Inferring population history from genealogies

Konrad R. Lohse

Submitted for the degree of Doctor of Philosophy

University of Edinburgh

2010

Declaration

This dissertation is submitted to the University of Edinburgh in accordance with the requirements for the degree of Doctor of Philosophy in the faculty of Science. Some of the work described in this thesis was only possible through collaborations, details of which are presented below. In each case, the majority of the work is my own.

Chapter 2

Jerome Kelleher implemented the algorithm to find the number of mutations on rootward branches and provided help with running the simulations.

Chapter 3

BEAST analyses were carried out together with James Nicholls.

Chapter 4 and 5

The EST data used for primer design was collected and aligned by Barb Sharanowski.

Unless otherwise stated, the remaining work and content of this thesis are entirely my own.

Signed:

Konrad Lohse

Acknowledgments

Huge thanks my two supervisors Graham Stone and Nick Barton for their continual inspiration, patience and support and most of all for sharing their unique expertise. I am immensely grateful for having had the complete freedom to pursue my interests and thoroughly enjoyed learning from them. This would have been impossible without their help.

Thanks also to the Barton and Stone groups (Jitka Polechova, Jack Hearn, Simon Aischbacher, James Nicholls, Harold de Vladar, Frazer Sinclair, Jerome Kelleher, Pablo-Fuentes Utrilla) for their enthusiasm, help and friendship.

A great many people have assisted in climbing various mountains both literally and conceptually. Matthias Müller and Ben Brummer kept me company during many days of strenuous beetle collecting in the Alps. Andrew Rambaut, James Nicholls and Philippe Lemey helped with BEAST. Riccardo Skiaky, Alfried Vogler, Arved Lompe, Martin Baehr, Alex Weir and Pedro Oromí shared their expertise on alpine Carabids. George Melika introduced me to the wonderful world of galls and the difficult business of identifying their occupants. Majide Tavakoli, Juli-Pujade Villar and James Cook kindly provided specimens. Barb Sharanowski, Darren Obbard and Mark Blaxter taught me how to design primers, Ahmed Raza took me through the cloning protocol and Ziheng Yang explained the entrails of his software.

I would like to thank the Biotechnology and Biological Science Research Council for funding in particular for granting an extension which allowed me to attend taught courses and training throughout my PhD and the Genetics Society for travel money.

Special thanks to the many beer and coffee drinking friends from Ashworth and elsewhere in particular Jerome Kelleher, Alex Hall, Mike Hickerson and Richard Harrison for stimulation and distraction.

Most of all I thank my wife Marie for all her love which kept me sane and my daughter Carla who — despite being born in the middle of this — turned out to be the most amazing and relaxed baby.

Publications

The following papers have arisen from this thesis and are included in the Appendix:

- Lohse, K. & Kelleher, J. 2009. Measuring the degree of starshape in genealogies — Summary statistics and demographic inference. *Genetics Research*, **91**: 281–292.
- Lohse, K., Sharanowski, B., Stone, N.G. 2010. Quantifying the Pleistocene history of the oak gall parasitoid *Cecidostiba fungosa* using twenty intron loci. *Evolution*, *in press*.

I have also pursued some other projects on related topics that were not directly part of my thesis. One of these resulted in the following paper which is included in the Appendix.

- Lohse, K. 2009. Can mtDNA barcodes be used to delimit species? A response to Pons et al. (2006). *Systematic Biology*, **58**(4): 439–442.

Contents

1 Introduction **1**

1.1 Classic models of population structure . . . 2

1.2 Phylogeography . . . 3

1.3 The neutral coalescent . . . 3

1.4 Inference methods . . . 6

1.5 Thesis outline and aims . . . 8

2 Measuring the degree of starshape in genealogies — Summary statistics and demographic inference **12**

2.1 Summary Statistics . . . 14

2.1.1 Genealogical ratios . . . 15

2.1.2 Starting from the limiting case . . . 17

2.2 Methods . . . 18

2.2.1 Summary statistics and demographic model . . . 18

2.2.2 Power test . . . 18

2.2.3 Likelihood method . . . 19

2.3 Results . . . 20

2.3.1 Single locus . . . 20

2.3.2 Multiple loci . . . 22

2.4 Discussion . . . 24

2.4.1 General limits to demographic inference . . . 24

2.4.2 Pairwise measures . . . 27

2.4.3 Recombination and topological uncertainty . . . 28

2.4.4 Conclusions . . . 29

3 Inferring the colonisation of a mountain range - refugia vs. nunatak survival in high alpine ground beetles 30

3.1 Materials and Methods . . . 36

3.1.1 Sampling . . . 36

3.1.2 Molecular work . . . 36

3.1.3 Phylogenetic analysis . . . 37

3.1.4 Bayesian inference of relationships among populations . . . 38

3.1.5 Testing topological constraints . . . 40

3.2 Results . . . 40

3.2.1 Phylogenetic analysis . . . 40

3.2.2 Bayesian inference of relationships among populations . . . 41

3.2.3 Testing topological constraints . . . 43

3.3 Discussion . . . 47

3.3.1 Patterns and causes of phylogeographic structure . . . 48

3.3.2 Locations as states in gene trees . . . 49

4 Developing EPIC primers for chalcid Hymenoptera from EST and genomic data 51

4.1 Methods . . . 53

4.1.1 Choice of nuclear loci and EST libraries . . . 53

4.1.2 Primer design . . . 54

4.1.3 Screening amplification success . . . 54

4.1.4 Divergence, diversity and information content . . . 55

4.2 Results . . . 57

4.2.1 Screening amplification success . . . 57

4.2.2 Divergence, diversity and information content . . . 57

4.3 Discussion . . . 60

5 Quantifying the Pleistocene history of the oak gall parasitoid *Cecidostiba fungosa* using twenty intron loci 63

5.1 Methods . . . 67

5.1.1 Choice of loci . . . 67

5.1.2 Molecular methods . . . 67

5.1.3 Model of population divergence and population sampling strategies . . . 68

5.1.4 Alignment and mutation rate . . . 69

5.1.5 Recombination tests and gene tree reconstruction 70
5.1.6 Maximum Likelihood analysis 71
5.1.7 Bayesian estimation of divergence times and ancestral population sizes 72
5.2 Results 72
5.2.1 Gene trees 72
5.2.2 Maximum likelihood analyses 73
5.2.3 Bayesian estimation of divergence times and ancestral population sizes 76
5.2.4 Gene divergence times 82
5.3 Discussion 82
5.3.1 Ancestral N_e and sampling 86
5.3.2 Assumptions and extensions of the model 87
5.3.3 Towards a multilocus approach to community phylogeography 88

6 Topological probabilities in models of divergence with gene flow 90
6.1 Model and derivation 92
6.1.1 Asymmetric migration 93
6.1.2 Limit cases 96
6.1.3 Symmetric migration 97
6.2 Discussion 100

7 Discussion 103
7.1 Conclusions 103
7.1.1 Chapter 2 103
7.1.2 Chapter 3 104
7.1.3 Chapter 4 104
7.1.4 Chapter 5 105
7.1.5 Chapter 6 106
7.2 Outlook 106

Abstract

This thesis investigates a range of genealogical approaches to making quantitative inferences about the spatial and demographic history of populations with application to two insect systems: A local radiation of high alpine ground beetles (Carabidae) in the genus *Trechus* and major refugial populations of the oak gall parasitoid *Cecidostiba fungosa* (Pteromalidae).

i) Summary statistics, which make explicit use of genealogical information are developed. Using simulations their power to detect a history of population growth is shown to be higher than that of standard measures such as Tajima's D for single and multilocus data. The improvement arises from the fact that in contrast to pairwise measures, the new statistics are minimally confounded with the topology.

ii) A Bayesian method to reconstructing character states is used to infer the Pleistocene history of populations of high alpine *Trechus* sampled along a single mountain range from mitochondrial and nuclear data. Despite evidence for some incomplete lineage sorting, a simple model of a series of extreme founder events out of two refugia during or before the last glacial maximum provides a good fit to the data.

iii) A large set of exon-primed, intron-spanning (EPIC) loci is developed for Hymenoptera from EST and genomic data. Amplification success is screened on a range of Hymenopteran species associated with two insect-plant interactions: Oak galls and figs.

iv) Borrowing model-based approaches developed to quantify species divergence, the new EPIC loci are used to investigate the relationships between three major European refugia in the oak gall parasitoid *C. fungosa*. These analyses reveal strong support for an eastern origin, effective ancestral population sizes comparable to insect model species and evidence for recent population divergence during the last interglacial. The results also suggest that there is significant information in minimal samples provided a large number of loci are available.

v) Results for the probability of gene tree topologies are derived for a model of divergence with gene flow between three populations. I outline how the asymmetries in the frequency of gene tree topologies may be used to distinguish incomplete lineage sorting from migration and discuss the results in the context of next generation sequence data from *D. melanogaster* and humans and Neanderthals.

Chapter 1

Introduction

Understanding the spatial and demographic history of populations and species has been central to evolutionary theory from its very beginning. In fact, geographic distribution is the only topic to which Darwin devotes two entire chapters in the Origin of Species (Darwin, 1859). This initial work is either theoretical or based on comparisons of species distributions and considers the implications of geography for the process of speciation and the factors shaping the composition of regional faunas and floras (Darwin, 1859; Wallace, 1876; Jordan, 1905; Holdhaus, 1954). Molecular data provide an independent source of information about the history of species that has enabled researchers to put many of those early ideas to the test. For instance, evolutionary biologists have used sequence data to ask whether particular climate episodes have acted as drivers of speciation (e.g. Klinka & Zink, 1997; Knowles, 2001) or to what extent ecologically linked taxa share spatial histories (e.g. DeChaine & Martin, 2006; Hayward & Stone, 2006). In other cases, such as our own species and its diseases, being able to make historical inference from sequence data is obviously of direct cultural and medical relevance (e.g. Lemey *et al.*, 2009; Green *et al.*, 2010). Alternatively, studies aiming at identifying genes under selection may not be directly interested in population history, but nevertheless rely on realistic null models against which the signature of past selection acting on particular genes can be tested.

Given this broad and varied interest in spatial and demographic history and the difficulty in choosing between the potentially infinite number of histories, it is is perhaps not surprising that the study of structured populations is a historically divided field (Hey & Machado, 2003). This division has — at least in the past — been a practical one between the study of model organisms and humans, for which genetic tools and data are abundant, and non-model organisms, which may have very interesting histories but few available genetic resources to infer them. However, there is a deeper, conceptual divide between phylogeography, which emphasizes the information contained in genealogies, and population genetics, which sees genealogies as essentially random outcomes of genetic drift.

Although the composition of this thesis does by necessity reflect these different traditions, a general goal throughout has been to bridge the gap between them by investigating how genealogical approaches can be used to improve population genetics inferences and, *vice versa*, how population genetics methods and sampling schemes may be applied to make robust phylogeographic inferences in non-model systems. While each chapter includes its own specific introduction, the general introduction below outlines the conceptual differences between population genetics and phylogeography. It will become clear that coalescent theory provides an elegant quantitative framework that naturally encompasses both the "tree-thinking" of phylogeography and the sound, population genetics view of genetic drift. The large number of recent reviews on this topic (Maddison, 1997; Nichols, 2001; Hey & Machado, 2003; Knowles, 2004, 2009; Machado *et al.*, 2005; Degnan & Rosenberg, 2009; Edwards, 2009; Nielsen & Beaumont, 2009; Hickerson *et al.*, 2010) bear witness to the fact that this synthesis is now well under way.

1.1 Classic models of population structure

Classic population genetics theory studies the effects of mutation, drift, selection, recombination and dispersal on allele frequencies. In the simplest case of a large, randomly mating population of constant size with discrete, non-overlapping generations, drift can be described by a single quantity, the effective population size N_e (Fisher, 1930; Wright, 1931); for example the variance in allele frequency of two alleles with frequency p and q, increases at rate $pq/2N_e$ per generation (the factor of 2 enters because it is standard to assume a diploid population). Furthermore, a randomly chosen gene copy has chance of $1/2N_e$ of going to fixation in the population and, if it does so, takes on average $4N_e$ generations. This null model, also known as the Wright-Fisher model was extended early on to investigate the effects of population structure. Perhaps the simplest model of structure is the symmetric island model, a set of subpopulations or demes which are connected to each other through migration occurring at rate m per generation. Sewall Wright (1931) derived results for the distribution of allele frequencies under this model showing that the between population component of the variance of allele frequencies (F_{ST}) is inversely proportional to the scaled migration rate $M = 2N_em$. Wright's famous equilibrium solution for F_{ST}, a measure of genetic differentiation between populations (Wright, 1951), has been widely used and abused (Whitlock & McCauley, 1999) to estimate the number of migrants from allozyme and microsatellite data. Because the symmetric island model assumes that demes are statistically exchangeable (in other words migrants are equally likely to disperse into any deme), it does not contain any measure of geographic distance. More realistic models that capture spatial structure include models of stepping stone migration (Wright, 1943; Weiss & Kimura, 1965; Malécot, 1969) and continuously distributed populations (Wright, 1943). In both cases, migration and hence reproduction are more likely to occur between individuals from neighboring demes (stepping stone model) or nearby locations (geographic continuum), leading to a pattern of isolation by distance (Wright, 1943). Importantly, classic population genetics results for these models are phrased in terms of allele frequencies as the population evolves forwards in time.

