some northern French sequences, and the branching order of the main lineages is unresolved (right-hand phylogeny in Fig. 8). Similarity between Spanish sequences and those from south-western France remains strongly supported. Thus, addition of further sequences does not improve our ability to infer which lifecycle is ancestral and which derived.

Can eastern and western A. kollari *hybridize?*

Although the genetic transition between eastern and western distributions in south-western France is relatively abrupt (Fig. 6), allozyme data identified no hybrids near the contact area, and no hybrid zone. Only a single individual (from Mortain, north-west France) could be identified as a likely hybrid on the basis of allozyme data. This individual possessed alleles diagnostic of both east and west, combining allele 2 at αGPD1 (diagnostic of the western distribution) with allele 6 at 6PGD (diagnostic of the eastern distribution). A further three individuals from north-west France (one from each of St. Malo and Forêt de Crécy, and a second individual from Mortain) also possessed allele 2 for αGPD1. If we assume that this allele evolved only once, these individuals must be the result of long-range gene flow from the western distribution. The western distribution alleles could be present either in the offspring of purely Spanish immigrants, or in hybrids. Two possible hybrids (the two individuals from Mortain) were sequenced for the cytochrome *b* fragment. Each possessed a haplotype (numbers 13 and 14, Table 2) indicating origin in the eastern distribution. Because their haplotype is maternally inherited, these individuals must have acquired their western distribution αGPD1 allele by mating of a western distribution male with an eastern distribution female.

Discussion

Genetic relationships among native range populations of Andricus kollari

Allozyme and DNA sequence data suggest that eastern and western distributions of *Andricus kollari* have been

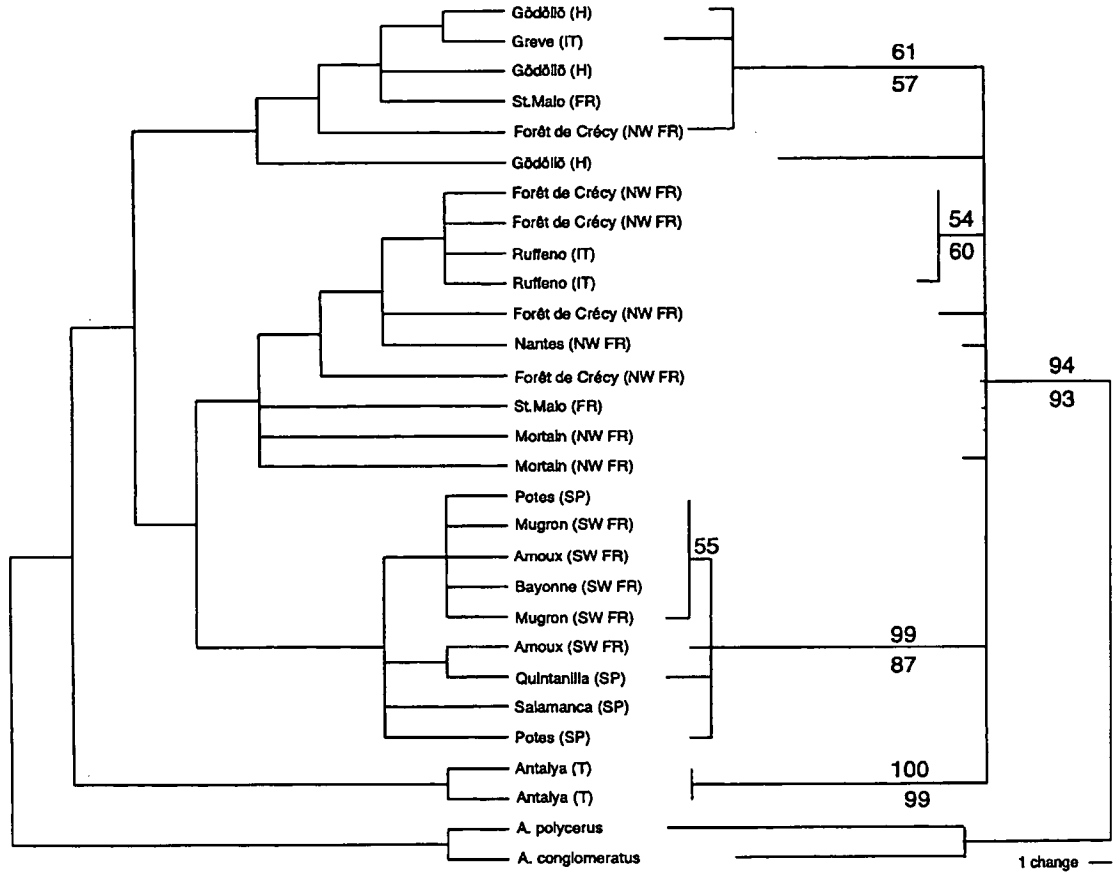


Fig. 8 Relationships, based on cytochrome *b* sequence data, among *Andricus kollari* populations from both native and recently invaded regions. The left-hand cladogram is the strict consensus of 55 shortest trees generated by maximum parsimony (length 73 steps, CI = 0.767, RI = 0.902). With the exception of a single trivial rearrangement, the topology of the best maximum likelihood tree is identical (–log likelihood = 934.41). The right-hand phylogram shows the majority rule consensus tree for 100 bootstrap replicates generated by maximum likelihood (–log likelihood = 959.46). The bootstrap topology returned by maximum parsimony is identical (length 91 steps, CI = 0.615, RI = 0.798). Numbers on the cladogram are bootstrap percentages, for maximum parsimony above the branch and maximum likelihood below.

genetically separated for a substantial period of time. Both regions contain private allozyme alleles, and each consists of a mutually exclusive set of cytochrome *b* haplotypes. It is probable, therefore, that the Iberian peninsular has acted as a discrete glacial refuge for several glacial cycles prior to the present interglacial. The small number of sequences sampled to date further suggest that three lineages diverged in the eastern distribution at approximately the same time that Iberian populations became isolated, perhaps corresponding to simultaneously isolated refugial populations in Italy, the Balkans and Turkey.

The complete separation of mtDNA lineages between east and west suggests greater genetic subdivision than is suggested by F_{ST} values calculated from allozyme data. While this difference may in part be due to the difference in sampling effort between mitochondrial and nuclear markers, it may also reflect the larger effective population size, and longer coalescence times, of nuclear alleles relative to mitochondrial markers. This means that allozyme alleles common to both regions (which tend to reduce F_{ST}-based estimates of differentiation) may represent ancestral polymorphisms retained in both regions, rather than the results of contemporary gene flow.

Population genetic evidence for the generality of a sexual generation in the lifecycle of A. kollari

Many European oak gallwasps appear to lack sexual generations in their lifecycle (Ambrus 1974; Nieves-Aldrey 1987; Csóka 1997). In most cases, inference of a purely parthenogenetic lifecycle is due to ignorance of any sexual generation, rather than demonstrated persistence of purely parthenogenetic reproduction. Painstaking rearing experiments have shown that the asexual generation of at least one oak gallwasp, *A. quadrilineatus*, includes both females able to give rise directly to a second generation of parthenogenetic

females, and females giving rise to a sexual generation (Folliot 1964). Such intraspecific variance in reproductive mode makes it possible that a species that has a sexual generation in one part of its range could lack it in another (geographical parthenogenesis). Geographical parthenogenesis has been proposed for three oak gallwasps (Yasumatsu 1951; Bailey & Stange 1966; Abe 1986), and is known to occur in rose gallwasps (Stille 1985a,b; Plantard *et al.* 1998, 1999). Though geographical parthenogenesis was proposed as a possibility for Iberian *A. kollari* (Csóka *et al.* 1998), allele frequency data for Spanish populations strongly support Pujade-Villar's (1991, 1992) conclusion that *A. kollari* has a sexual generation in Iberia. Many other gallwasps are known only from asexual generations, and application of a similar population genetic approach should reveal which are true purely parthenogenetic species, and which are half of a cyclically parthenogenetic lifecycle (Atkinson 2000).

Which lifecycle is ancestral?

A taxonomically diverse set of gallwasp species in the genera *Andricus*, *Callirhytis*, *Neuroterus* and *Synophrus* exploit *Quercus cerris* in the east and *Q. suber* in the west (Trotter & Cecconi 1904; Dalla Torre & Kieffer 1910; Nieves-Aldrey 1987, 1992). This raises two general questions. First, which represents the ancestral sexual generation host for each of these species? Do they share a common ancestral host? Second, did the shifts from one host to another occur at more or less the same time for the whole set, perhaps driven by a common set of circumstances, or have they occurred at different times in different species? Our analysis of *A. kollari* is a first step towards answering these broader questions of the origin of host association in oak cynipids.

Our cytochrome *b* data cannot resolve which of eastern or western distributions gave rise to the other. Although both MP and ML analyses suggest that Spanish cytochrome *b* sequences are derived with respect to an eastern distribution ancestor (left-hand figures in Figs 7, 8), there is no bootstrap support for this conclusion. What is clear is that lineages currently associated with *Q. suber* and *Q. cerris* diverged 1–2 million years ago, and thus the host shift, whichever its direction, may be relatively ancient.

An alternative approach to the question of which sexual generation host is ancestral is to take a phylogenetic view of the occurrence of *Andricus* lifecycles involving the two oak species. At least five host-alternating *Andricus* species have a sexual generation on *Q. cerris* where this oak is native. Three are members of a single closely related clade (*A. corruptrix*, *A. kollari*, and *A. lignicola*), while two are members of more distantly related species groups (*A. quercuscalicis* and *A. gemmea*) (Stone & Cook 1998; Cook *et al.* 1998). Of these six, only two (*A. kollari* and *A. gemmea*) are found in Iberia. In contrast, no host-alternating *Andricus* species are yet known that are restricted to Iberia. Taken at face value, this distribution would suggest that host alternation evolved in the native range of *Q. cerris*, followed by a host shift to *Q. suber* by one (or possibly two) species. However, this pattern may merely reflect a lack of knowledge of lifecycles in Iberian cynipids, and further work on this issue is required before conclusions can be drawn.

Origins of populations of A. kollari *north of glacial refuges*

Allozyme and sequence data show that populations north of the Pyrenees, Alps and Tatras are divided into two sets. We suggest that these sets are probably associated with escape from refuges on two different timescales.

Populations in south-western France. Populations in the extreme south-west of France are genetically indistinguishable from Spanish populations, sharing the same private allozyme alleles and cytochrome *b* haplotypes. This region of south-west France corresponds closely to the distribution of *Q. suber* at the western end of the Pyrenees (Fig. 1), and is compatible with a dependence of western distribution *A. kollari* on this host. As described above, *Q. suber* occupied a refuge in Iberia during the last glacial period, and populations in southern France are the result of natural range expansion from Iberia following retreat of the ice. Because the level of sequence divergence between Iberia and the other refuges suggests divergence long before the current interglacial, *A. kollari* has probably been associated with *Q. suber* in Iberia for multiple glacial cycles. Therefore, populations of *A. kollari* in south-western France probably represent the results of natural postglacial range expansion by Iberian *A. kollari* already dependent on *Q. suber*, 8000–10 000 years ago. The northern limit to this range expansion was probably imposed by the inability of *Q. suber* to spread beyond the geographical limits of the hot humid or subhumid Mediterranean climates it requires (Camus 1938).

Populations across the rest of northern Europe. All other populations in northern Europe show close affinity to populations in Italy and Hungary. Haplotypes from populations in north-western France lie in both of the lineages found in Italy and Hungary, and pairwise F_{ST} values indicate high gene flow to these refuges (Fig. 3). Exploitation of planted *Q. cerris* has apparently been entirely by *A. kollari* migrants from areas in which this oak is the natural host of the sexual generation. This confirms that, as for other oak gallwasps (Stone & Sunnucks 1993; Sunnucks & Stone 1996; Csóka *et al.* 1998), any geographical variation across northern continental Europe in available genotypes of section Quercus hosts has not prevented range expansion from eastern refuges.

Are there host races in A. kollari?

If Iberian *A. kollari* spread into south-western France following natural dispersal of cork oak, they were there for thousands of years before human introduction of *Q. cerris* and would have been exposed to Turkey oak immediately following its introduction. In contrast, immigrants from the east could only reach south-western France following dispersal across other *Q. cerris* patches, resulting in a substantial time lag (Stone & Sunnucks 1993). Despite a head start, western distribution *A. kollari* have remained restricted to the distribution of *Q. suber*, suggesting that Iberian asexual generation females may be unable to exploit *Q. cerris*.

Exploitation of novel *Q. cerris* requires: (ii) that asexual generation females adapted to location of *Q. suber* can locate *Q. cerris*, and receive suitable oviposition stimuli; (ii) that host tissues are located in the narrow time window during which they are developmentally susceptible to gall induction; and (iii) that their sexual generation offspring can successfully induce galls on the novel host (Craig *et al.* 1993; Brown *et al.* 1996; Abrahamson & Weis 1997). Inability of western *A. kollari* to exploit *Q. suber* could result from failure in any one of these requirements. The significance of host genotype for gallwasps remains little understood. Studies of oak hybrid zones have shown that hostplant genotypes have major effects on the abundance of specific oak gallwasp species (Boecklen & Spellenberg 1990; Aguilar & Boecklen 1991), but whether the causes of such variation act before or after oviposition remains unknown. Host races have been demonstrated in at least two gallwasps, one on oaks (Abe 1988, 1991) and one on roses (Stille 1985b). In both systems, host races are characterized by demonstrable divergence in the cues which elicit oviposition. Rearing experiments by Folliot (1964, personal communication) and J. Pujade-Villar (personal communication) suggest that similar divergence has evolved between eastern and western *A. kollari*. They found that asexual female *A. kollari* from Brittany (almost certainly of the eastern race) oviposited on *Q. cerris*, but never *Q. suber*, while females from Catalonia (almost certainly of the western race) would lay eggs in *Q. suber* but never *Q. cerris*. The split between *Q. cerris* and *Q. suber* is relatively ancient (Manos *et al.* 1999), and the two species may possess quite different profiles of the host secondary metabolites often used by phytophagous insects as host location and oviposition cues (Thompson 1994; Abrahamson & Weis 1997). As for *A. mukaigawae* (Abe 1988, 1991), asexual generation *A. kollari* adapted to one of these hosts may not be able to recognize the alternative host as an oviposition site. Although it has not yet been investigated for *A. kollari*, it is possible that even were oviposition on the alternative host to occur, gall induction might not be successful.