1.2 Phylogeography

In contrast, the field of phylogenetics, which seeks to reconstruct the evolutionary relationships between species and has its roots in systematics, is fundamentally backwards-looking. In a seminal paper Avise (1987) proposed that mitochondrial DNA sequences sampled at different locations within species could be used to reconstruct genealogies, which in turn should be informative about the underlying geographic history. Avise's paper marks the beginning of the field of phylogeography and features an illuminating figure depicting the fundamental connection between deep level phylogenies and population level pedigrees through a series of increasing magnifications (Avise, 1987, figure 1). Although Avise's expressed hope was that phylogeography could bridge the gap between systematics and population genetics, the field initially developed largely in isolation from population genetics. Phylogeography's focus on mitochondrial DNA and its embrace of cladistic methods, which seemed to provide the obvious tools for the analysis of trees, if anything deepened the divide. While the emphasis on the information contained in genealogies and the adoption of phylogenetic methods (Nei & Kumar, 2000) to reconstruct them, constituted an important step, the phylogeographic inference of historical scenarios itself remained a largely descriptive exercise. Attempts to formalize this inference include Templeton's nested-clade phylogeographic analysis (NCPA), a method that relies on summary statistics to measure the spatial spread of clades in a genealogy (Templeton *et al.*, 1995). While the significance of the correlation between genealogy and geography is assessed using randomization tests, likely historical scenarios are inferred qualitatively and clade by clade through an inference key, in a process similar to key-based taxonomic identification (Knowles, 2002).

1.3 The neutral coalescent

A few years before the field of phylogeography took off, population geneticists underwent a crucial shift from thinking in terms of allele frequencies forwards in time to considering the ancestry of samples backwards in time. This focus on the ancestry of samples, which has important precursors in Malécot's notion of identity by descent (Malécot, 1969), was driven by the increasing availability of genetic data. Ewens' sampling formula, describing the frequency distribution of allelic types in a sample under the infinite alleles mutation model (Ewens, 1972), marks the first step towards viewing genetic drift backwards in time. The formulation of the neutral coalescent by Kingman (1982) and Hudson (1983) as the mathematical description of the ancestral process of a sample from a Wright-Fisher population completes this transition. In the words of Wakeley (2008), "The demonstration that a relatively simple ancestral process exists for a sample was a major advance in population genetics." The elegance and simplicity of the neutral coalescent is indeed striking. For a sample of n lineages, the chance that a pair shares a common ancestor (i.e. coalesces), in any generation is given by the number of possible pairs $\binom{n}{2} = n(n-1)/2$ and the effective population size N_e. More precisely, the rate of coalescence is $\lambda = \binom{n}{2}/2N_e$. Scaling time in units of $2N_e$ generations, the times between successive coalescence events (T_i), where i denotes the number of lineages in each interval, have the following probability density function:

$$f(T_i) = \binom{i}{2} e^{-}\binom{i}{2} t \tag{1.1}$$

The fundamental property of the neutral coalescent is that genealogies are highly random both in topology and branch lengths. In fact, since the Wright-Fisher model assumes random mating, all lineages are equally likely to coalesce and thus all topologies are equally probable. Similarly, the variance in the time between successive coalescence events, which determine the branch lengths of the genealogy, is considerable. The variance of an exponentially distributed variable is $1/\lambda^2$, so the time until the last coalescent event has variance $2N_e^2$. Furthermore, the cumulative distribution function of $f(T_i)$ is very wide. For example, there is a 5% chance in total that the coalescence of the last pair of lineages takes less than $0.025 \times 2N_e$ or more than $3.7 \times 2N_e$ generations.

The power and great success of the coalescent is threefold. Firstly, it provides a convenient null model against which patterns observed in sequence data can be tested. In particular, it is straightforward to derive the full distribution of two basic measures of the size of a genealogy: the time to the most recent common ancestor of the sample (Tavaré, 1984; Takahata & Nei, 1985); and the total size of a genealogy (Tavaré, 1984). The latter, in turn, leads to an expression for the distribution of the number of segregating sites under the infinite sites mutation model (Kimura, 1969; Watterson, 1975; Tavaré, 1984; Wakeley, 2009). Secondly, separating the ancestral process from the occurrence of mutations and focusing on the history of a sample rather than the entire population, makes it extremely efficient to simulate sequence data under arbitrary histories and mutation models (Hudson, 1993, 2002). Finally, analytical work has extended the coalescent to more realistic population histories including equilibrium and non-equilibrium models of structure and changes in population size (Griffiths & Tavaré, 1994). In fact, many classic population genetic results can be easily and perhaps more intuitively understood in the language of the coalescent theory. For instance F_{ST} can be defined as the relative difference in expected coalescence time between a pair of genes sampled at random from the population as a whole T_T and a pair sampled from the same deme T_0 (Hudson *et al.*, 1992; Charlesworth *et al.*, 2003),

$$F_{ST} = \frac{T_T - T_0}{T_T} \tag{1.2}$$

For the symmetric island model the expected coalescence time of a pair sampled from the same deme is given by the total effective population size, i.e. $T_0 = 2dN_e$, where d is the number of demes (Slatkin, 1991). The time to coalescence for a pair sampled from the whole population is increased by the time it takes them to find themselves in the same deme, i.e. $T_T = T_0 + (d-1)^2/2dm$. Substituting into eq. 1.2 yields Wrigth's solution in the limit of large deme numbers (Charlesworth *et al.*, 2003).

The main result of extending the coalescent to models of population structure is the demonstration that the process is remarkably robust to a variety of complications and — in many cases — can be recovered through simple approximations and rescaling. Wakeley (1998, 1999) showed that given a large number of demes, the ancestry of a sample from a symmetric island model can be separated into two phases: An initial, instantaneous phase of coalescence and migration (termed the scattering phase); and a later phase during which the ancestry follows the neutral coalescent with a rate that is given by the number and size of demes and is inversely proportional to the rate of migration between them (collecting phase). While the effect of island-model population structure is to increase the effective population size, more realistic types of structure, in particular those involving fluctuations in deme size and local extinctions, tend to decrease N_e (Whitlock & Barton, 1997; Wakeley & Aliacar, 2001; Charlesworth *et al.*, 2003). Similar separations of time-scale have been applied to more general variants of the island model (Wakeley, 1999, 2001; Matsen & Wakeley, 2006), metapopulation models (Wakeley & Aliacar, 2001; Wakeley, 2004a, 2009) and models of populations in a two-dimensional continuum (Wilkins, 2004). A basic result of this theoretical work is that even under models in which lineages are most likely to coalesce in their neighborhood, the majority of the ancestry of a sample, and hence the backbone of a reconstructed genealogy, is highly random both in terms of its topology and branch length. Irwin (2002) used coalescent simulations to show that in species distributed along a linear habitat such as a shore line or a mountain range, deep phylogeographic breaks can arise at random locations without barriers to dispersal. This is particularly worriesome for traditional phylogeographic inference, which readily interprets such breaks in mitochondrial genealogies as evidence for past historical events or barriers to gene flow. In general, the basic insight of coalescent theory, that the same history can lead to very different genealogies and *vice versa*, implies that large numbers of genealogies are required to make robust inferences about population history. The statistical power of analysing a large number of loci is illustrated by the detailed inferences about human history that can be made from just a few complete genomes (Chen & Li, 2001; Yang, 2002; Rannala & Yang, 2003; Patterson *et al.*, 2006; Ebersberger *et al.*, 2007), most strikingly from the recent Neanderthal sequences (Green *et al.*, 2010).

1.4 Inference methods

Despite the advances in coalescent theory outlined above, deriving results which can be used to analyse phylogeographic data under realistic models of structure has been hampered in two ways. Firstly, formulating a model that captures the movement of individuals in continuous space in a way that is consistent forwards and backwards in time and ensures some form of density regulation (Felsenstein, 1975), has proven to be a major challenge (Barton & Wilson, 1995; Barton *et al.*, 2002; Wilkins, 2004), although progress has been made recently (Barton *et al.*, 2010). Secondly, even without a full description of geography, finding the joint distribution of coalescent times and topology is difficult simply because of the vast number of possible tree topologies even for moderate samples. The total number of possible coalescent histories (i.e. trees with time-ordered nodes) is given by the product over the number of possible coalescence events at each time interval, i.e. $\prod_{i=2}^{n} \binom{n}{2}$ (Wakeley, 2008) and thus grows much faster than exponentially with sample size. For instance, a sample of size 10, may have 2,571,912,000 possible histories. Coalescent results for models of population structure are therefore commonly restricted to samples of size two and even then, analysis can be challenging in particular for non-equlibrium models. For example, the full distribution of pairwise coalescence times under the non-equilibrium analog of the symmetric island model (i.e. a panmictic population which has become subdivided into a set of island-model demes at some recent time and not reached migration-drift equilibrium) has only been found recently (Wilkinson-Herbots, 2008). While numerical likelihood methods to estimate parameters under models of divergence from minimal samples exist (Yang, 2002; Wilkinson-Herbots, 2008; Wang & Hey, 2010), the integration over the large number of possible genealogies for larger samples is not tractable analytically (Felsenstein, 1988; Hey & Nielsen, 2007) and is generally achieved using approximate methods such as Markov chain Monte Carlo simulations (Kuhner *et al.*, 1995; Nielsen & Wakeley, 2001; Rannala & Yang, 2003; Hey & Nielsen, 2004), importance sampling (Griffiths & Tavaré, 1994), or summary statistics (Becquet & Przeworski, 2007; Hickerson *et al.*, 2007). While these approaches are powerful and have been successfully applied to make historical inferences in a wide range of organisms (e.g. Kliman *et al.*, 2000; Jennings & Edwards, 2005; Won *et al.*, 2005; Hickerson *et al.*, 2006; Becquet & Przeworski, 2007; Carstens *et al.*, 2009; Muster *et al.*, 2009; Hey, 2010a) including our own species (Rannala & Yang, 2003; Hey, 2005), they are often computationally intensive and the complexity of the algorithms involved makes it difficult to fathom which features of the data are informative about past processes. The situation is perhaps worst for approximate Bayesian methods (Beaumont *et al.*, 2002), which rely on summary statistics to compare the fit of observed data to simulations and, ultimately, to estimate the posterior distribution of model parameters. Because statistics are chosen empirically and arbitrary cut-offs are used both to decide which simulation replicates are informative about the fitted model, and to restrict priors, it can be very difficult to assess how much information about a particular model there is in the data.

These theoretical difficulties may in part explain the slow uptake of coalescent theory by phylogeography, despite its obvious implications for the interpretation of genealogies. Inference methods for the analysis of spatial samples under realistic models of structure simply do not exist and rejecting an unrealistic null model such as that of a panmictic Wright-Fisher population hardly yields much insight into population history. Moreover, the spatial processes phylogeography seeks to understand (e.g. range expansions and local extinctions) implicitly assume correlations across loci, something that is not captured by standard coalescent models. Perhaps an equally serious obstacle has been the practical difficulty of obtaining sequence data for multiple nuclear loci in most organisms. However, progress has been made in two ways. Firstly, the fact that genealogies differ from the population or species history, even if this is tree-like itself, has now been absorbed into phylogenetics and phylogeography (Pamilo & Nei, 1988; Maddison, 1997; Nichols, 2001; Edwards, 2009). Species trees include as an additional dimension the effective sizes of all populations involved. This crucial set of parameters, which determines the rate of coalescence of genealogies nested within the species tree and thus the probability of gene tree - species tree incongruence was missing from Avise's original figure (Avise, 1987, figure 1). By making simplifying assumptions about the ancestral N_es, it is possible to infer properties of the underlying species tree from a set of time-measured gene trees (Degnan & Salter, 1995; Degnan & Rosenberg, 2009; Maddison & Knowles, 2006; Liu & Pearl, 2007; Kubatko *et al.*, 2009) or from sequence data directly (Yang, 2002; Rannala & Yang, 2003; Heled & Drummond, 2009). Secondly, phylogeographers now routinely use coalescent simulations to assess the fit of observed genealogies to simple *a priori* models, such as contrasting models of population divergence (Knowles, 2001). The sobering conclusion of most statistical phylogeographic studies is that the power to distinguish even between very simple models is limited (e.g. Knowles, 2001; DeChaine & Martin, 2006). Finally, rigorous evaluation of the performance of nested-clade phylogeographic analysis (NCPA) using coalescent simulations demonstrated a high frequency of false positives (Knowles, 2002; Panchal & Beaumont, 2007; Knowles, 2008). This together with the realisation that NCPA lacks any quantitative basis (Knowles, 2002) has led to its abandonment by phylogeographers, despite Templeton's attempt to rebrand NCPA as a "coalescent-based method of statistical inference" (Templeton, 2010).