Hybridization between eastern and western races of A. kollari

Although our data show the two lifecycle races of *A. kollari* to meet at a relatively clearly defined boundary in south-western France, we found no individuals in this region combining alleles characteristic of eastern and western distributions. The absence of hybrid heterozygotes despite the high frequency and diversity of refuge-specific private alleles suggests that hybrids are either genuinely rare, or extremely local, or both. Because both eastern and western races share some alleles, allozyme data will underestimate the frequency of hybridization; analysis of haplotype distributions is a far more powerful tool, and is still at an early stage in this system. The identification of two individuals in north-western France with Iberia-specific alleles and mitochondrial haplotypes characteristic of the eastern distribution shows that there is no absolute barrier to introgression. Several aspects of the biology of oak gallwasps, however, may contribute to partial prezygotic barriers.

First, sexual generations of *Andricus* and of other gall-inducers mate as the sexual females emerge from their galls (G. Stone, unpublished data; Stille 1985b; Waring *et al.* 1990). Because the sexual generations of the two races of *A. kollari* develop on different host plants, mating is thus likely to be highly assortative. Host-mediated assortative mating is well-known in phytophagous insects, and among gall-inducers has been demonstrated to contribute to genetic isolation of host races in a japanese oak gallwasp *A. mukaigawae* (Abe 1991), and in the the tephritid gall fly *Eurosta solidaginis* (Abrahamson & Weis 1997). To overcome such assortment, either: (i) asexual generation females must induce galls on the unfamiliar sexual generation host, and their sexual offspring mate with members of the other race; or (ii) sexual individuals must disperse from *Q. cerris* to *Q. suber*, or vice versa, and mate. As discussed above, asexual generation females may be unable to exploit nonfamiliar sexual generation hosts, and dispersal of sexual adults seems a more likely source of hybrids. Although small cynipids can be carried for long distances in aerial currents (Hardy & Cheng 1986; Ros-Farré & Pujade-Villar 1998), any dispersing sexual adults would still have to recognize and land on the nonfamiliar oak host, and locate a mate.

Second, the sexual generation wasps are very short-lived (less than a week for *A. quercuscalicis*) and a small difference in the phenology of sexual generation galls on their respective hosts would effectively prevent interbreeding (allochronic isolation). Phenological differences are known to be important in the restriction of genetic exchange between races of other herbivorous insects (e.g. Mitter *et al.* 1979; Feder *et al.* 1993), including an oak gallwasp (Abe 1991). The phenology of sexual generations of *A. kollari* on *Q. cerris* and *Q. suber* remains unknown, but it is quite possible that they do not overlap.

Given that hybridization occurs despite such considerations, a hybrid zone between the host races of *A. kollari* may exist in south-western France. The resulting asexual generation females are the first to experience any detrimental effects of interracial matings. Unlike the sexual generation females, which gall the same section Quercus hosts regardless of their refugial origin, the asexual generation females must return either to *Q. suber* or to *Q cerris* to lay their eggs. Any host-specific differences associated with host location or oviposition (as suggested by the experiments described above for *A. kollari*) may be disrupted by hybridization. Even if oviposition does occur, studies in another galling system showed that hybrid eggs either fail to induce galls, or that resulting larvae die at an early stage of development (Abrahamson & Weis 1997). If there is similarly strong negative selection against hybrid *A. kollari*, only first generation hybrids are expected. Because hybrids are probably generated only when sexual adults disperse between the two sexual generation hosts, any hybrid zone is expected to track closely those areas in which *Q. suber* and *Q. cerris* come into close contact and to have a breadth determined by the dispersal distance of sexual generation females. South-western France is a hotspot of hybrid zones in a range of taxa (Taberlet *et al.* 1998; Hewitt 1999). If it exists, this *A. kollari* hybrid zone would be unusual in that its position can be explained by the postglacial fortunes of the two sexual generation hosts.

Q. suber and *Q. cerris* are naturally sympatric in Italy (Fig. 1), and though the gallwasp fauna is generally less well-studied in Italy than it is in either Iberia or the Balkans, *A. kollari* is present in the area of overlap (Trotter & Cecconi 1904; Dalla Torre & Kieffer 1910). This raises the question of whether the sexual generation of *A. kollari* might also be found on *Q. suber* in this region. The limited data we have on Italian *A. kollari* link populations from the north of this refuge with the Balkans, but the lifecycle and genetic make-up of populations in the south remains completely unknown. *Q. suber* populations in North Africa and Italy are thought to have been derived by range expansion from an Iberian refuge one or more interglacial periods ago (Reille *et al.* 1996; Toumi & Lumaret 1998). If the eastern and western lifecycles of *A. kollari* had already diverged by this time, it is possible that western race migrants accompanied *Q. suber* across Africa and into Italy. Here they would have encountered eastern race populations already exploiting *Q. cerris*. This raises the possibility of a second, more ancient, contact zone between two host-races of *A. kollari* in Italy.

Acknowledgements

Graham Stone was supported by grants from the NERC EDGE thematic programme (GR9/03553), the Royal Society (574006), the British Ecological Society (Small Ecological Project Grant number 1462) and The E.P. Abraham Research Fund of Oxford University. José-Luis Nieves-Aldrey was supported by a grant from the DGES (PGC, PB97-1241). We would like to thank all of those who helped at various times in the collection of samples, particularly Olivier Plantard, Alex Rowe and an army of Hungarian Foresters. We thank Susan Haley and Gordon Brown for their help in the allozyme electrophoresis, Jill Lovell for her help with sequencing, and Sylvain Piry for his patient, long-range help with NJBP.exe. We thank Roger Folliot and Juli Pujade-Villar for their many comments on the biology of Spanish *Andricus kollari*. We thank Karsten Schönrogge, Paul Sunnucks and an anonymous referee for their constructive comments on the submitted manuscript.

References

Abe Y (1986) Taxonomic status of the *Andricus mukaigawae* complex and its speciation with geographic parthenogenesis. *Applied Entomology and Zoology*, **21**, 436–447.

Abe Y (1988) Two host races in *Andricus mukaigawae* (Mukaigawa) (Hymenoptera: Cynipidae). *Applied Entomology and Zoology*, **23**, 382–387.

Abe Y (1991) Host race formation in the gall wasp *Andricus mukaigawae*. *Entomologia Experimentalis et Applicata*, **58**, 15–20.

Abrahamson WG, Weis AE (1997) *Evolutionary Ecology Across Three Trophic Levels: Goldenrods, Gallmakers and Natural Enemies*. Princeton Monographs in Population Biology Number 29. Princeton University Press, Princeton, USA.

Aguilar JM, Boecklen WJ (1991) Patterns of herbivory in the *Quercus grisea* × *Quercus gambelii* species complex. *Oikos*, **64**, 498–504.

Ambrus B (1974) Cynipida-gubacsok-cecidia cyniparum. In: *Fauna Hungariae*, Vol. 12 *(Hymenoptera 2), Part 1/a (Serial Number 116)*. Academic Press, Budapest, Hungary.

Askew RR (1984) The biology of gallwasps. In: *The Biology of Galling Insects* (ed. Ananthakrishnan TN), pp. 223–271. Oxford and IBH Publishing Co, New Delhi.

Atkinson R (2000) The genetic analysis of natural history, reproductive strategy and population structure in European oak gallwasps (Hymenoptera: Cynipidae). PhD Thesis. University of Edinburgh, UK.

Bailey SF, Stange LA (1966) The twig wasp of Cork Oak — Its biology and control. *Journal of Economic Entomology*, **59**, 663–668.

Beijerinck MW (1902) Über die sexuelle Generation von *Cynips kollari*. *Marcellia*, **1**, 13–20.

Belkhir K (1999) *GENETIX, Version 4.0. A windows programme for population genetic analysis*. Laboratoire Génome, Populations, interactions UPR 9060 du CNRS, Université Montpellier 2, Place E. Bataillon, 34095 Montpellier Cedex 05, France.

Bensasson D, Zhang DX, Hewitt GM (2000) Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. *Molecular Biology and Evolution*, **17**, 406–415.

Bernatchez L, Wilson CC (1998) Comparative phylogeography of nearctic and palearctic fishes. *Molecular Ecology*, **7**, 431–452.

Boecklen WJ, Spellenberg R (1990) Structure of herbivore communities in two oak (*Quercus* spp.) hybrid zones. *Oecologia*, **85**, 92–100.

Brower AVZ (1994) Rapid morphological radiation and convergence among races of the butterfly *Heliconius erato* inferred from patterns of mitochondrial DNA evolution. *Proceedings of the National Academy of Sciences of the USA*, **91**, 6491–6495.

Brown JM, Abrahamson WG, Way PA (1996) Mitochondrial DNA phylogeography of host races of the goldenrod ball gallmaker, *Eurosta solidaginis* (Diptera: Tephritidae). *Evolution*, **50**, 777–786.

Camus A (1938) *Les Chênes*, Vol. 3. Monographie du genre *Quercus*. Lechevallier, Paris.

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. *American Journal of Human Geneticsi*, **19**, 233–257.

Cook JM, Stone GN, Rowe JA (1998) Patterns in the evolution of gall structure and life cycles in oak gall wasps (Hymenoptera: Cynipidae). In: *The Biology of Gall-Inducing Arthropods* (eds Csóka G, Mattson W, Stone GN, Price PW), pp. 261–279. General Technical Report NC 199. North Central Research Station USDA, St. Paul, Minnesota.

Craig TP, Itami JK, Horner JD, Abrahamson WG (1993) behavioural evidence for host-race formation in *Eurosta solidaginis*. *Evolution*, **47**, 1696–1710.

Csóka G (1997) *Plant Galls*. Agroinform Publishing, Budapest, Hungary.

Csóka G, Stone GN, Atkinson RJ, Schönrogge K (1998) The population genetics of postglacial invasions of northern Europe by cynipid gallwasps (Hymenoptera: Cynipidae). In: *The Biology of Gall-Inducing Arthropods* (eds Csóka G, Mattson W, Stone GN, Price PW), pp. 280–294. General Technical Report NC 199. North Central Research Station USDA, St. Paul, Minnesota.

Dalla Torre KW, Kieffer JJ (1910) Cynipidae. In: *Das Tierreich. Eine Zusammenstellung und Kennzeichnung der Rezenten Tierformen*, pp. 1–891. Der Deutsche Zoologischen Geselschaft, Berlin.

Docters van Leuwen WM (1959) Generatiewisseling en wisseling van waardplant bij vier Nederlandse galwespen. *De Levende Natuur*, **2**, 149–161.

Dumolin-Lapègue S, Demesure B, Fineschi S, Le Corre V, Petit RJ (1997) Phylogeographic structure of White oaks throughout the European continent. *Genetics*, **146**, 1475–1487.

Feder JL, Hunt TA, Bush L (1993) The effects of climate, host plant phenology and host fidelity on the genetics of apple and hawthorn infesting races of *Rhagoletis pomonella*. *Entomologia Experimentalis et Applicata*, **69**, 117–135.

Felsenstein J (1993) PHYLIP *(Phylogenetic inference package), Version 3.5c*. Department of Genetics, University of Washington, Seattle, USA.

Ferris C, Oliver RP, Davy AJ, Hewitt GM (1993) Native oak chloroplasts reveal an ancient divide across Europe. *Molecular Ecology*, **2**, 337–344.

Ferris C, Oliver RP, Davy AJ, Hewitt GM (1995) Using chloroplast DNA to trace postglacial migration routes of oaks into Britain. *Molecular Ecology*, **4**, 731–738.

Folliot R (1964) *Contribution à l'étude de la biologie des cynipides gallicoles (Hymenoptera, Cynipoidea)*. PhD Thesis. University of Rennes, France.

Garnier-Géré P, Dillman C (1992) A computer program for testing pairwise linkage disequilibria in subdivided populations. *Journal of Heredity*, **83**, 239.

Gómez A, Carvalho GR (2000) Sex, parthenogenesis and genetic structure of rotifers: microsatellite analysis of contemporary and resting egg bank populations. *Molecular Ecology*, **9**, 203–214.

Goudet J (1995) FSTAT, Version 1.2: a computer program to calculate *F*-statistics. *Journal of Heredity*, **86**, 485–486.

Guo SW, Thompson EA (1992) Performing the exact test of Hardy–Weinberg proportions for multiple alleles. *Biometrics*, **48**, 361–372.

Hails RS, Crawley MJ (1991) The population dynamics of an alien insect: *Andricus quercuscalicis* (Hymenoptera: Cynipidae). *Journal of Animal Ecology*, **60**, 545–562.

Hardy AC, Cheng L (1986) Studies in the distribution of insects by aerial currents.3. Insect drift over the sea. *Ecological Entomology*, **1**, 283–290.

Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174.

Hewitt GM (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society*, **58**, 247–276.

Hewitt GM (1999) Postglacial re-colonisation of European biota. *Biological Journal of the Linnean Society*, **68**, 87–112.

Houard C (1912) Les zoocécidies do nord de l'Afrique. *Annales de la Societé Entomologique de France*, **81**, 1–236.

Huelsenbeck JP, Hillis DM (1993) Success of phylogenetic methods in the 4 taxon case. *Systematic Biology*, **42**, 247–264.

Huntley BH (1990) European vegetation history: palaeovegetation maps from pollen data — 13 000 years BP to present. *Journal of Quaternary Science*, **5**, 103–122.

Jalas J, Suominen J (1987) *Atlas Florae Europeae. Distribution of Vascular Plants in Europe*, Vol. 2. Cambridge University Press, Cambridge.

Jermiin LS, Crozier RH (1994) The cytochrome-*b* region in the mitochondrial DNA of the ant *Tetraponera rufoniger* — sequence divergence in Hymenoptera may be associated with nucleotide content. *Journal of Molecular Evolution*, **38**, 282–294.

Manos PS, Doyle JJ, Nixon KC (1999) Phylogeny, biogeography and processes of molecular differentiation in *Quercus* subgenus *Quercus* (Fagaceae). *Molecular Phylogenetics and Evolution*, **12**, 333–349.

Marsden-Jones EM (1953) A study of the lifecycle of *Adleria kollari* Hartig, the marble or Devonshire gall. *Transactions of the Royal Entomological Society of London*, **104**, 195–222.

Mitter C, Futuyma DJ, Schneider JC, Hare DJ (1979) Genetic variation and host-plant relations in a parthenogenetic moth. *Evolution*, **33**, 777–790.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89**, 583–590.

Nei M, Maruyama T, Chakraborty R (1975) The bottleneck effect and genetic variability in populations. *Evolution*, **29**, 1–10.

Nichols RA, Hewitt GM (1994) The genetic consequences of long distance dispersal during colonisation. *Heredity*, **72**, 312–317.

Nieves-Aldrey JL (1982) Contribución al conocimiento de los cinípidos gallícolas (Hym. Cynipidae) de la encina y el alcornoque en la provincia de Salamanca. *Boletín de la Asociación Española de Entomología*, **5**, 59–74.

Nieves-Aldrey JL (1987) Estado actual del conocimiento de la subfamilia Cynipidae (Hymenoptera, Parasitica, Cynipidae) en la Península Ibérica. *Eos*, **63**, 179–195.

Nieves-Aldrey JL (1992) Revisíon de las especies europeas del género *Callirhytis* (Hymenoptera, Cynipidae). *Graellsia*, **48**, 171–183.

Peakall R, Beattie AJ (1991) The genetic consequences of worker ant pollination in a self-compatible clonal orchid. *Evolution*, **45**, 1837–1848.

Petit RJ, Pineau E, Demesure B, Bacilieri B, Ducousso A, Kremer A (1997) Chloroplast DNA footprints of postglacial recolonisation by oaks. *Proceedings of the National Academy of Sciences of the USA*, **94**, 9996–10001.

Plantard O, Rasplus J-Y, Mondor G, Le Clainche I, Solignac M (1998) Wolbachia-induced thelytoky in the rose gallwasp *Diplolepis spinosissimae* (Giraud) (Hymenoptera: Cynipidae), and its consequences on the genetic structure of its host. *Proceedings of the Royal Society of London B*, **265**, 1075–1080.

Plantard O, Rasplus J-Y, Mondor G, Le Clainche I, Solignac M (1999) Distribution and phylogeny of *Wolbachia*-induced thelytoky in Rhoditini and 'Aylacini' (Hymenoptera: Cynipidae). *Insect Molecular Biology*, **8**, 185–191.

Posada D, Crandall KA (1998) Modeltest: testing the models of DNA substitution. *Bioinformatics*, **14**, 817–818.

Pujade-Villar J (1991) *Contribució al coneixement dels cinípids cecidógens dels arbres i arbursts de Catalunya, dels cinípids associats a aquests i dels seus parásits*. Unpublished PhD Thesis. University of Barcelona, Spain.

Pujade-Villar J (1992) *Andricus kollari* (Hartig) (Insecta: Hymenoptera: Cynipidae) Part II: Consideracions sobre el seu cicle biológic. *La Sitja Del Llop*, **3**, 12.

Raymond M, Rousset F (1995) GENEPOP (Version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Reille M, Andrieu V, de Beaulieu J-L (1996) Les grands traits de l'histoire de la végétation des montagnes méditerranéennes occidentales. *Ecologie*, **27**, 153–169.

Richardson BJ, Baverstock PR, Adams M (1986) *Allozyme Electrophoresis; a Handbook for Animal Systematics and Population Studies*. Academic Press, New York.

Ros-Farré P, Pujade-Villar J (1998) Estudio mediante una trampa malaise de la comunidad de cinipidos cecidogenos e inquilinos de Santa Coloma, Andorra (Hymneoptera, Cynipidae). *Ecología*, **12**, 441–454.

Saitou N, Nei M (1987) The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 1406–1425.

Schönrogge K, Walker P, Crawley MJ (1998) Invaders on the move: parasitism in the galls of four alien gallwasps in Britain (Hymenoptera: Cynipidae). *Proceedings of the Royal Society of London, B*, **256**, 1643–1650.

Simon J-C, Baumann S, Sunnucks P *et al.* (1999) Reproductive mode and population genetic structure of the ceral aphid *Sitobion avenae* studied using phenotypic and microsatellite markers. *Molecular Ecology*, **8**, 531–545.

Stille B (1985a) Population genetics of the parthenogenetic gall wasp *Diplolepis rosae* (Hymenoptera: Cynipidae). *Genetica*, **7**, 145–151.

Stille B (1985b) Host plant specificity and allozyme variation in the parthenogenetic gallwasp *Diplolepis mayri* and its relatedness to *D. rosae* (Hymenoptera: Cynipidae). *Entomologia Generalis*, **10**, 87–96.

Stone GN, Cook JM (1998) The structure of cynipid oak galls: patterns in the evolution of an extended phenotype. *Proceedings of the Royal Society of London, B*, **265**, 979–988.

Stone GN, Schönrogge K, Crawley MJ, Fraser S (1995) Geographic variation in the parasitoid community associated with an invading gallwasp, *Andricus quercuscalicis* (Hymenoptera: Cynipidae). *Oecologia*, **104**, 207–217.

Stone GN, Sunnucks PJ (1993) The population genetics of an invasion through a patchy environment: the cynipid gallwasp *Andricus quercuscalicis*. *Molecular Ecology*, **2**, 251–268.

Sunnucks PJ, de barro PJ, Lushai G, MacLean N, Hales D (1997) Genetic structure of an aphid using microsatellites: cyclic parthenogenesis, differentiated lineages, and host specialisation. *Molecular Ecology*, **7**, 339–353.

Sunnucks PJ, Hales DF (1996) Numerous transposed sequences of mitochondrial cytochrome oxidase I–II in aphids of the genus *Sitobion* (Hemiptera: Aphididae). *Molecular Biology and Evolution*, **13**, 510–524.

Sunnucks PJ, Stone GN (1996) Genetic structure of invading insects and the case of the knopper gallwasp. In: *Frontiers of Population Ecology* (eds Floyd RB, Sheppard AW, De Barro PJ), pp. 485–496. CSIRO, Canberra, Australia.

Suomalainen E, Saura A, Lokki J (1987) *Cytology and Evolution in Parthenogenesis*. CRC Press, Boca Raton, Florida, USA.

Swofford DL (1998) PAUP* 4.0 — Phylogenetic Analysis Using Parsimony (*and other methods). Sinauer Associates, Sunderland, MA, USA.

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: *Molecular Systematics* (eds Hillis DM, Moritz C, Mable BK), pp. 407–514. Sinauer, Sunderland, Massachusetts, USA.

Swofford DL, Selander RB (1981) BIOSYS 1: a FORTRAN program for the comprehensive analysis of data in population genetics and systematics. *Journal of Heredity*, **72**, 281–283.

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

Thompson JN (1994) *The Coevolutionary Process*. University of Chicago Press, Chicago USA.

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

Toumi L, Lumaret R (1998) Allozyme variation in cork oak (*Quercus suber* L.): the role of phylogeography and genetic introgression by other Mediterranean oak species and human activities. *Theoretical and Applied Genetics*, **97**, 647–656.

Trotter A, Cecconi G (1904) Cecidotheca italica. *Marcellia*, **3**, 76–81.

Waring GL, Abrahamson WG, Howard DJ (1990) Genetic differentiation among host-associated populations of the gallmaker *Eurosta solidaginis* (Diptera: Tephritidae). *Evolution*, **44**, 1648–1655.

Weir BS (1991) *Genetic Data Analysis*. Sinauer, Sunderland, Massachussetts, USA.

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Yaltirik F (1984) *Türkiye Meseleri Teshis kilavuzu (The oak trees of Turkey)*. Yenilik Basimevi, Istanbul.

Yang Z (1996) Among-site variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution*, **11**, 367–372.

Yasumatsu K (1951) A new Dryocosmus injurious to chestnut trees in Japan (Hym. Cynipidae). *Mushi*, **22**, 89–93.

Zhang DX, Hewitt GM (1996) Highly conserved nuclear copies of the mitochondrial control region in the desert locust *Schistocerca gregaria*: some implications for population studies. *Molecular Ecology*, **5**, 295–300.

Graham Stone has a long-standing interest in the phylogeography and evolution of oak cynipids, and in particular in the consequences of human dispersal of Turkey oak for the distributions and population genetic structure of host-alternating species. Rachel Atkinson and Antonis Rokas are both graduate students in the gallwasp group in Edinburgh. Rachel is working on interspecific comparisons of population structure in oak gallwasps, ranging from broad studies of phylogeography to population genetic analyses of the number of females laying eggs in a single oak bud. Antonis Rokas is working on a sequence-based phylogeny of oak cynipids, and has a special interest in the phylogenetic distribution of *Wolbachia* symbionts in gallwasps. Gyuri Csóka has a lifelong interest in all aspects of the biology of gallwasps, particularly in patterns of host oak exploitation. José-Luis Nieves-Aldrey has published many taxonomic papers on gallwasps and their associated communities, and has just completed a major monograph on this group in the Iberian Peninsular.

Supplementary material

The following material is available from http://www.blackwell-science.com/products/journals/suppmat/mec/mec1211/mec1211sm.htm

Appendix I

Sampling effort and allele frequencies at the 13 polymorphic loci for populations of *Andricus kollari*. Degrees latitude and longitude are decimalized

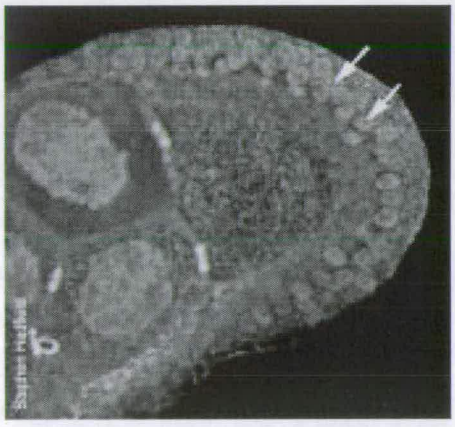
Journal Club

Selfish element wars in a wasp battleground

Bacteria of the genus *Wolbachia* are masters in the art of sex-determination manipulation. By exploiting the haplodiploid system of sex determination (in which haploid eggs develop into males and diploid eggs develop into females) in certain Hymenoptera, *Wolbachia* are able to convert unfertilized eggs, which would normally develop as males, into females. As a consequence, a sexually reproducing wasp population might, on infection with *Wolbachia*, become parthenogenetic. Furthermore, it is a well-established fact that the life style of many parasitoid wasp species favours a female-biased sex ratio. Therefore, one would expect in *Trichogramma kaykai*, a minute wasp that parasitizes the eggs of a lycaenid butterfly and which has a typical sex ratio consisting of 22–29% males, that infection with *Wolbachia* would mean just one thing: switch to parthenogenesis. Alas, this is not the case, with *Wolbachia* infection levels being stably maintained in frequencies of between 5% and 25%. In a recent paper, Stouthamer *et al.*[1] exploited this system for some elegant experimental and theoretical work.

Three hypotheses could explain such infection levels; inefficient transmission of *Wolbachia*; presence of a suppressor gene that either kills *Wolbachia* or balances its sex-ratio-distorting effect; or another non-Mendelian balancing factor acting in the opposite direction from *Wolbachia* (i.e. favouring a male-biased sex ratio). Plotting *Wolbachia* transmission frequencies versus a range of infection costs is enough to show that, for the range of infection frequencies observed in field populations of *T. kaykai*, the two possible outcomes are either the fixation or loss of the *Wolbachia* infection (rather than a maintenance of a stable equilibrium of about 5–25%). Adding to this, the recent demonstration (by the same research group) that *Wolbachia* might be horizontally transmitted at an appreciable frequency (37%) and inefficient transmission of *Wolbachia* as the sole explanation for the low levels of infection becomes unlikely.

Alternatively, low infection levels of *Wolbachia* could be explained by the presence of suppressor genes that could either kill the bacteria or eliminate their effect on the host. Modelling carried out by Stouthamer *et al.*, which takes into account the peculiarities of the *T. kaykai* system, shows that the presence of a suppressor gene would be a plausible theoretical outcome; however, detailed crossing protocols failed to demonstrate its presence.



A third scenario invokes the presence of a sex-ratio factor acting in the opposite direction from *Wolbachia*, favouring male production. Such elements are collectively called paternal sex ratio (PSR) factors and have been identified in previous studies on parasitic wasps. PSRs are paternally transmitted and eggs fertilized with sperm from PSR-bearing males become male instead of female (exactly the opposite effect of that of *Wolbachia*). By crossing experiments and chromosome karyotypes, Stouthamer *et al.* present compelling evidence for the major effect of a PSR factor keeping *Wolbachia* infection levels low.

This study is an elegant example of the occurrence of genomic conflict in a natural population. As our knowledge of the characterization and capabilities of these amazing reproductive parasites increases, ever more exciting cases of genomic parasitism will appear.