Nielsen and Beaumont (2009) point out a more subtle but equally serious problem with NCPA: The inference key often suggest multiple rather vague answers. Faced with this subjective choice, researchers tend to inadvertently settle on scenarios that match their prior expectations thereby over-interpreting the data. This phenomenon, which is known in psychology as the Forer effect (Forer, 1949), may explain why some studies have found such striking congruence between quantitative inferences made using coalescent methods and results obtained from NCPA (Sunnucks *et al.*, 2006; Nielsen & Beaumont, 2009). However, the use of quantitative inference methods *per se* by no means safeguards against the self-delusion *a la* Forer. This is illustrated by a recent study by Tanabe *et al.* (2010) which uses approximate Bayesian Computation to investigate the demographic history of a set of seven populations of malaria parasite (*Plasmodium falciparum*) sampled from Africa and Asia. Tanabe *et al.* (2010) simulate expansion histories under a one-dimensional stepping stone model and use θ_π, the average pairwise diversity within populations, as a summary statistic to estimate the origin and timing of the expansion. They assume a uniform prior (with bounds set at 1,000 and 100,000 years) for the onset of this expansion and test the effect of three different mutation rates (assuming a divergence time between *P. falciparum* and its closest relative the chimpanzee malaria parasite *P. reichenowi* of 10,000 years, 2.5 MY and 6 MY). Finding that the highest mutation rate leads to a very poor fit to the data, the authors conclude that " *P. falciparum* had already infected humans before the out-of-Africa expansion." However, in reality there is no information to separately estimate mutation rate and expansion time in these data and the poor fit to the high mutation rate scenario can be entirely explained from the lower prior bound chosen for the expansion time. Thus, apart from choosing to exclusively focus on a one-dimensional stepping stone model, subjective choices are made at various steps in the analysis including the summary statistic, the cut-offs on the (supposedly uninformative) priors of model parameters, the acceptance criterion and the three particular mutation rates investigated. While some of these choices may be justifiable given prior knowledge of the system, others are clearly arbitrary. Thus, any model based analysis faces the difficulty of deciding on a set of plausible models (Carstens *et al.*, 2009), which are simple enough to be distinguishable using the data at hand, but nevertheless capture relevant aspects of the underlying history.

The above overview is necessarily incomplete and omits many of the historical twist and turns in the development of spatial and demographic inference methods. Such more arcane history-of-the-field reasons for the popularity of particular methods can be surprisingly long-lived despite their arbitrariness and may be solely driven by the availability of bioinformatics software. For example, phylogeographic

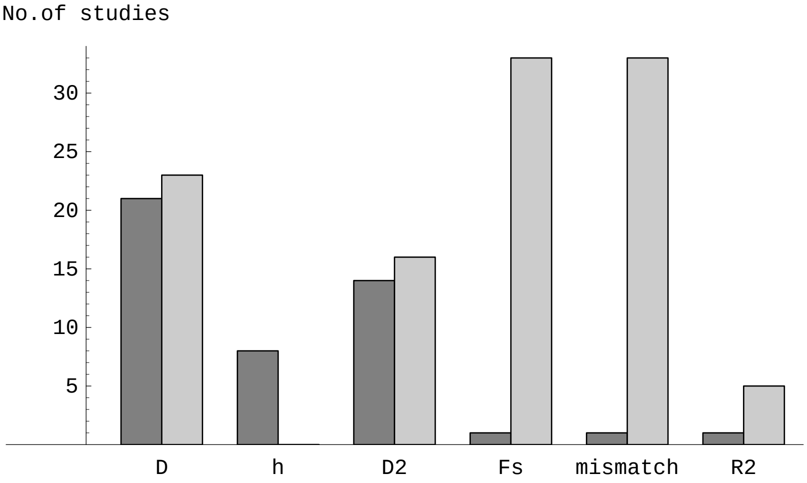

Figure 1.1: A literature search on Web Of Science returned 68 studies that use summary statistics for demographic inference. The search criterion was any pairwise combination from the following sets of key words: i) Demography, demographic history/inference, population growth/expansion and ii) summary statistics, neutrality tests. Studies were classed as population genetic (dark grey bars) or phylogeographic (light grey bars) depending on whether they featured a reconstructed genealogy and the summary statistics used for demographic inference were recorded: From left to right, Tajima's D (Tajima, 1989), H (Fay & Wu, 2000), D_2 (Fu & Li, 1993), F_S (Fu, 1996) and the mismatch distribution (Slatkin & Hudson, 1991). Population genetics and phylogeographic studies differ both with regard to the sampling scheme and statistics used: Population genetics studies (22), mainly on *Drosophila* and human populations, typically have relatively small sample sizes (N=10-30), but analyse data from multiple loci and use Tajima's D and D_2 to evaluate demographic models. Phylogeographic studies (46), are mostly based on a single mitochondrial gene sampled for a large number of individuals and most frequently (34) use "visual inspection of mismatch distributions" or Fu's F_S to assess population growth.

and population genetic studies generally use different summary statistics to test for population growth. While population geneticists prefer statistics that considering properties of the underlying genealogy such as Fu & Li's D (Fu & Li, 1993), phylogeographers are fond of pairwise mismatch distributions (Fig. 1.1). This is ironic not only because summary statistics based on pairwise measures are among the least powerful, but also because they most fundamentally ignore the underlying genealogy (Felsenstein, 1992), a fact which is discussed at length in the original paper introducing mismatch distributions (Slatkin & Hudson, 1991). The only explanation for this odd preference is that mismatch distributions were first applied to mitochondrial data (Slatkin & Hudson, 1991; Harpending, 1994; Schneider & Excoffier, 1999) and that convenient software is available to perform simulations (Schneider *et al.*, 2000). Thus, while coalescent theory has achieved much in terms of integrating phylogeographic and population genetics approaches to historical inference, this synthesis is clearly far from complete.

1.5 Thesis outline and aims

This thesis considers the use of genealogies for historical inference from a variety of angles and applies model-based phylogeographic analysis to two contrasting insect systems: a local radiation of dispersal-limited, high alpine ground beetles (Carabidae, genus *Trechus*) sampled from a single mountain range in the Southern Alps; and three major refugial populations of the highly dispersive wasp *Cecidostiba fungosa* (Pteromalidae) parasitizing oak galls.

Chapter 2 focuses on the simplest case of a panmictic population and asks how past population growth, which distorts genealogies towards a starshape can be inferred most efficiently from sequence data using summary statistics. Felsenstein (1992) pointed out that pairwise measures, which underly many commonly used neutrality tests such as Tajima's D (Tajima, 1989), are inefficient because of their inherent sensitivity to the topology of the underlying genealogy, which in a panmictic population is entirely random. The challenge therefore is to construct summaries that separate effects of the topology from relevant branch length information. Using coalescent simulations under a history of exponential growth, the power of standard summary statistics is compared to that of two types of new measures which are derived by explicitly considering the underlying genealogy: i) genealogical ratios based on the number of mutations on the rootward branches, which, given an outgroup sequence can be inferred using a simple algorithm; and ii) statistics that use properties of a perfectly starshaped genealogy. A likelihood-based method (Griffiths & Tavaré, 1994) is taken as an upper bound of statistical power for comparison.

Chapter 3 is in many ways a traditional phylogeographic study. Twelve populations of high alpine carabid beetles (genus *Trechus*) were sampled from the Orobian Alps in Northern Italy. While summits along the northern ridge of this mountain range were surrounded by the icesheet as small ice-free islands of habitat, so-called sky-islands or nunataks during the last glacial maximum, southern areas remained unglaciated. The aim was to consider how mitochondrial (*Cox1* and *Cox2*) and nuclear (*PEPCK*) sequence data can be used to infer the spatial history of this local radiation. Rather than drawing qualitative inferences from the reconstructed genealogies of the two loci, the fit to two simple *a priori* models of population history is assessed: prolonged survival of Northern populations *in situ*; and recent recolonisation from Southern populations. Extreme versions of these scenarios make alternative predictions about the topology of genealogies. While isolation eventually leads to reciprocal monophyly of populations, a series of extreme founder events results in a pattern of nested paraphyly, which is informative about the order of population founding. Bayesian inference methods are used in two ways. Firstly, directional location state changes in the genealogy are modeled to find the most likely sequence of putative founder events under the recolonisation model. Secondly, the fit of the data to the two scenarios is quantified by testing the expected mono and paraphyly constraints. Because location states are inferred jointly with the genealogy and mutational parameters, the analysis takes genealogical uncertainty into account. It also allows us to assess the contribution of incomplete lineage sorting and migration.

Realizing the power of jointly analysing data from a large number of loci in a model-based framework motivated the development of intron-spanning primers for Hymenoptera (where intronic regions can be sequenced straightforwardly in haploid males). Chapter 4 describes how 40 conserved genes, mainly ribosomal proteins, were chosen from Hymenopteran and insect EST data to develop degenerate primers. The aim was to find loci suitable for comparative multispecies studies of natural Hymenopteran communities, i.e. which amplify across a wide taxonomic range. Amplification success was assessed in two communities; gall wasps (Cynipidae) and their associated Chalcid parasitoids; and tropical fig wasps (Aagonidae) and their associated non-pollinating wasps (Pteromalidea). Taxa were chosen at increasing distance from *Nasonia* which was used for primer design, i) within Pteromalidae, ii) within Chalcidoidae (Eupelmidae, Eulophidae, Eurytomidae, Ormyridae, Torymidae), and iii) for a selection of distantly related gall and fig wasp hosts. To assess the usefulness of these loci for phylogeographic studies, genetic diversity between major Palearctic refugia was estimated for two species of oak gall parasitoids; *C. fungosa* and *Mesopolobus amaenus* (Pteromalidae).

In chapter 5, 20 of the new EPIC loci are used to quantify the Pleistocene history of the oak gall parasitoid *C. fungosa*. The longitudinal spread of temperate organisms into refugial populations in Southern Europe is generally assumed to predate the last interglacial. However few studies have attempted to quantify this process using explicit models and multilocus data. Maximum likelihood and Bayesian methods methods originally developed to quantify species divergence are used to infer the order of population splitting and estimate divergence times and ancestral population sizes for three major refugial populations (Middle East, the Balkans and Iberia). To determine how quantitative inferences can be made most efficiently from multilocus data, the power of minimal sampling (a single haploid male per population) is compared with that of more extensive samples of three individuals per population.

The fundamental symmetry in the two incongruent histories under the three population divergence model translates into symmetries in the expected frequency of site counts which can be easily tested in genome wide alignments. Chapter 6 extends the three population divergence model analytically to include gene flow involving the older population. Slatkin & Pollack (2008) showed previously that ancestral population structure in divergence models can lead to asymmetries in the frequency of triplet topologies. Using an analogous matrix approach, the probabilities of triplet topologies are derived for the case of symmetric and asymmetric migration. Potential applications of these results for the analysis of genomic data from *Drososphila melanogaster* (Obbard *et al.*, 2009) and a recent study on Neanderthal-human divergence (Green *et al.*, 2010) using this model are discussed.

Chapter 2

Measuring the degree of starshape in genealogies — Summary statistics and demographic inference

Published as: Lohse, K. Kelleher, J. (2009).
Genetics Research, **91** (281–292)

The motivation for studying the impact of past demography on sequence data is two-fold. Firstly, changes in population size are interesting in their own right, being intimately linked to processes such as speciation or geographic range shifts. Secondly, the standard neutral model (SNM) of a randomly mating Wright-Fisher population of constant size and discrete generations, hardly ever describes the patterns of diversity found in natural populations. Thus, studies aiming to detect loci under selection are faced with the considerable challenge of fitting realistic demographic models against which selection can be tested (e.g. Glinka *et al.*, 2003; Hamblin *et al.*, 2004; Haddrill *et al.*, 2005; Ometto *et al.*, 2005; Thornton & Andolfatto, 2006). Since the rate of coalescence is inversely proportional to the effective population size, it is clear that demographic changes must leave a detectable signature in genealogies (Felsenstein, 1992). In general, positive population growth distorts genealogies towards a starshape with shorter internal branches, resulting in more low frequency variants and a unimodal rather than multi-peaked mismatch distribution (Slatkin & Hudson, 1991; Harpending, 1994; Schneider & Excoffier, 1999). In contrast to selective processes which act on single genetic variants, demography affects the whole genome, so one expects to find a concordant signature across loci (Tajima, 1989; Galtier *et al.*, 2000).

Approaches to demographic inference fall into three broad categories; (for a review see Emerson *et al.*, 2001). Firstly, likelihood methods, which are available for bottleneck and exponential growth models, make use of all the information in a sample by integrating over a large set of likely genealogies (Griffiths & Tavaré, 1994; Kuhner *et al.*, 1995). Although optimal in terms of statistical power and accuracy, likelihood estimation is computationally intensive and requires a fully specified alternative model. Therefore realistic growth histories often remain analytically intractable. Secondly, there are tree-based methods, which take the branch length information of a reconstructed tree as their starting point. Assuming that sequence evolution is clock-like, the number of lineages can be plotted against time and the shape of this trajectory compared to its neutral expectation (Nee *et al.*, 1995; Pybus *et al.*, 2002). Despite their conceptual appeal, these methods neglect any uncertainty in tree topology and are thus only as good as the reconstructed tree they are based on. Furthermore they cannot deal with recombination by definition. Finally, there are classical neutrality tests, most of which do not explicitly consider the genealogy but instead use more immediate aspects of the data such as the frequency spectrum of mutations, e.g. Tajima's D (Tajima, 1989) and Fu $\&$ Li's D (hereafter referred to as D_2) (Fu & Li, 1993), the haplotype distribution, e.g. Fu's F_S (Fu, 1996; Innan *et al.*, 2005), or the mismatch distribution, e.g. the raggedness statistic (Slatkin & Hudson, 1991). Compared to likelihood estimation, summary statistics are straightforward to calculate and their distribution can be simulated under almost any growth model.