1 Stouthamer, R. *et al.* (2001) Selfish element maintains sex in natural populations of a parasitoid wasp. *Proc. R. Soc. London B Biol. Sci.* 268, 617–622

Antonis Rokas
a.rokas@ed.ac.uk

Behaviour without function: egg rejection in the absence of parasitism

Avian brood parasites and their hosts provide one of the classical examples of coevolution. One famous aspect of this is the mimicry of host egg size, shape and coloration by some cuckoos. This parasitic tactic is seen as a response to hosts that have evolved the ability to detect and reject odd-looking eggs in their nest. Surprisingly, some passerines that are not currently challenged by brood parasites still show this rejection behaviour when experimenters artificially introduce alien eggs into their nests. This rejection can reach levels of near 100% even when the added eggs are very similar to their own.

Stephen Rothstein[1] has now studied two of these populations in detail. He demonstrates that the most plausible explanation for this ability in the hosts is that the ancestors of these birds were subject to natural parasitism, rejection then developed as a response to this, and the trait has been retained despite the parasitism fading out.

In support of this theory, Rothstein was unable to find a cost of this rejection behaviour in either population. It is generally supposed that such behaviour was unlikely to be retained in the absence of parasitism, because it would cause hosts to occasionally reject one of their own eggs after misidentifying it as alien. Rothstein's contention is that this might not be the case, and the trait can be retained because, in the absence of parasitism, it is selectively neutral. He suggests that there is no clear

http://tree.trends.com 0169-5347/01/$ – see front matter

Of course, a similar question can be approached from the evolutionary end of biology. How, for example, can we use the selective pressures generated from anthelmintic treatment to generate less pathogenic worms? In a characteristically stimulating paper, Read and his colleagues examine the prepatent period of worms. This is the all-important time between infection and worm maturation that determines how large a worm can grow and the tradeoff between the fecundity benefits of size against the risk of dying. Although drug application affects worm burdens, it also affects age-specific schedules that in turn can influence fitness and could lead to an important impact on worm selection. This is a different approach from that used by Cornell and Grenfell, but one that provides an alternative view and an intriguing starting point that should stimulate other workers to take these ideas further.

The other chapters are peppered with some interesting questions and intriguing approaches to parasitology, some of which I will use to liven up sections of my teaching. I must also make sure that some of my postgraduate students read these papers, which give some fresh ways of looking at parasitological issues. We have not reached the synthesis we seek, but this book and its focus on the enigmatic term R_0 certainly helps us down that route.

Peter Hudson
Institute of Biological Sciences, University of Stirling, Stirling, UK FK9 4LA.
e-mail: p.j.hudson@stir.ac.uk

What is a character?

The Character Concept in Evolutionary Biology
edited by Günter P. Wagner. Academic Press, 2000. £54.00 hbk (xxiii + 622 pages)
ISBN 0 12 730055 4

Günter Wagner, in the edited volume *The Character Concept in Evolutionary Biology*, has managed not only to raise perhaps one of the most interesting biological questions, but also to assemble an array of superb scientists to attempt to answer it. The subject matter of the book is succinctly laid out in the Introduction by Lewontin: 'How are we to recognise the *true* characters of organisms rather than imposing upon them arbitrary divisions that obscure the very processes that we seek to understand? That is the question of this book. No issue is of greater importance in the study of biology'. Few would disagree.

Given that certain concepts in evolutionary biology have had a long and frustrating history (e.g. species concepts), some researchers might be negatively preoccupied with attempts to address broad questions with little understanding of the underlying mechanisms. But, from the first chapter, Wagner sets out his objectives: the purpose of theories is to induce research that makes them obsolete (and the history of species concepts justifies such an assertion).

The character problem can be seen as a special case of a problem permeating all sciences, namely the question of which units, if any, are the fundamental ones that play a causal role in a process (the natural kinds). In the literature of evolutionary biology, the nature of characters has received extensive treatment under investigations into another famous 'what is' question, that of homology (e.g. Ref. 1). Wagner suggests that an expansion of the homology question 'how do I identify the correspondence between the characters in two species (if any)?' to the broader 'what are the natural units organisms are composed of and what is their origin and evolution?' offers greater scope to the investigation. This is no small step, but I agree with Wagner that it should (at least sometimes) be taken.

The volume is divided into five sections, the first of which attempts to identify the roots of the character concept and the second offers a range of new approaches to and/or definitions of the issue. The remaining three sections include attempts to develop protocols for identification of characters; to understand their architecture as shaped by molecular, developmental and environmental factors; and to elucidate their evolutionary origins. All sections are well written although, depending on the individual reader's background, some will be more readily digestible than others.

But practicing evolutionary biologists might, perhaps justifiably, think that the questions addressed in this book are too philosophical. In other words, why do practicing biologists need a character concept (and will they suffer from the lack of one)? But the importance of understanding the origin and evolution of characters becomes obvious. Whole areas of evolutionary biology (e.g. comparative biology, phylogenetics and developmental evolution) are fundamentally dependent on the correct identification of characters. Take, for example, morphological phylogenetics. The ability to correctly identify characters that are independent and homologous (rather than homoplasious) can profoundly bias the result of the analysis[2] and cases in which morphological characters are functionally or developmentally correlated are far from rare.

Comparative developmental biologists are also interested in understanding exactly what a character is. A good example is offered by the complex eyespot patterns seen on the wings of butterflies (Chapter 15). In this case, the term 'character' can be applied to a single feature of an eyespot, a single eyespot, a subset of eyespots or the complete pattern. Years of comparative genetical, developmental and morphological work have shown that the most appropriate use of the term 'character' is for the complete eyespot pattern. This issue is not trivial semantics. An explanation, at the genotypic and phenotypic level, of morphological patterns – such as the butterfly eyespots – is crucial for a research programme into how characters originate and evolve. It also facilitates a better description of the transition from genotype to phenotype.

Of what use is this book going to be in the ever-growing body of literature that concerns evolutionary biology? I would not recommend it to anybody who expects to find rigid answers to the questions asked or who is not willing to, more than occasionally, tackle compact and dense paragraphs. But for those who can afford to spend some time and thought on some very interesting problems, or for graduate students in need of stimulating questions, this book is a must.

Antonis Rokas
Institute of Cell, Animal and Population Biology, Ashworth Laboratories, University of Edinburgh, West Mains Road, King's Buildings, Edinburgh, UK EH9 3JT.
e-mail: a.rokas@ed.ac.uk

References
1 (1999) *Homology*, John Wiley & Sons
2 Poe, S. and Wiens, J.J. (2000). Character selection and the methodology of morphological phylogenetics. In *Phylogenetic Analysis of Morphological Data* (Wiens, J.J., ed.), pp. 20–36, Smithsonian Institution Press

Heredity 87 (2001) 294–304 *Received 3 November 2000, accepted 1 February 2001*

Understanding patterns of genetic diversity in the oak gallwasp *Biorhiza pallida*: demographic history or a *Wolbachia* selective sweep?

ANTONIS ROKAS*, RACHEL J. ATKINSON, GORDON S. BROWN, STUART A. WEST & GRAHAM N. STONE

Institute of Cell, Animal and Population Biology, Ashworth Laboratories, University of Edinburgh, King's Buildings, West Mains Road EH9 3JT, U.K.

The endosymbiont *Wolbachia* can be responsible for selective sweeps on mitochondrial DNA variability within species. Similar signals can also result from demographic processes, although crucially the latter affect nuclear as well as mitochondrial loci. Here we present data on *Wolbachia* infection status and phylogeographic patterning for a widely distributed insect host, the oak gallwasp *Biorhiza pallida* (Hymenoptera: Cynipidae). Two hundred and eighteen females from eight European countries were screened for *Wolbachia*. All individuals from Hungary, Italy, France, U.K., Ireland, Switzerland, Sweden, and northern and southern Spain were infected with a single group A strain of *Wolbachia*, while populations in central Spain were not infected. A mitochondrial marker (cytochrome *b*) shows low variation and departure from neutrality in infected populations, but greater variation and no deviation from neutrality in *Wolbachia*-free populations. This pattern is compatible with a *Wolbachia*-induced selective sweep. However, we also find parallel differences between infected and uninfected populations for nuclear markers (sequence data for ITS1 and ITS2). All markers support the existence of a deep split between populations in Spain (some free of *Wolbachia*), and those in the rest of Europe (all infected). Allelic variation for five allozyme loci is also consistent with the Spain–rest of Europe split. Concordant patterns for nuclear and mitochondrial markers suggest that differences in the nature and extent of genetic diversity between these two regions are best explained by differing demographic histories (perhaps associated with range expansion from Pleistocene glacial refugia), rather than a *Wolbachia*-associated selective sweep.

Keywords: cytochrome *b*, gallwasp, phylogeography, range expansion, selective sweep, *Wolbachia*.

Introduction

Bacteria of the genus *Wolbachia* are intracellular microorganisms that infect the reproductive tissues of arthropods and nematodes (O'Neill *et al.*, 1997; Werren, 1997; Stouthamer *et al.*, 1999). They are inherited cytoplasmically (i.e. passed from mother to daughter) and alter reproduction in their arthropod hosts in a number of ways, including cytoplasmic incompatibility, male killing, feminization and imposition of parthenogenesis (see O'Neill *et al.*, 1997 and chapters therein). *Wolbachia* are extremely common, infecting 16–22% of insects (Werren *et al.*, 1995; West *et al.*, 1998; Werren & Windsor, 2000), with a recent study indicating that the percentage might be even higher (Jeyaprakash & Hoy, 2000). In addition to immediate reproductive modifications, *Wolbachia* infection has a range of longer term evolutionary impacts on host taxa (O'Neill *et al.*, 1997; Werren, 1997; Stouthamer *et al.*, 1999).

Most research on *Wolbachia*, particularly for insect hosts, has concentrated on understanding the phylogenetic distribution and extent of infection (Werren *et al.*, 1995; West *et al.*, 1998; Werren & Windsor, 2000), using a small number of individuals for each species. The few studies examining geographical variation in levels of *Wolbachia* infection of a single host have revealed spatial patterns in the presence/absence of the bacterium and the occurrence of multiple infections (Turelli *et al.*, 1992; Solignac *et al.*, 1994; Plantard *et al.*, 1998; Malloch *et al.* 2000). Studies that combine analyses of

*Correspondence. E-mail: a.rokas@ed.ac.uk

spatial variation in *Wolbachia* infection, and of genetic diversity in the host, are of particular relevance because of the potential impact of *Wolbachia* on the host's genetic structure (described below). In this paper, we address the potential impact of *Wolbachia* infection on large-scale genetic patterning in an insect host, the oak gallwasp *Biorhiza pallida*.

When an advantageous mutation is driven through a population to fixation (a process known as a selective sweep), much of the neutral variation at linked loci is eliminated during the process (Maynard Smith & Haigh, 1974). The genetic variants at the linked loci that are initially paired with the advantageous mutation, although neutral in themselves, 'hitchhike' to fixation. Spread of a specific *Wolbachia* strain through a host population can have a directly analogous effect on variation in other cytoplasmically inherited markers, such as mitochondrial DNA. Evidence from *Drosophila* suggests that a single mtDNA haplotype may become widespread in the host population through hitchhiking with a successful *Wolbachia* strain (Turelli *et al.*, 1992; Ballard *et al.*, 1996). Selective sweeps on mtDNA not only reduce haplotype diversity but also cause the remaining set of host haplotypes to deviate from predictions based on neutrality (Johnstone & Hurst, 1996).

These potential impacts of *Wolbachia* infection on mitochondrial markers are particularly important because the latter are often employed as phylogeographic markers under the assumption of neutral evolution (Johnstone & Hurst, 1996). Low variability in mtDNA, and departure from neutrality, can also be caused by demographic processes such as range expansions or population bottlenecks (Avise, 2000). Demographic and *Wolbachia*-related explanations for spatial patterning in mtDNA variation can be distinguished by comparing patterns seen for nuclear and mitochondrial markers. Unlike mitochondrial markers, nuclear markers are typically inherited in a Mendelian fashion and are not expected to show any change in nucleotide diversity or deviation from neutrality in response to *Wolbachia* infection, although this depends on the phenotypic effect that *Wolbachia* has on its host. More specifically, only if sexual reproduction is involved can this assumption hold. If the phenotypic effect of *Wolbachia* on its host is the induction of parthenogenesis (and assuming that genetic exchange with sexual relatives is rare), concordance between nuclear and mitochondrial markers of the host may be expected.

In contrast to a *Wolbachia*-induced selective sweep, demographic processes cause changes in the extent and nature of variation for both mitochondrial and nuclear markers, although mitochondrial markers are expected to show stronger responses due to their lower effective population size (Avise, 2000). This difference is the basis for demonstrating *Wolbachia*-associated selective sweeps using comparative studies of sequence variation in mitochondrial and nuclear markers (for example, Turelli *et al.*, 1992; Ballard *et al.*, 1996). For hosts infected with *Wolbachia*, concordance in spatial patterning of genetic diversity in nuclear and mitochondrial markers argues against a significant causative role for *Wolbachia* infection.

Undetected selective sweeps on mtDNA variation due to *Wolbachia* infection can thus generate at least two types of artefact if interpreted in a phylogeographic context. Firstly, loss of mtDNA diversity in part of the host species' range could be attributed to a demographic effect having a similar impact, such as a population bottleneck. Here, a process that in fact affects only cytoplasmically inherited markers could be mistaken for a demographic process affecting both mitochondrial and nuclear markers. Secondly, patterning in mtDNA variation generated by spatial patterning in *Wolbachia* infection could be interpreted mistakenly as indicative of the phylogeographic history of the host. These potential pitfalls are highly relevant to studies of insect phylogeography, because of the high proportion of insect species infected with *Wolbachia* (Werren *et al.*, 1995; West *et al.*, 1998; Jeyaprakash & Hoy, 2000; Werren & Windsor, 2000) and the widespread use of mitochondrial markers in phylogeographic reconstruction (Avise, 2000).

In this paper, we analyse spatial patterns of variation in *Wolbachia* infection, and mitochondrial and nuclear markers for a widespread European phytophagous insect, the oak-apple gallwasp *Biorhiza pallida* (Olivier 1791). *B. pallida* induces galls on oaks in the genus *Quercus* (Csóka, 1997). *B. pallida is* extremely widely distributed in the western Palaearctic, extending from Morocco in the west to Georgia in the east, and as far north as Sweden. *Wolbachia* infection was detected in central European populations of *B. pallida* as part of a broader phylogenctic survey of *Wolbachia* in oak gallwasps (Rokas, unpublished data). These two features of *B. pallida* make it a suitable taxon within which to examine large scale spatial patterns in *Wolbachia* infection, and any associated impacts on variation at both mitochondrial and nuclear markers.

Oak gallwasps are obligate parasites of oak trees, and their spatial patterns of genetic variation are expected to reflect to an extent the phylogeographic history of their oak hosts. During the Pleistocene (1.8 mya until 0.01 mya), cycles of glacials and interglacials led to repeated range contraction and expansion of many taxa across Europe, with many species surviving the glacials in refugial areas in the southern of Europe (Italy, the Balkans and Spain) (Huntley & Webb, 1989; Hewitt,

1999). The retreat of the last ice sheet across Europe at the end of the Pleistocene era was followed by northern range expansion by many organisms from one or more of these refugia, resulting in the distribution pattern we see today. A growing body of work on oak gallwasps, covering 12 species in the genera *Andricus* and *Cynips*, shows that their current distributions and geographical patterns of genetic variation are determined largely by two factors: (a) the number and location of regions that acted as glacial refugia, and (b) the extent to which alternative refugia have contributed colonists to post-glacial range expansion (Stone & Sunnucks, 1993; Sunnucks & Stone, 1996; Atkinson, 2000; Stone *et al.* 2001). Consistent features of all species studied include (a) the existence of refuge-specific allozyme alleles and mtDNA haplotypes, and (b) a decline in genetic diversity with increasing latitude and distance from refugia. Where both have been studied, patterns of genetic variation north of glacial refugia are similar for mitochondrial and nuclear markers (Stone *et al.* 2001), and are consistent with genetic subsampling of neutral variation associated with the range expansion process (Stone & Sunnucks, 1993; Sunnucks & Stone, 1996; Atkinson, 2000). Preliminary sampling of the species whose phylogeography has been studied also suggests that, unlike *B. pallida*, they are free of *Wolbachia* infection. Although precise phylogeographic scenarios vary among species, these studies indicate that in the absence of any impact of *Wolbachia* infection, similar concordance in spatial patterning in nuclear and mitochondrial markers represents a qualitative null expectation for *B. pallida*. In contrast, discordant patterning in nuclear and mitochondrial markers, and concordance in patterning of *Wolbachia* infection and mtDNA diversity would suggest a significant impact of *Wolbachia*.

We established spatial patterns in *Wolbachia* infection by using a polymerase chain reaction (PCR)-based screening technique for 218 *B. pallida* individuals sampled from 46 localities in eight countries across Europe (Fig. 1). Strain diversity of *Wolbachia* was assessed by sequencing a fragment of the *wsp* gene. Host genetic diversity was analysed using a mitochondrial sequence (a fragment of the cytochrome *b* gene) and two nuclear sequences (the internal transcriber regions ITS1 and ITS2). We extended the diversity of nuclear markers sampled by screening 270 individuals for five polymorphic allozyme loci. We use these data to address the following questions: (a) how many strains of *Wolbachia* are present in this host, and how many infection events have occurred?; (b) is infection with

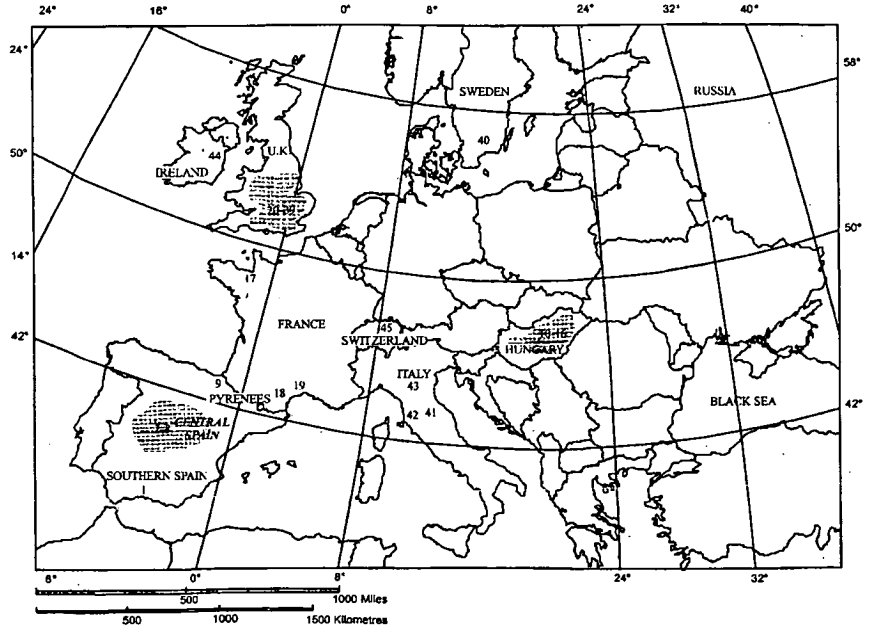


Fig. 1 Map of collection sites of populations of *Biorhiza pallida*. Spain: (1) Prado del Rey (2–8) Guadalix de la Sierra; Los Molinos; Zarzalejo; Avila-El Escorial; Cercedilla; Soto del Real; Villaviciosa. Northern Spain: (9) Puerto de Velate. Hungary: (10–16) Tiszaigar; Mátrafüred, Szentendre; Karcag; Bajna; Szeghalom; Visegrad. France: (17) Rennes; (18) St Jean pied de Porte; (19) Clermont-l'Hérault. U.K. (20–39) Cambridge; Hertford; Birnwood Forest; Oxford; Isle of Wight; Hampstead Heath; Fakenham; Broughton; SW Lincoln; Thetford; Elsfield; London; Ascot; Cawood; Chatham; West Shropshire; Grace Dieu Wood; Shorne; Hertingfordbury; Stoughton. Sweden: (40) Uppsala. Italy: (41) Chianti; (42) Volterra; (43) Casina. Ireland: (44) Dublin. Switzerland: (45) Luin. Populations in ***bold italics*** in central Spain (2–8) indicate absence of *Wolbachia* infection.

Wolbachia associated with lower sequence diversity, and departure from neutrality, for the mitochondrial marker?; (c) do the mitochondrial and nuclear markers show similar or discordant patterns of variation?; and (d) can we discriminate between demographic processes and a *Wolbachia*-associated selective sweep as possible causes of observed variation in host mitochondrial DNA?

Materials and methods

Collection and DNA extraction

Sexual generation galls of *B. pallida* were collected from 46 localities in eight European countries (Fig. 1, Table 1). *B. pallida* galls are multilocular (more than one offspring emerges from a single gall). To minimize screening of siblings we used one female from each gall, except where there was just a single gall from a particular location, when two females were screened. DNA was extracted from 218 female wasps as described by Stone & Cook (1998). To avoid contamination, each female wasp was soaked in 5% bleach and then serially rinsed in drops of sterile water prior to DNA extraction. With each DNA extraction three control extractions were performed using a *Nasonia Wolbachia*-positive strain, a *Nasonia Wolbachia*-negative strain and a no-DNA sample.

Wolbachia *screening*

Screening for *Wolbachia* was performed by PCR using *Wolbachia*-specific primers for the *ftsZ* cell-cycle gene (Werren *et al.*, 1995). These and all other PCRs were performed in a PTC-200 DNA engine (MJ Research Waltham, MA, USA). The forward primer was ftsZF1 (Werren *et al.*, 1995) and a new reverse primer WOLG-R was designed based on sequences available in GenBank. The sequence of WOLG-R (26 nucleotides) is 5′-GCA GVA TCA ACY TCA AAY ARA GTC AT-3′ (V = G/A/C, Y = C/T, *R* = A/G). The ftsZF1–WOLG-R pair amplifies the A, B and C groups of *Wolbachia*. Screening PCRs were attempted for sample DNA extractions at dilutions ranging from 1/10 to 1/100. Control PCRs were always performed. The PCR cycle for *ftsZ* was: one cycle of 94°C for 3 min, 55°C for 90 s and 72°C for 5 min, followed by 35 cycles of 94°C for 30 s, 55°C for 90 s, 72°C for 5 min and a final extension step at 72°C for 5 min. The PCRs were performed in 25 μL volumes, consisting of 1 μL DNA sample, 2.5 μL 10 × PARR buffer (Hybaid Ashford, UK), 1 μL $MgCl_2$ (25 mM), 0.5 μL dNTPs (10 mM), 0.35 μL of each primer (20 mM), 0.25 μL *Taq* (Promega Madison, WI, USA) and 19.05 μL of distilled, deionized H_2O. A 1% ethidium bromide-stained agarose gel was used for electrophoresis, and loaded with 15 μL of each reaction. To check that any *Wolbachia*-negative samples were not artefactual because of (a) failed DNA extraction; (b) presence of PCR inhibitors; or (c) incorrect DNA concentration, control PCRs with the general eukaryotic 28S rDNA primers 28Sf and 28Sr were performed as described in Werren *et al.* (1995). Of the 218 individuals screened for *Wolbachia* infection, 12 did not amplify for 28S rDNA and were discarded.

Table 1 Sampling regions of *Biorhiza pallida*. + *indicates* infection with *Wolbachia* – indicates absence of infection

Geographical region	Number of individuals tested	*Wolbachia*
Central Spain	34	–
Southern Spain	5	+
Northern Spain	5	+
Hungary	26	+
France	17	+
U.K.	95	+
Sweden	2	+
Italy	12	+
Ireland	5	+
Switzerland	5	+

PCR amplification and sequencing

All sequencing reactions were carried out at least twice (either with the forward and reverse primers for *wsp* and cytochrome *b* or twice with the forward primers for the ITS fragments) to minimize PCR artefacts, ambiguities and base-calling errors. Sequencing was carried out using the a sequencing kit and an automated sequencer (BigDye Terminator kit and ABI 377 sequencer; Perkin-Elmer Foster City, CA, USA). *Wolbachia* diversity in infected populations was assessed by PCR and sequencing of a fragment of the *wsp* gene (nine specimens were analysed: two from Hungary; one from southern Spain; two from France, southern and central; one from Switzerland; one from the U.K.; one from Ireland; and one from Italy). This is the most polymorphic gene so far isolated from *Wolbachia* (Zhou *et al.*, 1998) and hence, the most likely to distinguish between two closely related *Wolbachia* strains. *Wolbachia*-infected individuals were sequenced for *wsp* using the 81F and 691R primers following the methods described by Zhou *et al.* (1998). The total volume of three PCR reactions for each individual wasp was electrophoresed on a 1% agarose gel. The expected bands were cut from the gel and cleaned with a DNA extraction kit (QIAQuick gel extraction kit, no. 28704; Qiagen, Crawley, UK), and the clean DNA fragment was quantified and sequenced. Twenty-nine individuals (10 from central Spain; five

from southern Spain; three from the U.K.; three from southern France; four from Hungary; one from northern Spain; one from Switzerland; one from Ireland; and one from Italy) were sequenced for a 433-base pair (bp) fragment of cytochrome *b* showing 13 distinct haplotypes (Table 2). The cytochrome *b* fragment was amplified using the primers CB1 and CB2 as described by Stone & Cook (1998), and purified and sequenced as described above for *wsp*. Internal transcriber regions were amplified using the universal primers ITS4 and ITS5 (White *et al.*, 1990). The amplified fragment consisted of the internal transcriber regions, ITS1 and ITS2, and the 5.8S rDNA region of the rDNA array. Fifty-four clones of a 474-bp ITS2 fragment were sequenced for 13 individuals (three from central Spain, one from southern Spain, one from northern Spain, two from France, two from Italy, one from Hungary, one from Switzerland, one from Ireland and one from the U.K.) and 28 clones of a 635-bp fragment of ITS1 for six (of the 13) individuals (see below). The PCR cycle consisted of an initial denaturation step of 94°C for 2 min, followed by 35 cycles of 94°C for 15 s, 55°C for 60 s and 72°C for 2 min, and a final extension step of 15 min at 72°C. Reaction conditions were as for *Wolbachia* screening, except that 1.5 μL $MgCl_2$ and 18.55 μL of deionized, distilled H_2O were used. The rDNA array is present in multiple copies in the typical eukaryote genome and undergoes concerted evolution (Hillis & Dixon, 1991). However, for fast-evolving regions of the array, concerted evolution is not always perfect, resulting in intraindividual variation. To cheek for this variation, PCR products were cloned using a cloning kit (TOPO TA cloning kit no. 4500–01; Invitrogen Groningen, Netherlands) and 2–6 clones from each specimen were subsequently sequenced. Plasmids containing the fragment of interest were isolated using a commercial kit (QIAprep Spin miniprep kit, no. 27104; Qiagen). Plasmid DNA was subsequently quantified and sequenced.

Allozyme screening

Two hundred and seventy individuals were screened at five variable allozyme loci using cellulose–acetate gel electrophoresis (Zip Zone equipment; Helena Laboratories Gatshead, UK) as described in Stone & Sunnucks (1993) and Stone *et al.* (2001). Of an original set of 13 loci (see Stone *et al.* 2001) screened in *B. pallida*, five systems were found to be polymorphic: peptidase b (PEPb) (EC 3.4.11); aspartate aminotransferase (GOTm) (EC 2.6.1.1); 6-phosphogluconate dehydrogenase (6PGD) (EC 1. 1. 1.44); malate dehydrogenase (MDHs) (EC 1.1.1.37); and glucose-phosphate isomerase (GPI) (EC 5.3.1.9). PEPb, GOTm, 6PGD and MDHs were run on a sodium phosphate buffered gel (pH 6.3), and GPI was run using a Tris-EDTA–maleate–$MgCl_2$ buffered gel (pH = 7.6).