Considering the zoo of statistics available and their wide use, there are surprisingly few studies that systematically compare their power, and those that do mainly consider bottlenecks and single locus data (Simonsen *et al.*, 1995; Fu, 1996; Ramos-Onsins & Rozas, 2002; Depaulis *et al.*, 2003; Ramirez-Soriano *et al.*, 2008). However, joint analysis of multiple loci is not only necessary to distinguish between selective and demographic events (Galtier *et al.*, 2000) but also potentially far more powerful than inferences based on a single locus. An added advantage of multi-locus analysis is that both means and variances of summary statistics can be used for testing. Variance based tests were first developed for microsatellite data (Di Rienzo *et al.*, 1998; Reich *et al.*, 1999) but are now routinely used to analyse sequence data from multiple loci (Pluzhnikov *et al.*, 2002; Haddrill *et al.*, 2005; Heuertz *et al.*, 2006) or even species (Hickerson *et al.*, 2006).

A general conclusion that has emerged from simulation studies is that tests based on the number and distribution of haplotypes have more power to detect bottlenecks than statistics based on the average pairwise diversity (π), in particular Tajima's D (Ramos-Onsins & Rozas, 2002; Innan *et al.*, 2005; Ramirez-Soriano *et al.*, 2008). Earlier, Felsenstein made a theoretical argument for the inferiority of pairwise measures (Felsenstein, 1992). Their large variance under neutrality arises both from their sensitivity to the last coalescence event and the random genealogical topology (Tajima, 1983). Under the SNM more symmetric genealogies are on average associated with higher π and more ragged mismatch distributions than asymmetric genealogies. It is important to realise that this topological variance is independent of the already large variance in coalescence times inherent in the genealogical process. In other words "despite their aura of robustness" (Felsenstein, 1992), statistics based on π suffer from an unnecessarily large variance under neutrality, and hence have comparatively low power. Despite these results, D and mismatch distributions continue to be the methods of choice for demographic inferences in population genetics and phylogeography respectively.

Following Felsenstein's recommendation that "there is much to gain from explicitly taking the genealogical relationship of a sample into account" (Felsenstein, 1992), the aim of this study is to consider how genealogical information can be used for demographic inference in a summary statistics framework. Our premise here is that the mutation rate is sufficiently high relative to the per site recombination rate such that non-recombining blocks of sequences can be easily identified and treated as independent loci.

Given that there is usually not enough information in within-species sequence data to infer the full topology unambiguously it seems important to ask which part of the topology yields most information. The first part of the paper introduces some simple measures of starshape which are based on the properties of a rooted genealogy. Using simulations their power to detect a history of exponential growth is compared to standard neutrality tests for both the single and multi-locus case. We focus on the exponential growth model for two reasons. Firstly, although it is a frequently used demographic model, the power of summary statistics to detect exponential growth has been little investigated. Secondly, likelihood methods are available, which can be taken as an absolute "upper bound" of power for comparison. Such a direct comparison between summary statistics and the optimal likelihood methods is lacking so far.

2.1 Summary Statistics

Several neutrality tests compare two different estimators of the scaled mutation rate (Tajima, 1989; Fu & Li, 1993; Fay & Wu, 2000) $\theta = 4N_e\mu$, where μ is the mutation rate and N_e the effective population size, which capture different aspects of the data . Most prominently, Tajima's D is defined as the difference between θ estimated as π, and $\theta_w = S/a_n$ (Watterson's θ, where $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$, n is the sample size and S the total number of polymorphic sites in the sample), normalized by the standard deviation of this difference. Genealogies from growing populations typically have relatively more low frequency variants and hence tend to have a negative D.

While neutrality tests are commonly based on the frequency spectrum and π, it is instructive to consider departures from the SNM in terms of their effect on the genealogy. Such tree-thinking necessarily underlies summaries that make use of outgroup information, e.g. D_2 has a straightforward genealogical interpretation. Below two different ways of employing genealogical information in the construction of summary statistics are considered.

2.1.1 Genealogical ratios

The rationale behind D_2 is to distinguish between two classes of mutations: Those found on terminal branches, η_e and those on internal branches, η_i (Fig. 2.1) (Fu & Li, 1993). Suppose that some limited topological information can be inferred from the data. In particular we will for now assume that the placement of the root is known. It is then possible to distinguish mutations found on the two rootward branches, which we shall denote η_R. Under the infinite sites assumption these are all derived mutations which are shared by all individuals in either of the two sub-clades defined by the root. The advantage of considering the proximity of mutations to the root rather than the tips is twofold: Firstly, rootward branches cover a greater proportion of the time to the most recent common ancestor of the sample (T_{MRCA}) and should in general be more informative about past changes in population size. Under the SNM, on average half of the T_{MRCA} is taken up by the coalescence of the last two lineages (T_2) (Fig. 2.1), whereas in a growing population, the smaller population size in the past forces the last two lineages to coalescence much more rapidly. Secondly, the average length of a branch connected to the root is less dependent on the sample size than the average length of a terminal branch.

Ideally, one wants to know the total number of mutations that have occurred during T_2, rather than the number of mutations on both rootward branches, η_R which is larger and depends on the topology, i.e. the order of the first node on the longer of the two branches (Uyenoyama, 1997, Appendix).

One possibility is to only consider the shorter of the two rootward branches which has exactly length T_2. Thus the number of mutations found on this branch, η_{Rmin}, over θ_w constitutes a very simple measure of starshape.

$$X = \frac{\eta_{Rmin}}{\theta_w} \tag{2.1}$$

Such genealogical ratios have first been employed to study the effect of balancing selection on plant incompatibility loci (Uyenoyama, 1997). Being based on a single random event, X clearly neglects much of the information contained in the genealogy. Its power is limited by the probability of observing $\eta_{Rmin} = 0$ under neutrality. In other words, X is unlikely to be of much use in the case of a single locus.

Alternatively, one can ignore the uncertainty in node order and take the number of mutations found

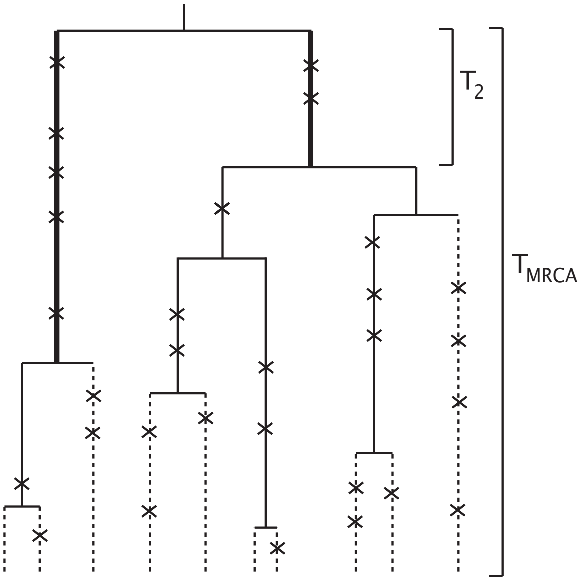

Figure 2.1: Random genealogy of a sample of 10 sequences. The root partitions the sample into two subclades of size 3 and 7. Rootward branches are shown as bold, terminal branches as dotted lines, mutations are represented as crosses. The time interval until the last coalescence event, T_2, is shorter than average under the SNM. In this example $S = 30$, $\eta_R = 7$, $\eta_{Rmin} = 2$ and $\eta_e = 14$.

on both rootward branches relative to θ_w.

$$X_1 = \frac{\eta_R}{\theta_w} \tag{2.2}$$

It is possible to construct various composite measures from the number of mutations found on different parts of the genealogy. Here we only consider one additional statistic, the relative difference between rootward and terminal mutations.

$$X_2 = \frac{\eta_R - \eta_e}{\theta_w} \tag{2.3}$$

The X statistics assume some knowledge of the tree topology which is usually unknown. Of course one could use some standard method of tree reconstruction and infer η_R and η_{Rmin} from the most likely topology. However, not only is it inefficient to reconstruct the full topology when all that is required is the placement of the root, conditioning on a single tree also ignores any topological uncertainty. We have therefore developed a simple scheme of inferring the root in a sample of polarized sequences that circumvents these problems.

Under the infinite sites assumption a necessary criterion for the root-node is that no mutations are shared between the two subsets on either side. One can show that if both branches connected to the root carry mutations, i.e. $\eta_{Rmin} > 0$ there exists exactly one bipartition of the sample with no mutational overlap. If however one or both of the rootward branches of the genealogy carry no mutations there may be multiple bipartitions that meet this criterion. In this case $\eta_{Rmin} = 0$ and the tree reconstructed from such a sample would have an unresolved polytomy at its base. To incorporate the topological uncertainty about the placement of the root we compute the average value of η_R over all partitions that are compatible with the criterion of no mutational overlap. Note that in contrast to most tree reconstruction algorithms which join similar sequences (i.e. start from the tips down the tree) our scheme is divisive (i.e. it starts from the root). To avoid having to considering all possible bipartitions of the sample ($2^{n-1} - 1$), we make use of the fact that any sequences that share mutations have to be on the same side of the root. By first binning sequences that share at least one mutation, we can directly calculate η_R and the number of possible partitions.

2.1.2 Starting from the limiting case

A different approach is to construct summaries which measure departures from the limiting case of a perfectly star-shaped genealogy. Star-shaped genealogies have some convenient properties which can be used for this. Assuming that outgroup information is available, one can record the number of terminal mutations in each sequence i (because lineages are exchangeable, the labeling is arbitrary), V_i. In a perfectly star-shaped genealogy all mutations must fall onto terminal branches by definition. Thus one expects the number of derived mutations in a sequence to be half the average pairwise diversity, i.e. $E[V_i] = \pi/2$. The statistic R_{2E} proposed by Ramos-Onsins and Rozas measures the average departure from this expectation.

$$R_{2E} = \frac{(\sum_{i=1}^{n} (V_i - \frac{\pi}{2})^2/n)^{1/2}}{S} \tag{2.4}$$

(Ramos-Onsins & Rozas, 2002, eq. 2). R_{2E} has proven superior to a wide range of summary statistics in detecting histories of bottlenecks (Ramos-Onsins & Rozas, 2002). However, because of its dependence on π, one may suspect it to suffer from a large variance under neutrality. We therefore consider a similar statistic which uses the observed S rather than π to assess the degree of starshape. Consider the total number of derived mutations in each sequence, D_i. Given that previous summary statistics such as H (Fay & Wu, 2000) have been derived from the unfolded site frequency spectrum, it may be helpful to note the connection between D_i and ξ_i, the number of derived mutations that occur i times in the sample here: $\sum_{i=1}^{n} D_i = \sum_{i=1}^{n-1} i\xi_i$. Using the fact that $E[D_i] = S/n$ in a star-shaped genealogy we can define a new statistic.

$$R_S = \frac{(\sum_{i=1}^{n} (D_i - \frac{S}{n})^2/n)^{1/2}}{S} \tag{2.5}$$

Since under neutrality a large proportion of mutations will be found on inner branches, i.e. be shared by many sequences, $E[D_i] > S/n$. In other words, R_S is such that smaller values are expected under a history of growth.

2.2 Methods

2.2.1 Summary statistics and demographic model

We carried out coalescent simulations in `ms` (Hudson, 2002) to compare the power of a range of summary statistics to distinguish between the SNM and a history of exponential growth. In addition to D, D_2, R_{2E} and the new statistics defined above, F_S, (Fu, 1996) and H (Fay & Wu, 2000) were considered. F_S is based on the number of haplotypes in the sample and has previously been found to be more powerful than statistics based on the frequency distribution (Fu, 1996; Ramos-Onsins & Rozas, 2002). H was conceived as a test for the effect of selection on linked neutral sites (Fay & Wu, 2000) and is not expected to have power to detect continuous growth. However, other demographic scenarios such as moderate bottlenecks may perturb genealogies in ways similar to genetic hitchhiking resulting in significant values of H.

We assume that the population size has grown exponentially with rate α to its current size N_0.

$$N(t) = N_0 e^{-\alpha t} \tag{2.6}$$

Following standard practice, this exponential growth is incorporated through a re-scaling of time (Slatkin & Hudson, 1991). We define a rescaled time T_{coal} relative to N_0 and α:

$$T_{coal} = \int_0^t \frac{e^{\alpha t}}{2N_0} dt = \frac{(e^{\alpha t} - 1)}{2N_0 \alpha} \tag{2.7}$$

This represents the total amount of genetic drift that has occurred. It is convenient to define a growth rate relative to N_0 as $A = 2N_0\alpha$, which gives:

$$T_{coal} = \frac{e^{\frac{At}{2N_0}} - 1}{A} \tag{2.8}$$

2.2.2 Power test

Critical values corresponding to $p = 5\%$ for each statistic were determined from 10 000 replicate genealogies simulated under the SNM for each of a wide range of S values (1-250) (Hudson, 1993; Braverman *et al.*, 1995; Ramos-Onsins *et al.*, 2007). Genealogies from growing populations were simulated conditional on θ. For each replicate the alternative hypothesis of positive growth was tested by comparing the observed value of a statistic to the critical value given the observed S. Power was estimated as the proportion of 10 000 replicate genealogies for which a statistic was below its critical value in a one-tailed test. Power to reject the SNM was recorded for a large range of parameter combinations. We compared the performance of statistics for different growth rates, ($0 < A < 50$), sample sizes ($n = 10, 50$) and values of θ (5-50). When varying θ, we chose a fixed value of $A = 8$. This seems compatible with growth rates estimated from empirical data. For example, variation at silent sites in the *Adh* region and X-linked genes in *D. pseudoobscura* is consistent with $A = 7$ (Schaeffer, 2002). While θ can be arbitrarily high for mitochondrial data, $\theta = 20$ may be unrealistic for nuclear loci in out-crossing species. Therefore, power was evaluated for a range of θ values ($5 - 50$) again keeping the growth rate fixed at $A = 8$.