Table 2 List of haplotypes for the 433-bp fragment of *cytochrome b* from *Biorhiza pallida*. Parsimony-informative sites are indicated by an asterisk (*) and singletons by a full stop (.)

			1	112	222	233	333	334	4
	111	566	990	170	124	902	224	771	1
	034	912	020	535	499	321	250	394	8
	**.	**.	.**	***	*..	***	***	.**	.
Haplotype 1	GTA	GTA	GAT	GTT	TAC	ATT	TGC	ATC	T
Haplotype 2	...	..G	...	...	.G.	...	...	...	.
Haplotype 3	...	...	...	...	...	...	...	...	C
Haplotype 4	...	...	...	...	...	...	...	G..	.
Haplotype 5	...	...	...	...	..T	...	.A.	...	.
Haplotype 6	...	...	...	...	...	...	...	..T	.
Haplotype 7	...	...	...	A..	...	...	..T	...	.
Haplotype 8	...	A..	...	A.C	C..	GAC	.A.	.AT	.
Haplotype 9	...	A..	A..	A.C	C..	GAC	.A.	.AT	.
Haplotype 10	...	A..	.G.	A.C	C..	GAC	.A.	.AT	.
Haplotype 11	A..	.A.	.G.	A.C	C..	.AC	CA.	..T	.
Haplotype 12	A.G	.A.	.G.	A.C	C..	.AC	CA.	..T	.
Haplotype 13	..C.	...	.GC	ACC	...	GAC	C.T	..T	.

Haplotype 1: St Jean pied de Porte, France; Clermont-l'Hérault, France; Szeghalom, Hungary; Broughton, U.K.; Luin, Switzerland; Szentendre, Hungary; Oxford, U.K.; Cambridge, U.K. Haplotype 2: Puerto de Velate, Northern Spain. Haplotype 3: Mátrafüred, Hungary. Haplotype 4: St Jean pied de Porte, France. Haplotype 5: Bajna, Hungary. Haplotype 6: Dublin, Ireland. Haplotype 7: Casina, Italy. Haplotype 8: Avila-El Escorial, central Spain. Haplotype 9: Guadalix de la Sierra, central Spain. Haplotype 10: Guadalix de la Sierra, central Spain; Zarzalejo, central Spain; Villaviciosa, central Spain; Los Molinos, central Spain; Soto del Real, central Spain. Haplotype 11: Cercedilla, central Spain; Zarzalejo, central Spain. Haplotype 12: Cercedilla, central Spain. Haplotype 13: Prado del Rey, southern Spain (five individuals).

Analysis of sequence data

Sequences were aligned by CLUSTALW (Thompson *et al.*, 1994) using the default settings. ITS1 and ITS2 alignments were manually checked to verify that there were no ambiguities. Departures from neutrality were tested using Tajima's D (Tajima, 1989) and Fu and Li's D^* and F^* (Fu & Li, 1993) statistics incorporated into the software program (DNASP, version 3.0) (Rozas & Rozas, 1999). For ITS1 and ITS2, sites with alignment gaps were excluded from the neutrality tests. All generated sequences were used (including the multiple clones from each individual).

Phylogenies for cytochrome *b* were generated using parsimony (MP) and maximum likelihood (ML) in PAUP* (version 4.0b3) (Swofford, 1999). MP was performed for 1000 bootstraps replicated with all sites equally weighted, using the tree bisection-rooting (TBR) option in a heuristic search. All ML analyses were heuristic searches using the TBR option. To reduce computational time, only 100 bootstrap replications were performed for each ML analysis. A hierarchical series of increasingly complex models of sequence evolution was employed to identify the model that made the data most likely using likelihood ratio test (LRT) statistics (reviewed by Huelsenbeck & Rannala, 1997). We tested, singly and together, the effects of unequal base frequencies, different rates between transitions and transversions (ti/tv) and rate variation over nucleotide sites (Hasegawa *et al.*, 1985). The assumption of among-site rate heterogeneity (Yang, 1993) and the enforcement of a molecular clock were also tested for the best-fitting model. The shape parameter α of the gamma distribution and the ti/tv ratio were calculated from a 50% majority rule consensus tree using unweighted parsimony. Whenever unequal base frequencies were employed, we used the empirical frequencies of the nucleotides (Yang *et al.*, 1994).

For ITS, phylogenies were constructed using combined data for ITS1 and ITS2. Each gap was coded as missing and as a fifth nucleotide character. Trees were constructed using MP (as above) and ML, using PUZZLE (Strimmer & von Haeseler, 1996). MP bootstraps were carried out as above for 100 replicates. ML was performed using the quartet puzzling algorithm with 10 000 puzzling steps, the HKY85 model of DNA substitution (Hasegawa *et al.*, 1985) and a gamma-shaped distribution to account for rate heterogeneity. All parameters were estimated from the dataset. A parsimony analysis was also performed using the gap insertions as unique indel (insertion/deletion) events. They were treated separately from the sequence dataset because no data exist concerning their frequency of substitution.

Results

Wolbachia *screening and diversity*

Of the 206 *B. pallida* successfully screened, 85% were infected with *Wolbachia*. Over the entire sampled range only populations in central Spain were not infected (Fig. 1, Table 1). All sequences were 564-bp long (GenBank accession number AF339629), with only two nucleotide positions polymorphic. Position 218 was polymorphic for T/C and position 230 was polymorphic for A/G in all the individual sequences. As the sequencing was done from a PCR fragment, it could not be determined whether these sites are genuinely polymorphic (not fixed yet, four alleles) or represent two different strains (two alleles). The low *wsp* diversity suggests that all infected European populations carry the same strain of *Wolbachia*.

Mitochondrial phylogeography

The 29 individuals sequenced for the cytochrome *b* fragment were polymorphic for 25 positions, 18 of which were parsimony informative (4.18%). This variation yielded a total of 13 discrete haplotypes (see Table 2). Likelihood ratio tests showed that the best-fitting model for the *B. pallida* data used empirical base frequencies, a ti/tv ratio equal to 10.44 and variation in rate among sites (using a gamma distribution with $\alpha = 0.0057$ and four rate categories). The estimated value $\alpha = 0.0057$ denotes a very strong rate variation (see Yang *et al.*, 1994). The assumption of a molecular clock could not be rejected at the 0.05% probability level (unconstrained model *versus* model with molecular clock enforced, $-\Delta L = 44.42$, d.f = 27, $0.01 < P < 0.05$).

All MP and ML analyses reveal a deep and well-supported split (based on 10 changes) between populations in central and southern Spain *versus* those in northern Spain and the rest of Europe (Fig. 2). Sequence diversity was higher in *Wolbachia*-free populations in central and southern Spain ($P = 0.0082$) than in the *Wolbachia*-infected populations throughout northern Spain and the rest of Europe ($P = 0.0029$). The 14 sequences from the populations in northern Spain and the rest of Europe showed significant deviation from predictions under neutrality (Tajima's $D = -2.09$, $P < 0.05$; Fu and Li's $D^* = -2.73$, $P < 0.05$ and $F^* = -2.93$, $P < 0.05$). In contrast, central and southern Spanish sequences, from either the 10 uninfected central Spanish individuals (Tajima's $D = 0.52$, $P > 10$; Fu and Li's $D^* = 0.62$, $P > 0.10$ and $F^* = 0.67$, $P > 10$), or the central and southern Spanish individuals combined (15 sequences) (Tajima's $D = 1.20$, $P > 0.10$; Fu and Li's

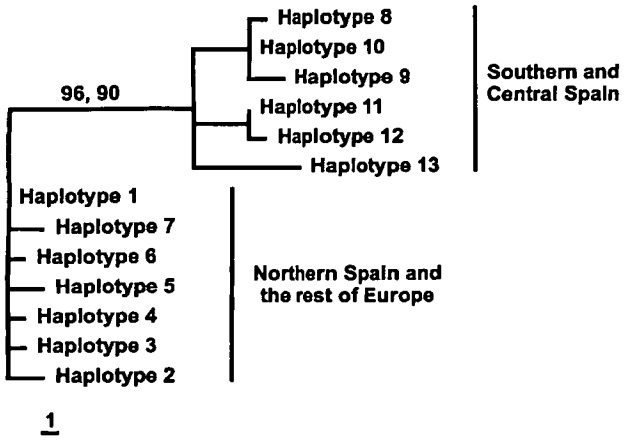


Fig. 2 Consensus phylogram (50% majority rule) for cytochrome *b* haplotypes using an ML model that accounts for unequal base frequencies, a different ti/tv ratio and rate variation among sites. See Table 2 for haplotype information. Values above the branches denote bootstrap support under ML and MP, respectively. The tree is 35 steps long with a consistency index (CI) of 0.714 and a retention index (RI) of 0.931. GenBank accession numbers: AF339616–AF339628.

$D^* = 0.89$, $P > 0.10$ and $F^* = 1.13$, $P > 0.10$), showed no significant deviation from neutrality.

ITS phylogeography

The 13 individuals sequenced for ITS2 and the six individuals sequenced for ITS1 and ITS2 revealed 43 parsimony-informative positions (Table 3). The length of the fragments varied between 456 and 467 bp for ITS2 and between 626 and 632 bp for ITS1. Length differences from clones from the same individual were usually much smaller or absent. The ML tree for ITS constructed with the quartet puzzling algorithm, and the MP tree coding alignment gaps as missing information, were both poorly resolved. However, MP analysis with gaps coded as a fifth base strongly supports the split of the population from central and southern Spain *versus* that from northern Spain and the rest of Europe revealed by the mitochondrial cytochrome *b* data (Fig. 3). MP analysis of the ITS indel dataset (with 10 of 12 indels being parsimony informative) revealed the same topology, although with lower bootstrap support.

Deviation from neutrality was tested separately for ITS1 and ITS2. Given that these are nuclear loci, within which recombination might be occurring, the estimates for deviation from neutrality are conservative. Significantly, regional variation in departures from neutral expectations parallel those seen in cytochrome *b*. For both loci, Fu and Li's D^* and F^* and Tajima's D showed significant departure from neutrality for populations outside Spain (ITS1: Tajima's $D = -1.92$, $P < 0.05$; Fu and Li's $D^* = -2.77$, $P < 0.05$ and $F^* = -2.92$, $P < 0.05$. ITS2: Tajima's $D = -1.98$, $P < 0.05$; Fu and Li's $D^* = 4.15$, $P < 0.05$ and $F^* = 4.05$, $P < 0.05$) but were nonsignificant for central Spanish specimens (ITS1: Tajima's $D = -1.01$, $P > 0.10$; Fu and Li's $D^* = -0.99$, $P > 0.10$ and $F^* = -1.07$, $P > 0.10$. ITS2: Tajima's $D = -0.77$, $P > 0.10$, Fu and Li's $D^* = 0.86$, $P > 0.10$ and $F^* = -0.95$, $P > 0.10$).

Allozyme variability

The sample sizes obtained from each location (maximum 12 galls, and so 12 females screened) were too small to allow meaningful analysis of allele frequencies. We thus limit our interpretation of the data to the presence/absence of specific alleles, as summarized in Table 4. Four alleles were found only in Spain. Of these, two were present only in central Spain; allele 1 at 6PGD was present only in one population, whereas allele 2 at 6PGD was present in four populations. The third allele was present in the only southern Spanish population (allele 4 at GPI) and one was present in all 10 Spanish populations (allele 1 at MDHs). Populations in the rest of Europe possessed two alleles absent from Spain. One was from a single individual from Stoughton, U.K. (allele 2 at GPI) and the other was very common in individuals from Phoenix Park, Ireland (allele 4 at MDHs). These regional differences in locally restricted alleles support the substantial genetic divergence between central and southern Spain *versus* northern Spain and the rest of Europe implied by the cytochrome *b* and ITS sequence analyses.

Discussion

Geographic variation in Wolbachia *infection*

Extensive screening revealed that, with the exception of central Spain, all European populations of *B. pallida* sampled are infected with *Wolbachia*. The apparent single strain infecting *B. pallida* appears to be relatively cosmopolitan; it is shared with three other European gallwasp species (Rokas, unpublished data), and database searches suggest that it also infects tse-tse flies. There are three possible explanations for existence of infected populations of *B. pallida* both north and south of a *Wolbachia*-free region in central Spain: (a) there was a single infection event and central Spanish populations have lost their infection; (b) there was a single infection in a common ancestor of the southern Spanish, northern Spanish and the rest of Europe *B. pallida* populations not shared with populations in central Spain; or (c) infection has occurred independently in southern Spain

Table 3 Parsimony-informative sites for ITS1 and ITS2. A dash (–) denotes a gap and a full stop (.) denotes an identical site. Individuals with the same parsimony-informative sites were grouped together (see below.) Numbers indicate number of individual and number of clone, respectively. All numbers refer to the individuals in Fig. 3. An asterisk (*) indicates individuals that were sequenced for both ITS1 and ITS2. Individuals without an asterisk were sequenced only for ITS2. Groups – **1:** 123.5; **2:** 140.10*; **3:** 117.3*, 123.11, 123.13, 48.4, 48.11, 53.4, 212.11, 35.3, 48.12, 53.11, 96.20*, 53.2, 48.7, 35.1, 48.3; **4:** 212.3, 212.9, 212.7, 212.8, 96.13*; **5:** 96.10*, 96.18*, 96.16*; **6:** 117.15*; **7:** 117.2*; **8:** 123.1; **9:** 141.6*, 140.5*, 141.10*, 141.3*, 48.5; **10:** 96.19*; **11:** 140.13*; **12:** 141.5*; **13:** 141.12*; **14:** 147.5*; **15:** 147.14*; **16:** 147.15*; **17:** 157.6, 157.16; **18:** 157.13, 157.7, 172.5; **19:** 147.19*; **20:** 147.13*; **21:** 32.5*; **22:** 32.7*, 32.3*, 32.8*; **23:** 32.10*; **24:** 147.3*; **25:** 123.3

	ITS1	ITS2
	2 2 2 3 4 4 4 4 4 5 5 5 5 5	1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4
	8 9 4 7 7 5 5 7 8 8 0 0 1 1 1	1 3 4 3 6 6 0 1 1 1 1 4 4 4 1 4 4 5 5 8 9 0 0 0 0 0 1 2
Group	5 4 5 6 7 5 2 2 3 4 5 6 3 4 5	5 8 8 9 1 2 6 0 1 2 3 6 7 8 6 8 9 0 1 7 9 5 6 7 8 9 6 2
1	? ? ? ? ? ? ? ? ? ? ? ? ? ? ?	T - - A - - T T A - - - - - A G T - - C T A - - - - - T
2	A A A - - T A A - - T A A T T	. - - . - - . . . - - - - - . . . G T . . . - - - - A .
3	A A A - - T A A - - T A A T T	. - - . - - . . . - - - - - . . . - - . . . - - - - A .
4	. . . A A C . T - -	C - C . - - G . . - - - - - . . . - - . . . - - - - A .
5	. . . A A . . T - -	C - C . - - G . . - - - - - . . . - - . . . - - - - A .
6	. . . - - . . T - - . . - - -	. - C G - - G . . - - - - - . . . - - . . . - - - - A .
7	. . . - - . . . - -	. - - G - - . . . - - - - - . . . - - . . . - - - - A .
8	? ? ? ? ? ? ? ? ? ? ? ? ? ? ?	. - - . - - . . . T A C A T . . . G T . A . - - - - A .
9	. . . - - . G T - - . . - - .	C - - . G T . - - - - C A T . . . G T . A . - - - - A .
10	. . . A A . . T - -	C - - . - - . . . T A C A T . . . G T . A . A - - - A .
11	. . . - - . . . - -	C - - . G T . - - - - C A T . . . G T . A . A - - - A .
12	. . . - - C G T - - . . - - -	C - - . G T . - - - - C A T . . . G T . A - - - - - A .
13	. . . - - . G T - - . . - - -	C - - . G T . - - - - C A T . . . G T . A - - - - - A .
14	. . - - - . . T T A - - - - -	C T - . G T . - - - - - - - . - - - - . A . A - - - A .
15	. . - - - . . T T A - - - - -	C T - . G T . - - - - - - - . - - - - . A . A - - A A .
16	. . - - - . . T T A - - - - -	C T - . G T . - - - - - - - . - - - - . A . - - - - A .
17	? ? ? ? ? ? ? ? ? ? ? ? ? ? ?	C A - . G T . - - - - - - - . - - - - T A G A A A A T C
18	? ? ? ? ? ? ? ? ? ? ? ? ? ? ?	C A - . G T . - - - - - - - . - - - - . A G A A A A T .
19	. T - - - . . T T A - - - - -	C A - . G T . - - - - - - - . - - - - . A - - - A A A .
20	. T - - - . . T T A - - - - -	C A - . G T . - - - - - - - . - - - - . A A - - - - A .
21	C . . - - . . T - - - - - - -	C T - . G T . - - - - - - - G - - - - . A . A - - - A .
22	C . . - - . . T T A - - - - -	C T - . G T . - - - - - - - G - - - - . A . - - - - A .
23	C . . - - . . T T A - - - - -	C T - . G T . - - - - - - - G - - - - . A - - - - - A .
24	. . - - - . . T T A - - - - -	C T - . G T . - - - - - - - G - - - - . A - - - - - A .
25	? ? ? ? ? ? ? ? ? ? ? ? ? ? ?	C - - . G T . - - - - C A T . . . - - . . . - - - - A .

and in Europe from the Pyrenees northwards. Too little is currently known about the mechanisms of *Wolbachia* transfer between lineages for these possibilities to be distinguished, although evidence for horizontal transfer of *Wolbachia* between and within species is accumulating (Vavre *et al.*, 1999; Huigens *et al.* 2000). We also know very little about the phenotypic consequences of *Wolbachia* infection in *B. pallida*. *Wolbachia* has been shown to induce parthenogenesis in some non-oak cynipids (e.g. Plantard *et al.*, 1999), but the presence of males in all our samples suggests that this does not occur in *B. pallida*. Additionally, the slow lifecycle of this host, and obligate development of the larva within oak tissues make the mating and curing experiments needed to examine *Wolbachia*-associated effects (such as for example, cytoplasmic incompatibility) extremely difficult (see also Plantard *et al.*, 1998; Plantard *et al.*, 1999).

Patterns of genetic diversity in B. pallida

Variability at the mitochondrial locus was higher in populations free of *Wolbachia* than in the northern Spanish and the rest of Europe populations infected with the symbiont. Similarly, infected northern Spanish and rest of Europe populations showed significant departures from neutrality, while central and southern Spanish populations showed no such signature. Such covariation between infection and sequence diversity is compatible with a *Wolbachia*-induced selective sweep.

Large-scale demographic processes such as range expansion can have a similar impact on mtDNA variation to that caused by a selective sweep in a population (Donnelly & Tavaré, 1995). Studies on other European oak gallwasps predict a significant impact of range expansion on spatial patterns in *B. pallida*, and should be regarded as a more parsimonious underlying cause if compatible with the data. As discussed above, demographic processes tend to generate qualitatively similar patterns at both mitochondrial and nuclear markers. This is exactly what we see in *B. pallida*, indicating that lower genetic diversity in central and northern Europe is far more likely to result from either historically low genetic diversity in nearby refugia, or loss of genetic diversity associated with range expansion, rather than a *Wolbachia*-associated selective sweep. *Wolbachia* infection may also generate similar patterns of genetic diversity in nuclear and mitochondrial markers, provided that the effect of *Wolbachia* on its host is to induce parthenogenesis. However, the presence of males in every population of *B. pallida* that we collected argues against a scenario of *Wolbachia*-induced parthenogenesis.

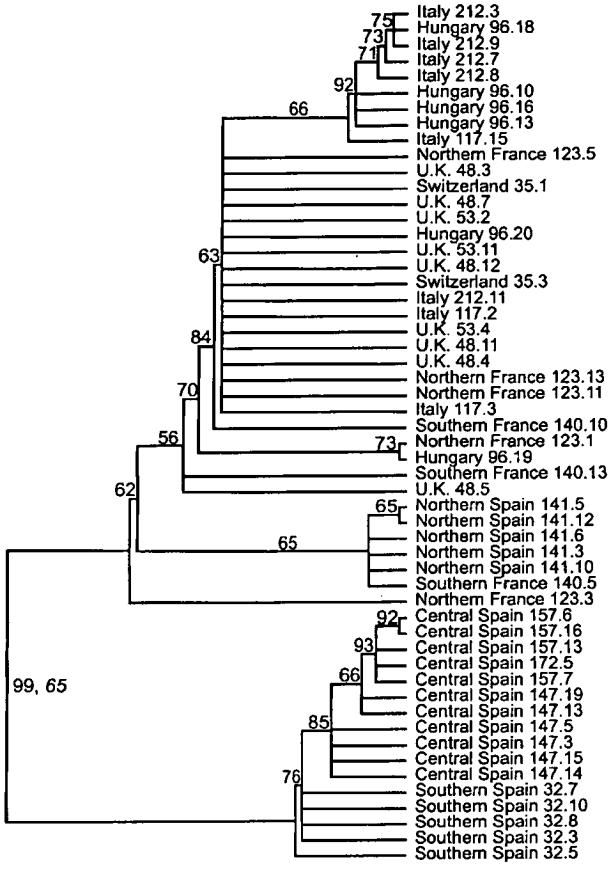


Fig. 3 Consensus tree (50% majority rule) for ITS using MP, with gaps coded as the fifth base. Numbers in branches denote bootstrap support under MP (alignment gaps as fifth base). For southern and central Spain *versus* northern Spain and the rest of Europe split, the bootstrap support given by MP analysis of the indel dataset is shown in ***italics***. Numbers next to countries' names indicate number of individual and number of clone, respectively. The tree is 141 steps long with a consistency index (CI) of 0.794 and a retention index (RI) of 0.925. GenBank accession numbers: ITS1, AF340069–AF340096; ITS2, AF340097–AF340150.

The dominant feature of genetic variation in *B. pallida* is the deep split between the central and southern Spanish population *versus* those of northern Spain and the rest of Europe, supported by all the *B. pallida* datasets. The division of these two groups is robustly supported by the cytochrome *b* (Fig. 2), and less robustly by the ITS (Fig. 3) analyses. The lower support observed in ITS is due to the fact that the informative sites are in the indel regions. When gaps are encoded as the fifth base or when they are treated as unique indel insertions, the split between central and southern Spain *versus* northern Spain and the rest of Europe is more strongly supported. Genetic discontinuities between areas north and south of the Pyrenees are known for many plants and animals, and many show hybrid zones at the Pyrenees (see reviews by Taberlet *et al.*, 1998; Hewitt, 1999). For many species, including gallwasps (Stone & Sunnucks, 1993; Stone *et al.*, 2001), expansion following retreat of the ice was principally from Italy and the Balkans, while Spanish populations failed to expand far into France. A similar scenario could explain the patterns seen in *B. pallida*. Region-specific haplotypes and nuclear alleles imply the existence of discrete refuge populations in Spain and in regions to the east. Some oak gallwasps show further differentiation between distinct refuges in Italy and the Balkans

Table 4 *Biorhiza pallida* allozyme alleles from 5 allozyme loci, from 4 areas of Europe. Locally restricted alleles are written in bold. Numbers in parentheses denote the number of individuals and number of populations screened, respectively

Area/allozyme	6PGD	GOTm	PEPb	MDHs	GPI
Central and southern Spain (43–10)	**1,2**,3,4	2,4	**1**,2,3,4	1,2,3	3,**4**
Southern France (20–3)	4	4	1,2,3,4	2	3
Italy, Hungary (51–11)	4	2,4	1,2,3,4	2	3
U.K., Eire, Northern France (156–30)	3,4	4	1,2,3,4	2,3,**4**	**2**,3

(Atkinson, 2000), but there is inadequate resolution in the *B. pallida* data to see if the same is true for this species. Range expansion from central or eastern Europe and associated rapid population growth would generate both the similarity among sites and the departure from expectations of neutrality seen outside Spain. Inability of Spanish populations to expand and spread beyond the Pyrenees would lead to absence of Spanish haplotypes or nuclear alleles from more northerly populations, and the absence of any deviation from neutrality. Inability to escape from the Spanish refuge has been clearly demonstrated in one other oak gallwasp, resulting from the evolution of oak-specific ecotypes whose hosts are essentially limited to Spain (Stone *et al.*, 2001). It is interesting to note that patterns of post-Pleistocene range expansion for gallwasps and for their host oaks do not match. Molecular evidence for oaks suggests that individuals from all three refugia (Spain, Italy and the Balkans) contributed to the colonization of central Europe (Ferris *et al.*, 1993; Dumolin-Lapègue *et al.*, 1997). In contrast, gallwasps expanded principally from Italy and the Balkans (see above; Atkinson, 2000; Stone *et al.*, 2001). This discordance may be explained by adaptation of gallwasp populations in the Spanish refuge to local endemic oak species that failed to expand further north (for more discussion see Atkinson, 2000; Stone *et al.*, 2001).

This phylogeographic hypothesis is supported by dating estimates based upon a *B. pallida* mitochondrial molecular clock, assuming a 2.3% divergence of mtDNA sequences per million years (Brower, 1994). There are two issues for which timing estimates are meaningful. The first concerns the timing of the event leading to the current levels of genetic diversity among northern Spanish and rest of Europe populations. Assuming that the 0–0.9% divergence that is observed among these populations is the result of substitutions that have occurred just after that event, we get an estimation between 0 and 390 000 years ago. This estimation roughly encompasses the range expansion by the oak hosts of *B. pallida* following the end of the Pleistocene (Ferris *et al.*, 1993; Dumolin-Lapègue *et al.*, 1997). The second issue is the split between central and southern Spain *versus* northern Spain and the rest of Europe. Levels of divergence between the groups of around 2.5% suggest that this split is ancient, pointing to a separation long before the end of the Pleistocene. Such a division into long-standing eastern and western refugia is supported by similar data for other oak gallwasp species, such as *Andricus kollari* and *Andricus quercustozae* (Atkinson, 2000; Stone *et al.*, 2001). Our study shows that while patterns of variation in mitochondrial sequence diversity in this system do not allow a *Wolbachia*-induced selective sweep to be discounted, consideration of nuclear marker diversity points to a demographic cause.

Acknowledgements

AR is partially funded by a NERC studentship and RJA holds a BBSRC studentship. GNS and GSB are supported by a grant from the NERC EDGE programme (GR9/03553), with additional support to GNS from the Royal Society (grant no. 574006), and the British Ecological Society (Small Ecological Project Grant, no. 1462). SAW holds a BBSRC fellowship. We especially thank Gil McVean and two anonymous referees for comments on the manuscript. Thanks to Mark Blaxter for unlimited access to Mac computers for running the ML analyses; to David Guiliano for providing DNA from *Wolbachia* strain C to test our WOLG-R primer; Jill Lovell for excellent technical support with the automated sequencing and John Werren for providing the *Nasonia* control strains. Special thanks to the members of the British Plant Gall Society, György Csóka, José-Luis Nieves-Aldrey, Felix Fontal-Cazalla, Gil McVean, Olivier Plantard and Jean-Yves Rasplus for help with gall collecting.

References

ATKINSON, R. J. 2000. *The Genetic Analysis of Natural History, Reproductive Strategy and Population Structure in European Oak Gallwasps (Hymenoptera: Cynipidae)*. PhD Thesis, University of Oxford, Oxford, UK.

AVISE, J. C. 2000. *Phylogeography: the History and Formation of Species*. Harvard University Press, Harvard, USA.