When using means and variances of summary statistics across loci, power was determined analogously to the single locus case. Critical values of 5% confidence of means and variances of statistics were determined from 10 000 replicate sets of loci with the exact same combination of S values. Although computationally expensive, this avoids making any assumptions about the distribution of mutation rates between loci. However, given that mutation rates vary along the genome, assuming the same θ for all loci to simulate the alternative history of growth seems unrealistic and may lead to overestimation of power. We checked for the influence of heterogeneity in mutation rates on power by repeating the multilocus power tests with θ values drawn from a gamma distribution with $\alpha = 2$ (Pluzhnikov *et al.*, 2002) and a scale parameter equivalent to a mean of $\theta = 20$. This combination of growth and mutation rates is roughly comparable to mutation rate estimates for nuclear loci in *Drosophila melanogaster* (Galtier *et al.*, 2000). As before we assumed no recombination within loci as well as absence of linkage between loci, i.e. replicate genealogies were simply treated as multiple loci.

2.2.3 Likelihood method

In practice, both θ and A are unknown, and their likelihood should, in principle, be estimated jointly. However, because of the non-independence of these two parameters, this is not a practical option. Following standard practice we alternated between maximum likelihood estimation of A and θ (Griffiths & Tavaré, 1994). First a maximum likelihood estimate (MLE) for θ under the SNM was estimated using the program GENETREE (http://www.stats.ox.ac.uk/griff/software.htm). In a second step this MLE for θ was fixed to run a likelihood surface for A. Finally, the MLE value for A was used to re-evaluate θ. This scheme yields two MLEs for θ for each replicate, one under the assumption of no growth and one given the most likely growth rate which were compared in a likelihood ratio test (LRT). We did not find that the MLE estimates for A and θ improved upon repeated reevaluation suggesting that a single round of estimation is sufficient for this moderate growth scenario. 100 000 runs were performed for each likelihood surface evaluation. Again, the proportion of replicate genealogies for which the null hypothesis could be rejected was taken as a measure of statistical power. Due to the long computing time, 100 replicates per parameter combination were used.

2.3 Results

2.3.1 Single locus

In general, both the likelihood method and summary statistics have low power to detect a history of moderate ($A < 8$) exponential growth for $n = 10$ (Fig. 2.2). As expected, the likelihood method is most powerful overall, although its superiority is surprisingly small. For example, based on the LRT the SNM is rejected for 30% of genealogies simulated under exponential growth of $A = 4$. In comparison, R_S and R_{2E} detect this history of growth in 23% of cases (Fig. 2.2).

Consistent with previous results, F_S, R_{2E}, and the new measure R_S, are considerably more powerful than both D and D_2 (Ramos-Onsins & Rozas, 2002; Ramirez-Soriano *et al.*, 2008). For $\theta = 20$, F_S is the most powerful statistic. The new measure R_S has consistently higher power than R_{2E}. As expected, H and X have no power to distinguish between the SNM and the growth case (not shown). However, the other two genealogical ratios perform surprisingly well. X_1 has higher power than D_2 and the power of X_2 is between that of R_{2E} and R_S (Fig. 2.2). The complete lack of power of D for $n = 10$ is somewhat surprising. Comparison with the result for $n = 50$ (Fig. 2.3) reveals that its performance is strongly dependent on sample size. We ran additional simulations (not shown) and found that for $n < 15$ extremely negative values of D are more likely under neutrality than under growth resulting in a rejection rate of the SNM of less than 5%. In other words, when n is small, the variance of D under neutrality is too large to detect exponential growth.

In general, all statistics have considerably higher power for $n = 50$ (Fig. 2.3). Interestingly, it never reaches 100% even when growth is extreme ($A = 50$). However, the relative effect of the sample size on power differs between statistics. For instance, X_1 improves relatively little in comparison to other measures. This is to be expected given that even small samples are likely to include the deepest split in the genealogy of the whole population (Saunders *et al.*, 1984). For $n = 10$, the power of all statistics decreases for histories of extreme growth ($A > 25$) (Fig. 2.2). This is due to the overall shortening of genealogies under rapid growth.

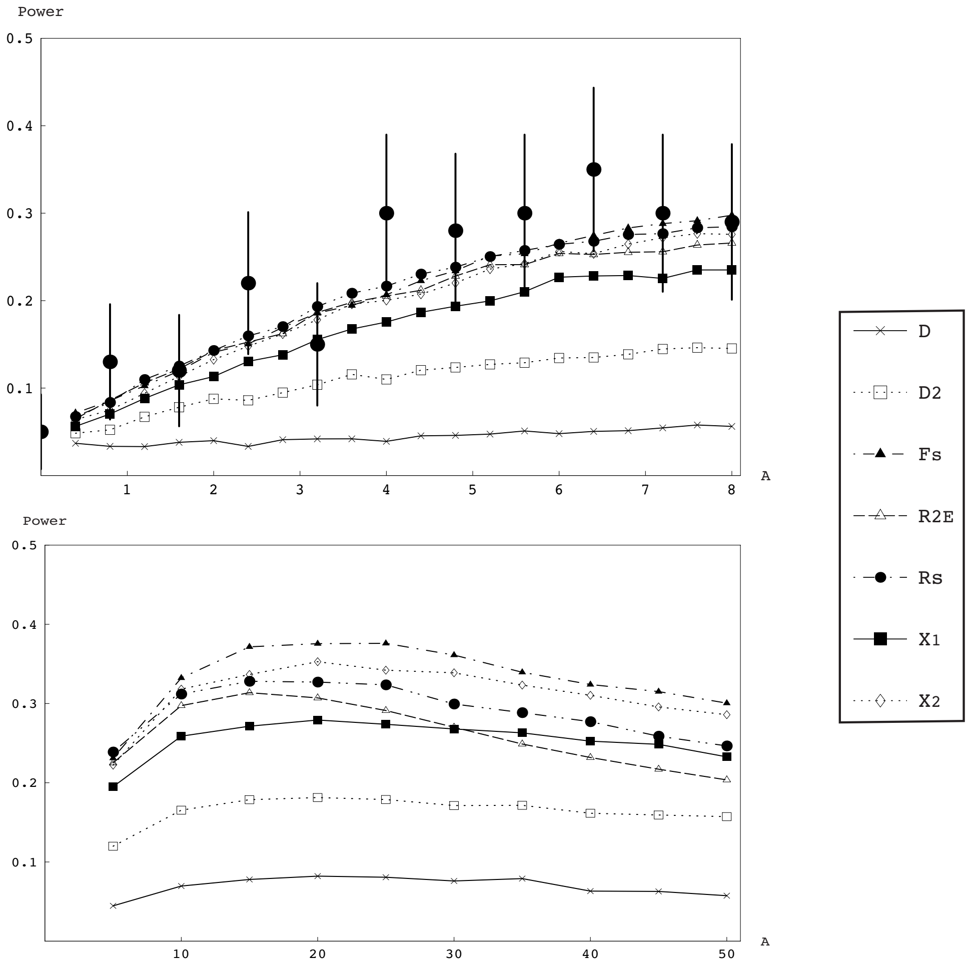

Figure 2.2: Power of summary statistics and likelihood method against exponential growth rate $A = 0 - 50$. $n = 10$, $\theta = 20$. Each point is based on 10000 replicate simulations. The power of the likelihood method was estimated from 100 replicates (see large filled circles and error bars)

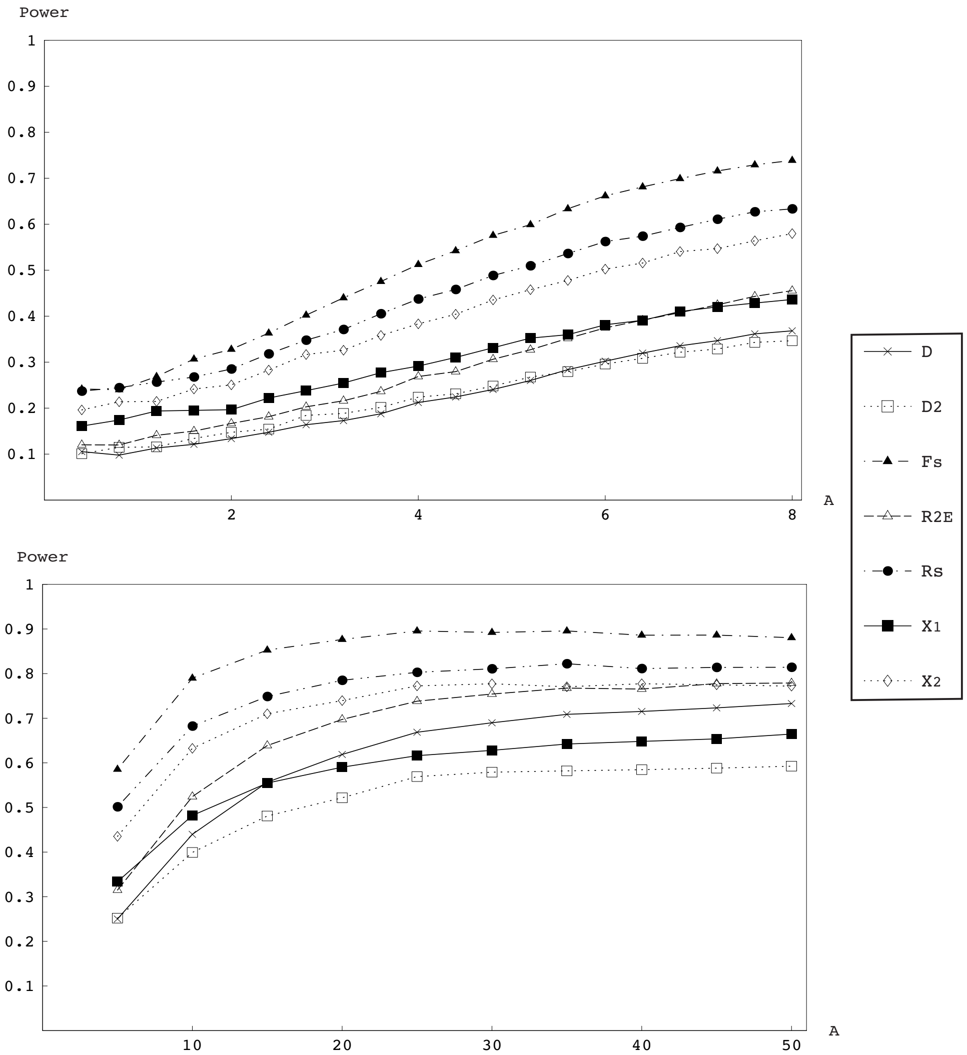

Figure 2.3: Power of summary statistics against exponential growth rate $A = 0 - 50$. $n = 50$, $\theta = 20$. Note the different range (0-1) on the y-axis compared to figure 2.2.

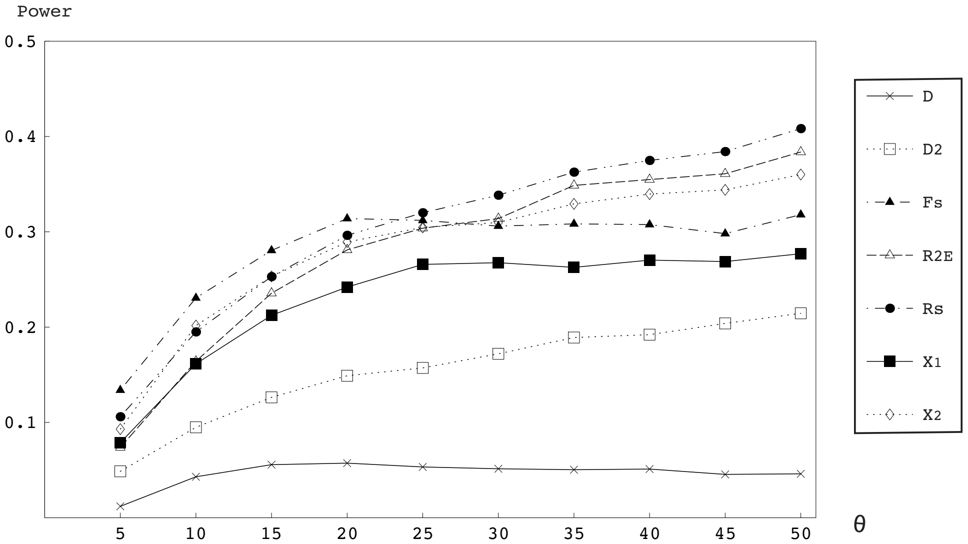

Figure 2.4: Power of summary statistics to detect a history of exponential growth ($A = 8$) against θ. $n = 10$.