BALLARD, J. W. O., HATZIDAKIS, J., KARR, T. L. AND KREITMAN, M. 1996. Reduced variation in *Drosophila simulans* mitochondrial DNA. *Genetics*, **144**, 1519–1528.

BROWER, A. V. Z. 1994. Rapid morphological radiation and convergence among races of the butterfly *Heliconius erato* inferred from patterns of mitochondrial DNA evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 6491–6495.

CSÓKA, G. 1997. *Plant Galls*. Agroinform Publishing, Budapest, Hungary.

DONNELLY, P. AND TAVARÉ, S. 1995. Coalescents and genealogical structure under neutrality. *A. Rev. Genet.*, **29**, 401–421.

DUMOLIN-LAPÈGUE, S., DEMESURE, B., FINESCHI, S. L. E., CORRE, V. AND PETIT, R. J. 1997. Phylogeographic structure of white oaks throughout the European continent. *Genetics*, **146**, 1475–1487.

FERRIS, C., OLIVER, R. P., DAVY, A. J. AND HEWITT, G. M. 1993. Native oak chloroplasts reveal an ancient divide across Europe. *Mol. Ecol.*, **2**, 337–344.

FU, Y.-X. AND LI, W.-H. 1993. Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

HASEGAWA, M., KISHINO, M. AND YANO, T. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160174.

HEWITT, G. M. 1999. Post-glacial re-colonization of European biota. *Biol. J. Linn., Society*, **68**, 87–112.

HILLIS, D. M. AND DIXON, M. T. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. *Q. Rev. Biol.*, **66**, 411–453.

HUELSENBECK, J. P. AND RANNALA, B. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science*, **276**, 227–232.

HUIGENS, M. E., LUCK, R. F., KLAASSEN, R. H. G., MAAS, M. F. P. M., TIMMERMANS, M. J. T. N. AND STOUTHAMER, R. 2000. Infectious parthenogenesis. *Nature*, **405**, 178–179.

HUNTLEY, B. AND WEBB, T. 1989. Migration – species' response to climatic variations caused by changes in the earth's orbit. *J. Biogeogr.*, **16**, 5–19.

JEYAPRAKASH, A. AND HOY, M. A. 2000. Long PCR improves *Wolbachia* DNA amplification: *wsp* sequences found in 76% of sixty-three arthropod species. *Insect Mol. Biol.*, **9**, 393–406.

JOHNSTONE, R. A. AND HURST, G. D. D. 1996. Maternally inherited male-killing microorganisms may confound interpretation of mitochondrial DNA variability. *Biol. J. Linn. Soc.*, **58**, 453–470.

MALLOCH, G., FENTON, B. AND BUTCHER, B. D. J. 2000. Molecular evidence for multiple infections of a new subgroup of *Wolbachia* in the European raspberry beetle *Byturus tomentosus*. *Mol. Ecol.*, **9**, 77–90.

MAYNARD SMITH, J. AND HAIGH, J. 1974. The hitch-hiking effect of a favorable gene. *Genet. Res.*, **23**, 23–35.

O'NEILL, S. L., HOFFMANN, A. A. AND WERREN, J. H., eds. 1997. *Influential Passengers: Inherited Microorganisms and Arthropod Reproduction*. Oxford University Press, Oxford, UK.

PLANTARD, O., RASPLUS, J. Y., MONDOR, G. L. E., CLAINCHE, I. AND SOLIGNAC, M. 1998. *Wolbachia*-induced thelytoky in the rose gallwasp *Diplolepis spinosissimae* (Giraud) (Hymenoptera: Cynipidae), and its consequences on the genetic structure of its host. *Proc. R. Soc. Lond. B*, **265**, 1075–1080.

PLANTARD, O., RASPLUS, J.-Y., MONDOR, G. L. E., CLAINCHE, I. AND SOLIGNAC, M. 1999. Distribution and phylogeny of *Wolbachia* inducing thelytoky in Rhoditini and 'Aylacini' (Hymenoptera: Cynipidae). *Insect Mol. Biol.*, **8**, 185–191.

ROZAS, J. AND ROZAS, R. 1999. DnaSP, version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics*, **15**, 174–175.

SOLIGNAC, M., VAUTRIN, D. AND ROUSSET, F. 1994. Widespread occurrence of the proteobacteria *Wolbachia and* partial cytoplasmic incompatibility in *Drosophila melanogaster*. *C. R. Acad. Sci. Paris*, **III** (317), 461–470.

STONE, G. N., ATKINSON, R., ROKAS, A., CSÓKA, G. AND NIEVES-ALDREY, J.-L. 2001. Differential success in northwards range expansion between ecotypes of the marble gallwasp *Andricus kollari:* a tale of two lifecycles. *Mol. Ecol.*, **10**, 1–18.

STONE, G. N. AND COOK, J. M. 1998. The structure of cynipid oak galls: patterns in the evolution of an extended phenotype. *Proc. R. Soc. Lond. B*, **265**, 979–988.

STONE, G. N. AND SUNNUCKS, P. 1993. Genetic consequences of an invasion through a patchy environment – the cynipid gallwasp *Andricus quercuscalicis* (Hymenoptera: Cynipidae). *Mol. Ecol.*, **2**, 251–268.

STOUTHAMER, R., BREEUWER, J. A. J. AND HURST, G. D. D. 1999. *Wolbachia pipientis:* Microbial manipulator of arthropod reproduction. *Ann. Rev. Microbiol.*, **53**, 71–102.

STRIMMER, K. AND VON HAESELER, A. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.*, **13**, 964–969.

SUNNUCKS, P. AND STONE, G. S. 1996. Genetic structure of invading insects and the case of the knopper gallwasp. In: Floyd, R. B., Sheppard, A. W. and De Barro, P. J., eds. *Frontiers of Population Ecology*, pp. 485–496. CSIRO Publishing, Australia.

SWOFFORD, D. L. 1999. *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods)*. Sinauer Associates.

TABERLET, P., FUMAGALLI, L., WUST-SAUCY, A. G. AND COSSON, J. F. 1998. Comparative phylogeography and postglacial colonization routes in Europe. *Mol. Ecol.*, **7**, 453–464.

TARMA, F. 1989. Statistical method for testing the neutral ruination hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

THOMPSON, J. D., HIGGINS, D. G. AND GIBSON, T. J. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl Acids Res.*, **22**, 4673–4680.

TURELLI, M., HOFFMANN, A. A. AND MCKECHNIE, W. S. 1992. Dynamics of cytoplasmic incompatibility and mtDNA variation in natural *Drosophila simulans* populations. *Genetics*, **132**, 713–723.

VAVRE, F., FLEURY, F., LEPETIT, D., FOUILLET, P. AND BOULETREAU, M. 1999. Phylogenctic evidence for horizontal transmission of *Wolbachia* in host parasitoid associations. *Mol. Biol. Evol.*, **16**, 1711–1723.

WERREN, J. H. 1997. Biology of *Wolbachia*. *Ann. Rev. Entomol.*, **42**, 587–609.

WERREN, J. H. AND WINDSOR, D. M. 2000. *Wolbachia* infection frequencies in insects: evidence of a global equilibrium? *Proc. R. Soc. Lond. B.*, **267**, 1277–1285.

WERREN, J. H., WINDSOR, D. AND GUO, L. 1995. Distribution of *Wolbachia* among neotropical arthropods. *Proc. R. Soc. Lond. B*, **262**, 197–204.

WEST, S. A., COOK, J. M., WERREN, J. H. AND GODFRAY, C. J. 1998. *Wolbachia* in two insect host–parasitoid communities. *Mol. Ecol.*, **7**, 1457–1465.

WHITE, T. L., BRUNS, T., LEE, S. AND TAYLOR, J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis, M. A. Gelfand, D. H. and Sninsky, J. J., eds. PCR. *Protocols. A Guide to Methods and Applications*, pp. 315–322. Academic Press, Inc., USA.

YANG, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, **10**, 1396–1401.

YANG, Z., GOLDMAN, N. AND FRIDAY, A. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol*, **11**, 316–324.

ZHOU, W., ROUSSET, F. AND O'NEILL, S. 1998. Phylogeny and PCR-based classification of *Wolbachia* strains using *wsp* gene sequences. *Proc. R. Soc. Lond. B*, **265**, 509–515.

testament to the probable costs to females of mating, and therefore differences in longevity between these females are not a result of mating.

But why do these males assess females on the basis of their future life expectancy? Seaweed flies, as the name suggests, lay their eggs in seaweed deposits that get washed up above the high tide mark on beaches. Such deposits are highly ephemeral and the longer the postmating survival of the female, the more likely she is to successfully lay her eggs. By choosing longer lived females, males are maximizing their chances of successfully leaving offspring. This is especially important in a mating system that is characterized by costly sexual conflict – by expressing such a choice males are best able to allocate their time and resources to reproduction. This example of sexual selection for longevity has important implications for the trade off between fecundity and life span, and future studies should consider these.

1 Dunn, D.W. *et al.* (2001) Male mating preference for female survivorship in the seaweed fly *Gluma musgravei* (Diptera: Coelopidae). *Proc. R. Soc. London B Biol. Sci.* 268, 1255–1258

David J. Stevens
d.stevens@bio.gla.ac.uk

Getting it right for the wrong reason

A statistical framework is necessary for almost every scientific field, but few disciplines had as much trouble as phylogenetics did in establishing one. The long and frustrating arguments over different methodologies and philosophies (cladistics versus phenetics versus evolutionary taxonomy in the earlier years – cladistics versus statistical inference more recently) might have waned but their echo still reverberates. In a short historical essay[1], Joe Felsenstein, one of the principal architects of the 'statistical revolution' in phylogenetics, tells the story of the growth of a much-troubled field with characteristic clarity and vigour. Those interested in the way that science (and scientists) proceeds and advances will find the essay illuminating, not least because it offers a nice illustration of a paradigm-shift in the making.

However, criticisms against statistical phylogenetic inference still remain. Swofford *et al.*[2] have now addressed one of the latest criticisms levelled against model-based methods (particularly likelihood); the notion that parsimony outperforms likelihood in reconstructing topologies belonging to the area of tree space known as the 'inverse-Felsenstein zone', which comprises trees with two long branches that are adjacent (Fig. 1a). This is exactly the reverse of the 'Felsenstein zone', the area of tree space where the two long-branched taxa are separated by a short internal branch (Fig. 1b). In Fig. 1b, likelihood outperforms parsimony, the latter producing strong support for the wrong tree. In both zones, the outcome from likelihood analysis is a star phylogeny (Fig. 1). Understanding the reasons behind potential biases of particular methods in the two zones is biologically relevant and significant; a glance at the phylogenetic literature of early eukaryotes will be convincing enough.

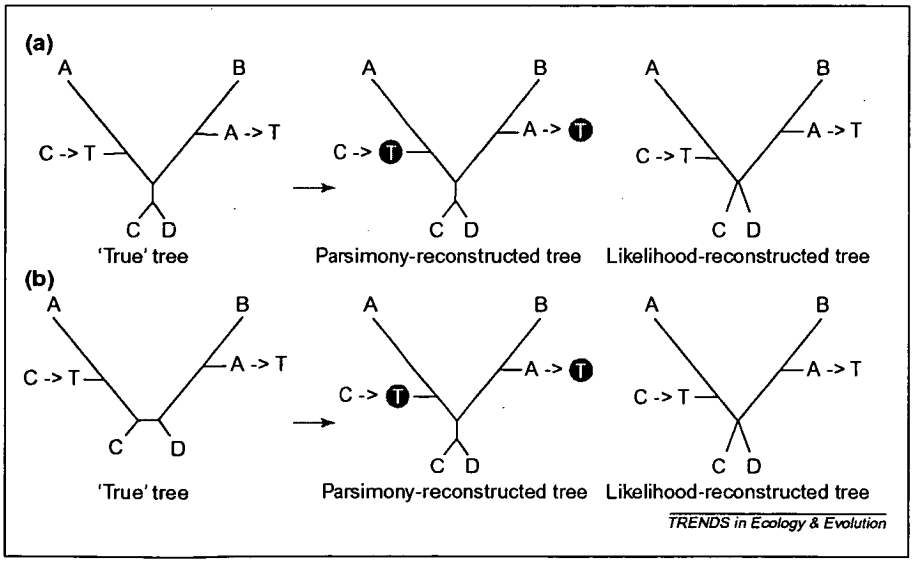


Fig. 1. Behaviour of parsimony and maximum likelihood in the (a) 'inverse-Felsenstein zone' and (b) 'Felsenstein zone'. Red dots indicate homoplasies that are mistakenly identified as homologies by parsimony.

Swofford *et al.* demonstrate convincingly that the reason behind the success of parsimony in the inverse-Felsenstein zone is the same for its failure in the Felsenstein zone. Homoplasies (i.e. parallel or convergent changes) are treated as homologies (i.e. changes revealing identity by descent), owing to the inability of parsimony to correct for multiple substitutions (red dots in Fig. 1). The extent of bias can only be appreciated in a numerical example. In the case of a four-taxon unrooted tree where there have been, on average, two changes in each of the two long branches and 0.2 changes on each of the three short ones, 97% of all apparent homologies will be misinterpreted homoplasies. It is perhaps an understatement to say that researchers will feel uncomfortable about their pet phylogenies after reading these values. By contrast, the star phylogenies produced by likelihood are a consequence of the low support of the data for the 'true' phylogeny, given that very few changes have happened on the short internal branch. Swofford *et al.* also show that UPGMA, a phenetic method whose assumption of an equal rate of change in all lineages is violated in both cases, is behaving in almost exactly the same manner to parsimony in the two zones. I doubt that parsimony advocates will find this last finding exonerating.

1 Felsenstein, J. (2001) The troubled growth of statistical phylogenetics. *Syst. Biol.* 50, 465–467
2 Swofford, D.L. *et al.* (2001) Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.* 50, 525–539

Antonis Rokas
a.rokas@ed.ac.uk