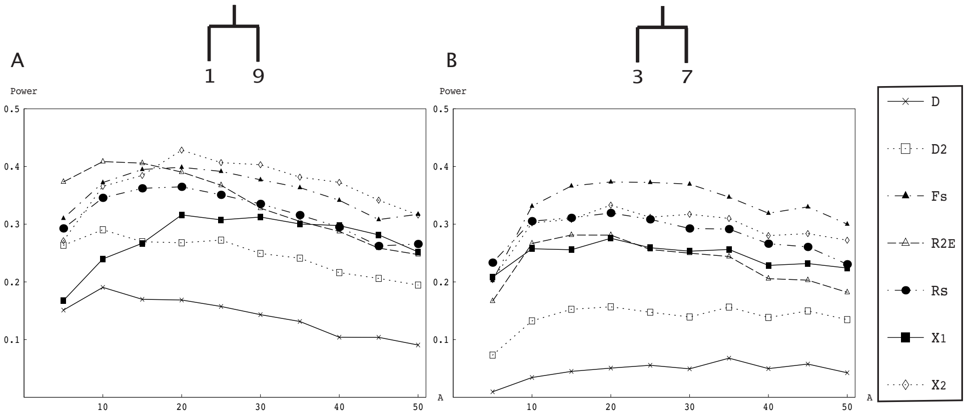

Figure 2.5: The effect of topological asymmetry on statistical power (simulation parameters as in 2.2). Genealogies of Fig. 2.2 were sorted according to the partition by the root (shown above plot). Only the most asymmetrical partition (9, 1) (a) and one other case (7, 3) (b) are shown. Results for the other three partitions were very similar to (B). Note that since lineages are exchangeable all asymmetrical partitions have the same probability $P_a = 2/(n-1)$ (Tajima, 1983, eq. 2).

The mutation rate has a relatively small influence on power. In general the power of all measures increases with θ (Fig. 2.4). However, the trajectories of X_1 and F_S level off while the power of the other statistics continues to improve with increasing values of θ. The power of F_S is limited by the number of haplotypes (which cannot exceed n).

To check how statistics are affected by the topological variance, genealogies simulated under the alternative history of growth were sorted according to the bipartition by the root and the proportion of significant values determined for each topology class. Figure 2.5 clearly shows that the two statistics based on π, D and R_{2E} as well as D_2 are sensitive to asymmetric topologies. The chance of observing a significant value increases markedly with topological asymmetry. This effect is most pronounced for D, which has no "power" to reject the SNM unless genealogies are very asymmetric and growth is weak. In contrast, the dependency of X_1 on the rootward partition is relatively slight and in the opposite direction, i.e. the chance of rejecting the SNM is smaller for asymmetric genealogies (Fig. 2.5).

2.3.2 Multiple loci

Compared to the relatively subtle effect both θ and n have on statistical power, increasing the number of loci improves power dramatically. In the mean based test all statistics apart from D have a power of close to 100% to detect a history of moderate exponential growth ($A = 8$) for 10 loci. However, the relative performance of statistics changes slightly compared to the single locus case. Notably, X_2 has higher power than all other summary statistics (Fig. 2.6). The power of X is slightly lower than that of

X_1 (not shown). Analogously to the results for a single locus, power increases both with more extreme growth scenarios and larger n (not shown).

As one may suspect, the increase in power with the number of loci is weaker for the variance test. More importantly, the relative performance of statistics is very different. By far the most powerful statistic in the variance test is X_1 followed by D and X (Fig. 2.7). This indicates a general trade-off. Statistics with a high variance under the SNM have comparatively low power in the single-locus case and the mean test, but high power in the variance test and *vice versa*.

Allowing for heterogeneity in mutation rates between loci affects both the relative performance of summary statistics and their overall power. As one may expect, heterogeneity in θ generally results in a decrease in power. In the mean based test the three X statistics are most affected. However, in the variance test the performance of X_1 is little affected. This statistic even has slightly higher power when mutation rates vary between loci. This appears to be due to the non-normal distribution of X_1 under growth. Genealogies with more than one possible root-partition generally have a very low value of X_1, since we take an average over all possible partitions most of which will be associated with $X_1 = 0$.

2.4 Discussion

It is important to distinguish between the general limitations that genealogical and mutational stochasticity impose on demographic inference from genetic data and problems associated with particular methods. Two main conclusions emerge from comparing the performance of the new "genealogical statistics" to classical neutrality tests and the LRT.

2.4.1 General limits to demographic inference

The signatures that changes in population size leave in genealogies are typically subtle compared to the randomness of the ancestral process. Thus all methods have low power to distinguish between the SNM and histories of moderate growth in the single locus case. A surprising finding of this study was that the full likelihood method only works marginally better than the most powerful summary statistics. Changes in N_e disproportionally affect the length of the basal branches of a genealogy. However, because these rootward branches also contribute most to the variance in total tree length, inferences based on a single locus will be weak at best. It is telling that the X statistics which only consider the last coalescence events in the history, outperform standard neutrality tests in the variance test when multiple realisations of this event, i.e. loci, are available. As has been argued before, most statistical power can be gained by

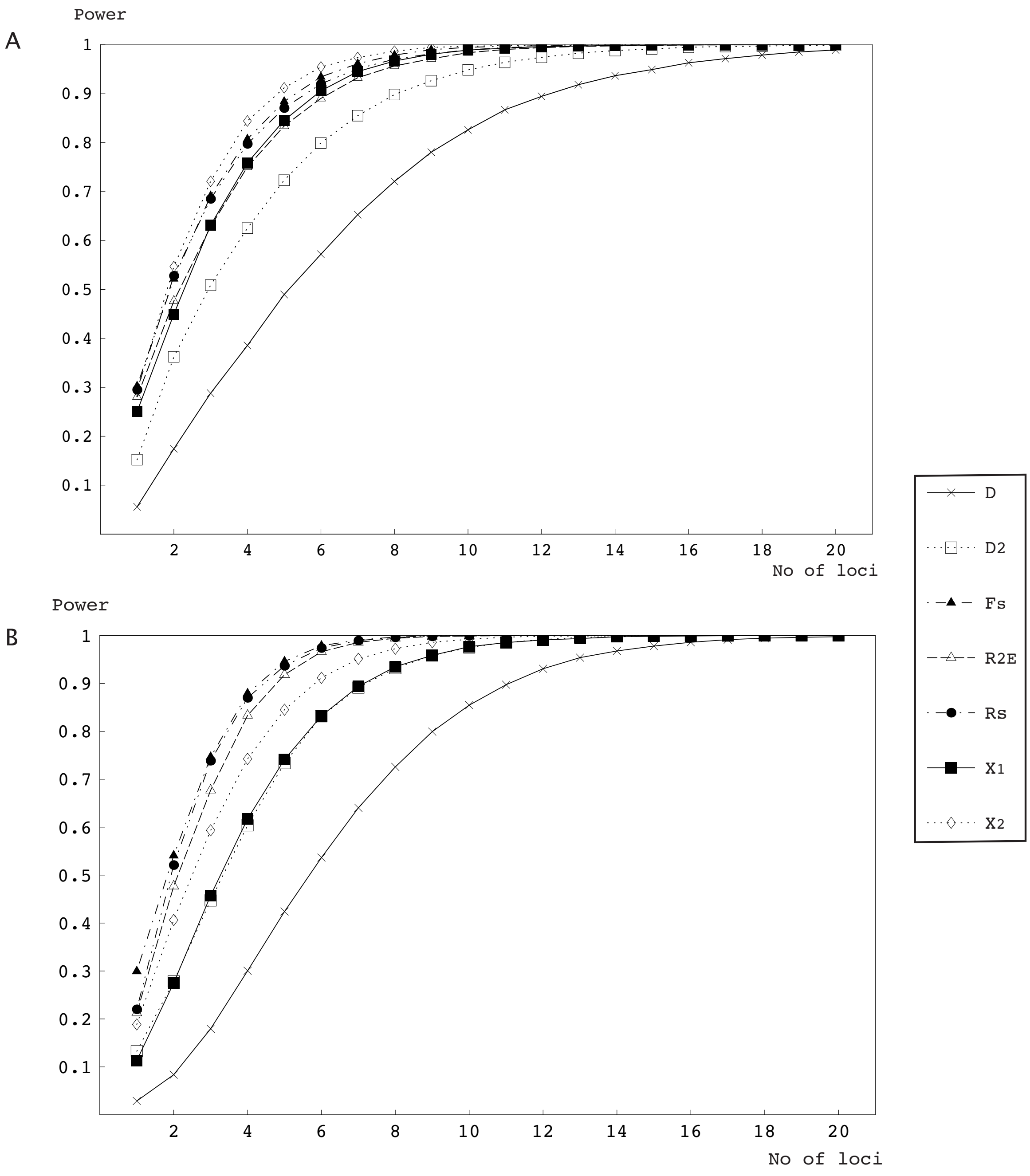

Figure 2.6: Power of summary statistics to detect a history of growth $A = 8$ using the mean across multiple loci against the number of loci, $n = 10$, A) $\theta = 20$ B) Assuming mutational rate heterogeneity (θ gamma distributed with $\alpha = 2$ and $E[\theta] = 20$).

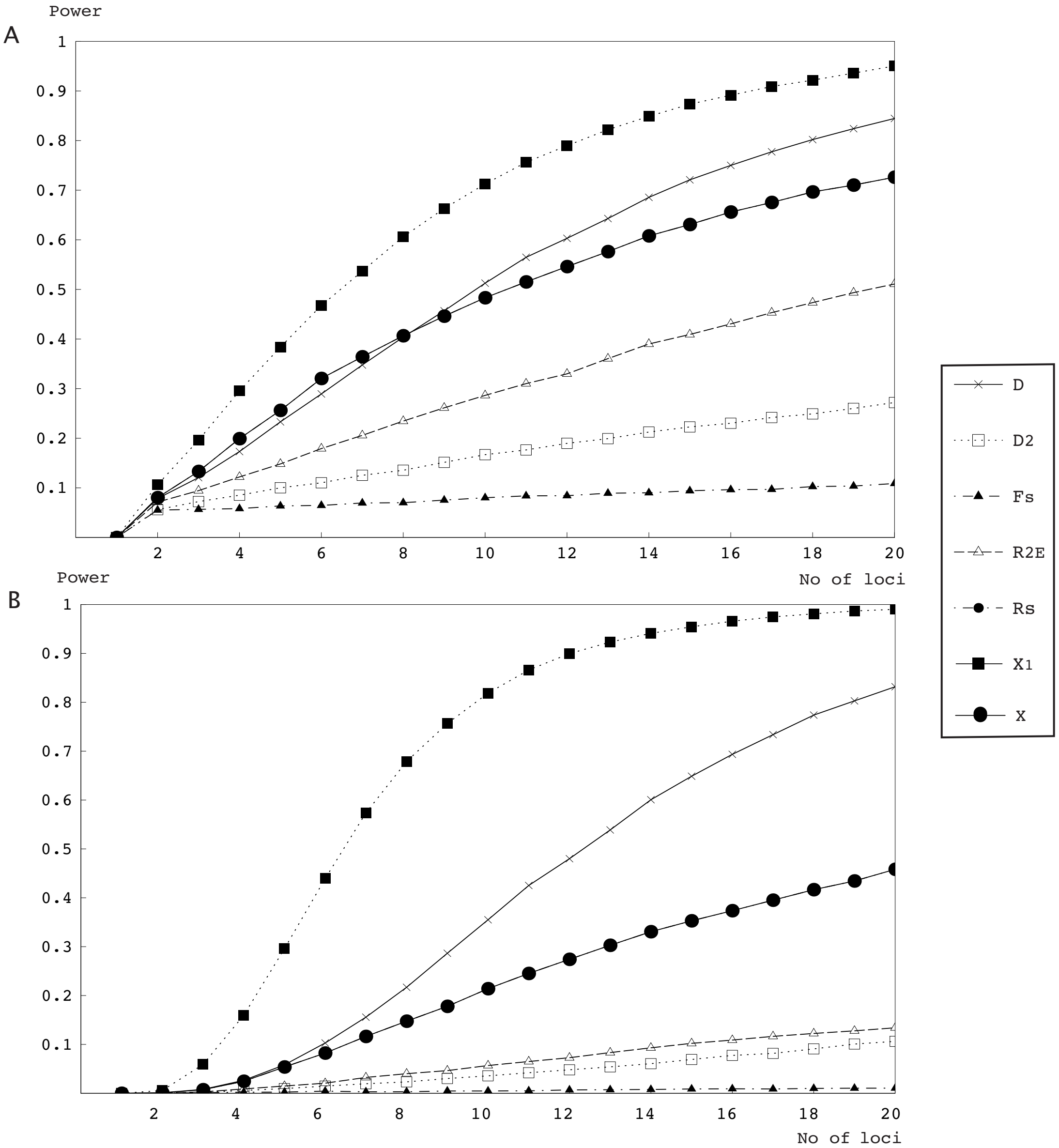

Figure 2.7: Power to detect a history of growth $A = 8$ using the variance of summary statistics across loci plotted against the number of loci. Assuming A) $\theta = 20 for all loci$ or B) mutational rate heterogeneity (θ gamma distributed with $\alpha = 2$ and $E[\theta] = 20$).

increasing the number of loci, which represent independent realizations of the ancestral process, rather than the sample size or the length of sequence (Felsenstein, 1992; Kliman *et al.*, 2000; Wakeley, 2004b).

2.4.2 Pairwise measures

Independent of the general limits to demographic inference, pairwise measures such as D have particularly low power to infer demography. This has been found in previous simulation studies, which consider other demographic scenarios such as strong bottlenecks and rapid logistic growth (Fu, 1996; Ramos-Onsins & Rozas, 2002; Ramirez-Soriano *et al.*, 2008). The fundamental flaw of pairwise measures can be best understood in terms of the underlying genealogy. In contrast to selection and population structure, changes in N_e on their own only alter the distribution of branch lengths without affecting the topology, which can be regarded as a random nuisance parameter. While the full topology can rarely be reconstructed, there is potentially a lot of topological information in sequence data. Thus the challenge that any efficient inference method has to meet is to separate this topological information from the relevant branch length information whilst taking topological uncertainty into account. Tree-based methods such as lineage-through time plots clearly fall short of the latter because they rely on a fully resolved topology. Pairwise measures on the other hand simply ignore the confounding effect of the topology (Felsenstein, 1992). It is thus easy to see why D has power only when sample sizes are large. While increasing sample size adds increasingly shorter external branches and therefore little additional information, it does reduce the chance of extremely asymmetric bipartitions by the root which are responsible for much of the variance in π and hence D.

Perhaps worryingly, this sensitivity to the topology not only translates into a loss of statistical power, but also means that negative D values may in fact be more informative about the topological asymmetry of the genealogy (which may be caused by other non-neutral forces, e.g. selection) underlying the sample than about past growth. In order to distinguish between the effects of selection and demography, topology needs to be separated from branch length information. One approach is to explicitly account for the topology information if possible. For instance one could determine confidence intervals of statistics conditional on the bipartition by the root if this is known. Not surprisingly, this improves the power of D, but has little effect on statistics that are not based on π (not shown). The alternative is to use measures which are less sensitive to the topology. F_S and other haplotype statistics have previously been shown to be more powerful than frequency spectrum statistics for this very reason (Depaulis *et al.*, 2003; Innan *et al.*, 2005). However, it has also been noted that F_S sometimes behaves erratically (Fu & Li, 1993; Ramos-Onsins & Rozas, 2002). As mentioned earlier, its power levels off with increasing θ (Fig. 2.4), because the sample size sets an upper bound to the number of haplotypes.

2.4.3 Recombination and topological uncertainty

The X statistics presented here fall somewhere between tree based methods and classical summary statistics. They exploit the fact that changes in population size disproportionally affect the relative length of the deepest branches in the genealogy and make use of topological information, without sacrificing the simplicity of the summary statistics framework. Given their high power in the multilocus case, how useful are such genealogical ratios in practice?

Recombination presents a fundamental problem to tree-based methods like the X statistics, which are defined only for non-recombining sequences. Similarly, likelihood methods which can deal with recombination are currently not available. To wrongly reconstruct trees from recombining data can potentially be severely misleading especially in the context of demographic inference. In fact, genealogical ratios similar to the ones presented here have been used to show that recombination can mimic the effect population growth has on the shape of inferred genealogies. Internal branches will appear relatively shorter and the tree overall more star-shaped (Schierup & Hein, 2000; Ramirez-Soriano *et al.*, 2008). Ideally one would like to model recombination explicitly when making demographic inferences. However estimates of recombination rates are usually associated with a large uncertainty. Furthermore, it is notoriously difficult to distinguish between recombination and back-mutations.

One approach to circumvent these problems is to test for recombination beforehand (e.g. using the four gamete test) and exclude recombinant regions from the analysis if necessary. One can then both condition on there being no within-locus recombination and afford to use more powerful statistics such as the ones presented here. This strategy of identifying non-recombining stretches of sequence is increasingly used to analyse multilocus data, (e.g. Galtier *et al.*, 2000; Jennings & Edwards, 2005). Fortunately, many organisms appear to have lower recombination rates than model species such as *Drosophila*. For instance in a recent study on Australian birds only six out of thirty loci of intergenic sequence showed evidence for recombination (Jennings & Edwards, 2005). How profitable this scheme is ultimately depends on the relative magnitude and distribution of recombination and mutation rates. Before the genealogical ratios can be used on multiple loci which have been pruned to exclude recombinant stretches, both the potential bias of such pruning and the effect of undetected recombination events on the genealogical ratios need to be properly evaluated. Interestingly, our method of inferring the root does in itself constitute a test for recombination and may help to focus on those recombination events that matter to the statistical test.

A related problem concerns the infinite sites assumption. Although the algorithm we have developed to compute the X statistics takes topological uncertainty into account, ignoring the possibility of back-mutations may underestimate the length of basal branches (Baudry & Depaulis, 2003). Although this source of error has been ignored here it should in principle be possible to account for back-mutations considering that they are independent of the assumptions of the genealogical process. In fact, any mutational model can be used to define statistics analogous to the genealogical ratios presented here. The problem with more complicated mutation models is in estimating the basal topology needed to calculate these measures.

2.4.4 Conclusions

In summary, the results confirm that only the most extreme demographic events leave a sufficient signature to be detectable in single locus data. Still, instead of the excessive and often non-quantitative employment of mismatch distributions, phylogeographic studies could benefit from using more powerful statistics such as R_S and R_{2E} to test demographic hypotheses. Conversely, population genetics studies of sequence data from multiple, unlinked loci could benefit from using summary statistics that incorporate genealogical information explicitly. When outgroup information is available and the assumptions of no within-locus recombination and infinite sites mutations can be justified, simple genealogical ratios are potentially more powerful than standard statistics. In taking the relative number of mutations found on specific parts of the genealogy as a measure of the degree of starshape, the demographic signal can be separated from irrelevant and confounding topological information. Extensions of this approach are feasible. For instance, one could consider the covariance between the number of basal and terminal mutations. Such simple statistics may be profitable for approximate likelihood or Bayesian approaches (Thornton & Andolfatto, 2006). There remains a need to understand the effect of pruning and undetected recombination events on tree reconstruction in general and tree-based measures such as the X statistics presented here in particular.

Chapter 3

Inferring the colonisation of a mountain range - refugia vs. nunatak survival in high alpine ground beetles

Published as: Lohse, K., Nicholls, J.A., Stone, N.G. (2010). *Molecular Ecology, in press*

Molecular phylogeographic studies have amply demonstrated the profound role of Pleistocene climate cycles in shaping the history of the fauna and flora in Europe (Hewitt, 2000). In general, temperate organisms survived glacial maxima in refugia south of the Pyrenees, Alps and Carpathians from which they recolonised more northern areas during interglacials. In contrast, it is less clear how cold-adapted, alpine organisms responded to Pleistocene climate change. Startled by the similarity of alpine species on different European mountain ranges, Darwin (1859) speculated, “By the time that the cold had reached its maximum we should have a uniform arctic fauna and flora, covering the central parts of Europe.” It is certainly tempting to assume that the Pleistocene history of alpine organisms is simply a reversal of the refugia/expansion dichotomy seen in temperate organisms, with range contractions into alpine refugia during interglacials followed by recolonisation of lower altitudes and latitudes during glacial periods. However, there are good arguments against such a simplistic scenario. Firstly, many alpine taxa are local endemics with poor dispersal abilities. For example, high alpine insects in otherwise winged taxa are often flightless (Hodkinson, 2005; Margraf *et al.*, 2007; Schmitt, 2009), which should greatly reduce their ability to undergo rapid range shifts. Secondly, current conditions in high alpine environments are not necessarily similar to those prevailing in the lowlands during glacial maxima. For instance, water is a limiting resource for many high alpine specialists and the dry conditions of the surrounding tundra during glacial maxima may have prevented large scale colonisation by alpine elements (Schmitt & Hewitt, 2006). Finally, glacial maxima lasted much longer than interglacials, so if extensive admixture of alpine organisms had occurred during the ice ages, their species diversity and geographic structure should generally be lower than that of temperate organisms, a pattern for which there is no evidence.

Two opposing views on the Pleistocene history of alpine biota emerged early in the development of the field of biogeography. The massif de refuge hypothesis holds that glacial survival of alpine species was restricted to large refugial areas at the periphery of the European Alps (Holdhaus, 1954), while the nunatak hypothesis proposes *in situ* survival on small ice-free islands of habitat surrounded by the ice-sheet, so-called nunataks or sky islands (Janetschek, 1956; Schmölzer, 1962). Early biogeographic studies have interpreted distribution patterns of high alpine species both in terms of the massif de refuge and the nunatak hypotheses. For instance, the absence of a number of high alpine groups from the Central Alps has been taken as evidence for very slow and incomplete postglacial recolonization originating from massifs de refuge at the periphery (Schweiger, 1969). In contrast, the extremely insular distributions of some small soil arthropods in the Central Alps are difficult to explain without invoking nunatak survival (Janetschek, 1956).

These two hypotheses also make contrasting predictions about patterns of genetic diversity within species. Under the nunatak hypothesis, ancestral variation should be more or less sorted into nunatak-specific lineages (Knowles, 2001). The rate of this process depends on effective population sizes and the time since isolation, the eventual endpoint being reciprocal monophyly (Fig. 3.1A). Furthermore, the nunatak hypothesis predicts that genetic diversity should be highest in previously glaciated areas. In contrast, under the massifs de refuge hypothesis, glaciated regions including nunataks were colonised during the current interglacial and, in general, genetic diversity should reflect refugial origin and decrease with distance from the massif de refuge (Fig. 3.1B).

Molecular studies, particularly on high alpine plants, have so far found patterns consistent with both massif de refuge and nunatak survival, although the majority of studies support the former. A meta-analysis of allozyme variation in twelve alpine plant species identified multiple, large massifs de refuge at the periphery of the Alps (Fig. 3.3A) (Schönswetter *et al.*, 2002, 2005) as well as putative nunatak survival in the Central Alps in a few species (Stehlik *et al.*, 2002). Similarly, the few molecular studies on high alpine insects in the European Alps to date (Margraf *et al.*, 2007; Pauls *et al.*, 2006; Schmitt & Hewitt, 2006) have mainly revealed genetic patterns in support of glacial survival in large and peripheral

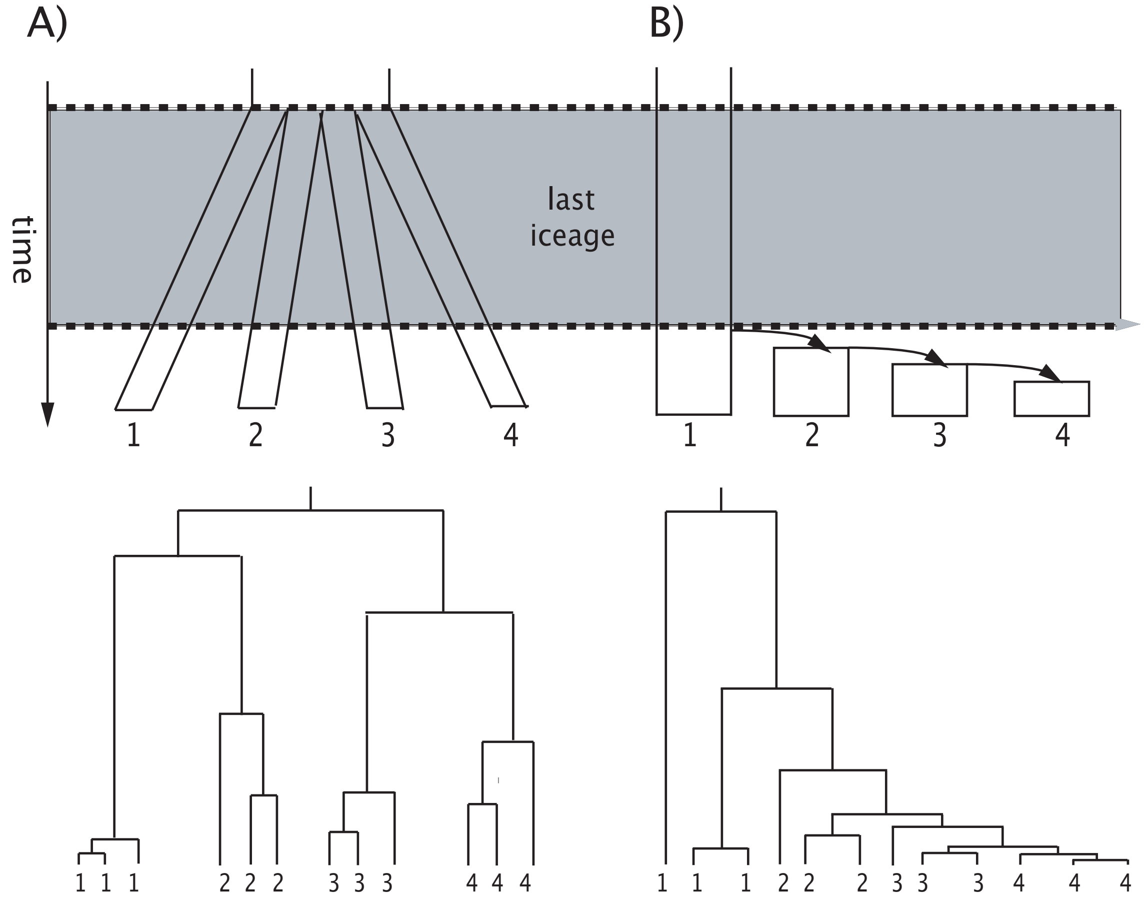

Figure 3.1: Schematic diagram of extreme population histories (above) leading to monophyletic or paraphyletic gene trees (below). A) If populations persist on multiple nunataks (1-4) and isolation is long and/or population sizes are sufficiently small, ancestral variation is sorted and populations in the gene tree are monophyletic. B) If populations are recolonized postglacially from a massif de refuge (1) through a succession of extreme founder events, populations form nested, paraphyletic clades in the gene tree. Note that in both A) and B) each location state in the genetree only ‘evolves’ once.

massifs de refuge, in most cases overlapping with those found in plants (Schönswetter *et al.*, 2005).

The balance of evidence for nunatak and massif de refuge scenarios is not only important in understanding the history of alpine species, it also has implications for their potential to adapt to local environments and to each other (Margraf *et al.*, 2007). For instance, long-term *in situ* survival on nunataks should increase local adapation and may ultimately lead to the formation of new species and communities (DeChaine & Martin, 2006). While molecular studies of high alpine taxa to date have generally aimed at resolving large-scale patterns and focused on widespread species (Pauls *et al.*, 2006; Schmitt & Hewitt, 2006; Margraf *et al.*, 2007), investigating the phylogeographic history of alpine specialists with more restricted ranges should add important resolution about the underlying processes. For example, given the complex topology of mountainous areas, it may be easier to identify which geographic features have acted as barriers to or corridors of dispersal over local scales.

Carabid beetles in the genus *Trechus* are small (2-5mm), generalist predators (Fig. 3.2) that offer ample opportunity to examine phylogeographic patterns over local scales. The genus contains more than 1000 currently described species worldwide and both species diversity and levels of endemism peak in mountainous regions (Barr, 1985; Lompe, 2004). The majority of the 60 or so Central European species are alpine or high alpine endemics with restricted ranges on the southern and northern slopes of the Alps (Jeannel, 1927; Schönmann, 1937; Focarile, 1949, 1950; Lompe, 2004).

Here we focus on a radiation of *Trechus* in the *pertyi* group in the Orobian Alps in Northern Italy (Figs. 3.3, 3.2) (Focarile, 1949, 1950). This local radiation of wingless, high alpine specialists provides an excellent test case for the nunatak and the massif de refuge hypotheses on a local scale. Firstly, the Orobian Alps constitute a geographically well-defined mountain range with sensible natural limits for a local sampling scheme: the Lago di Como in the West, the Camonica valley in the East and the Adda valley in the North (Fig. 3.3B). Secondly, the Orobian Alps are of particular interest for alpine biogeography because the maximal extent of the last glacial ice-sheet roughly divides the area in half (Jäckli, 1970) (Fig. 3.3B). Thus, while summits along the northern ridge of the Orobian Alps (pop. 2-11 Fig. 3.3B) were surrounded by the icesheet and isolated from each other as nunataks, southern summits such as Grignetta (pop. 1) and Pizzo Presolana (pop. 12) (Fig. 3.3B) remained ice-free (Jäckli, 1970) and could potentially have served as refugia during glacial maxima. Currently, high alpine *Trechus* can be found above 1800m around glacial lakes (Jeannel, 1927; Schönmann, 1937) throughout the entire Orobian Alps. In essence, this geographic set-up can be viewed as a miniature version of the pattern of nunataks and peripheral refugia in the Alps at large.

Reciprocal monophyly and polyphyly of populations are extremes in a continuum (Rosenberg, 2002). Basic coalescent theory shows that the time required for monophyly to arise after divergence depends on

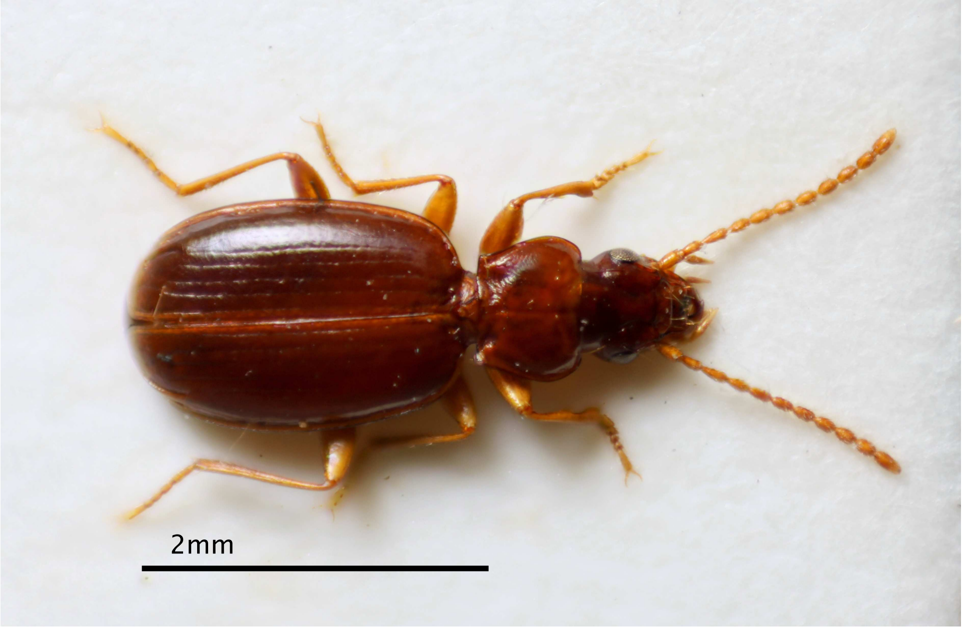

Figure 3.2: *Trechus brembanus* from Lago Verrobbio (pop. 5 in Fig. 3.3) in the western part of the Orobian Alps.

the long-term effective population size and has a very large variance (Tavaré, 1984; Hudson & Turelli, 2003). Thus if populations are large and/or stable, lineage sorting may take multiple ice ages or even predate the Pleistocene (Knowles, 2001). However, populations of Orobian *Trechus* are centred around small glacial lakes and it is difficult to imagine their effective sizes exceeding a few thousand females. In this case, the chance of monophyly as expected under the nunatak hypothesis is > 90% even after isolation for just a single glacial cycle (Hudson & Turelli, 2003) (Fig. 3.1A). Alternatively, lineage sorting can occur on a more recent time scale during a range expansion under the massif de refuge hypothesis if founder events are involved. In the simplest such case, each population is founded by just a single lineage without further gene flow between populations leading to a nested series of paraphyletic clades (Fig. 3.1B).

We sequenced two fragments of mitochondrial DNA (a total of 1431 bp) and 530 bp of nuclear sequence for a densely sampled set of populations in the Orobian Alps. We applied a recently developed Bayesian approach (Lemey *et al.*, 2009) that models directional location state changes (LSC) in gene trees. This approach, which was originally used to estimate migration rates from viral phylogenies (Lemey *et al.*, 2009; Ceiridwen *et al.*, 2010) was adopted to infer the most parsimonious set of LSC parameters connecting each population to one putative founder. Under a model of a series of extreme founder events, this set of LSC parameters determines the order of population recolonisation and thus the

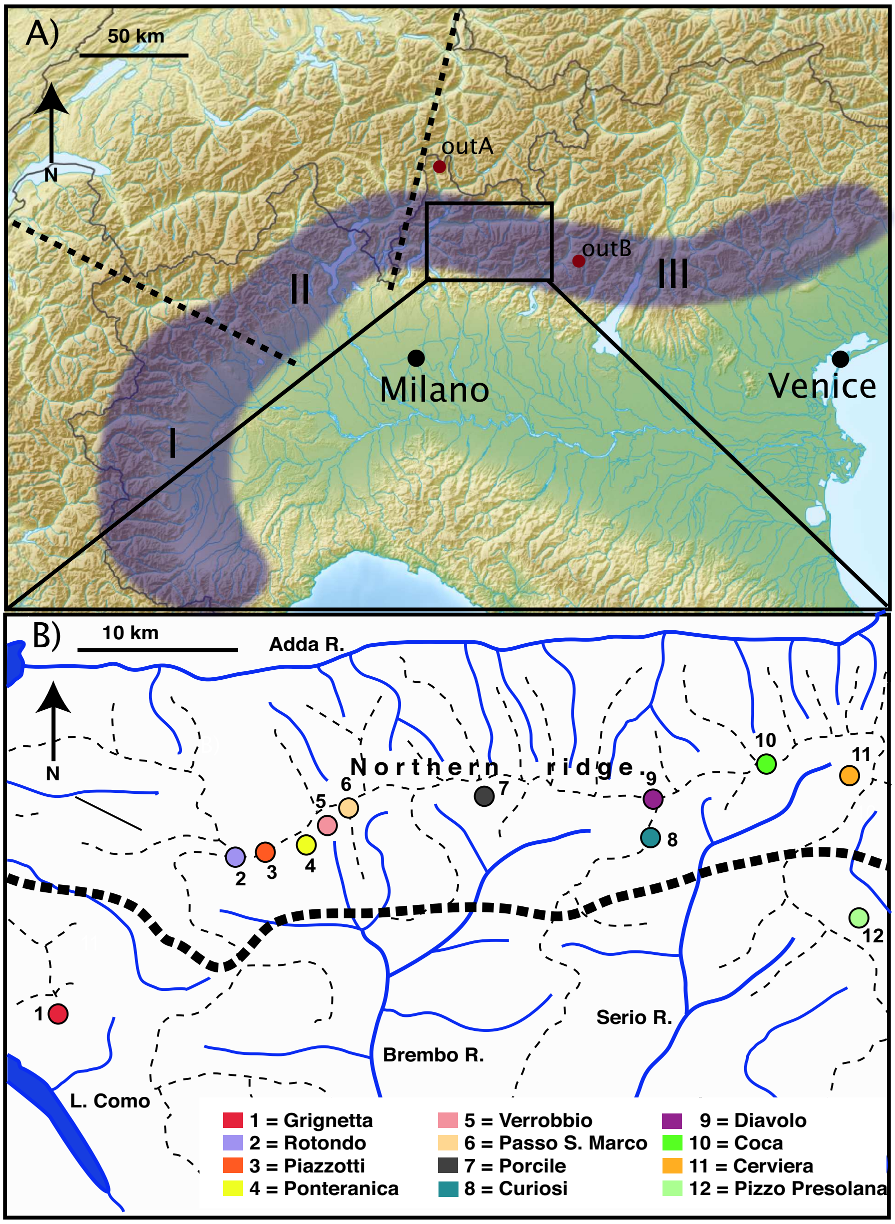

Figure 3.3: A) Main peripheral massifs de refuge (I-III) in the Western Alps inferred from a meta-analysis of genetic diversity in alpine plants are shown in purple (from Schönswetter *et al.*, 2005); breaks between refugia are indicated as dotted lines. The Orobian Alps are situated on the western edge of refugium III. Sampling localities of the geographic outgroups are indicated in red (outA= Passo di Spluga, outB = Adamello). B) Sampling localities of Trechus in the Orobian Alps. Watercourses are indicated in blue, ridges by thin dashed lines. With the exception of pop. 1 and 6, all localities are glacial lakes. The southern limit of the last glaciation (Jäckli, 1970) is indicated as a thick dashed line.

expected nesting of paraphyletic clades in the gene tree (Fig. 3.1B). Although this admittedly represents an extreme and simplistic cartoon of history, it does capture the directional aspect of recolonisation out of a massif de refuge. The great advantage of both the nunatak model and the extreme founder event model is that the expected monophyly and paraphyly in the gene tree can be tested explicitly to assess the importance of incomplete lineage sorting and/or migration (both of which lead to polyphyly). We first tested these constraints jointly for all populations and then individually for each population. Finally, we used a *Trechus*-specific, mutation rate estimate to date the age of the mitochondrial clades compatible with the extreme founder event model. This stepwise analysis allows us to address the following questions:

i) To what extent are populations on the northern ridge either reciprocally monophyletic as expected after prolonged isolation on nunataks, or paraphyletic as expected after a process of successive founder events out of one or multiple massifs de refuge?

ii) Is there evidence for incomplete lineage sorting and/or migration in the form of polyphyly?

iii) Do node ages of clades that meet the respective mono or paraphyly criteria under i) predate the last ice age as expected under the nunatak hypothesis, or are they postglacial as expected under the massif de refuge hypothesis?

3.1 Materials and Methods

3.1.1 Sampling

A total of 11 species in the *pertyi*-group have been described from the Orobian Alps, and most have allopatric distributions restricted to one or a few neighbouring mountain tops (Daniel & Daniel, 1898; Jeannel, 1927; Focarile, 1949, 1950). Their taxonomy is based on subtle differences in male genital morphology, a potentially unreliable set of traits that have been shown to vary even within populations (Faccini & Sciaky, 2002). Much of the taxonomic work on this group is linked with debates on alpine biogeography, making it difficult to gauge the extent to which species delimitations were based on vicariance hypotheses rather than morphological characters in the first place (Jeannel, 1927; Focarile, 1949, 1950). As a result, rather than sampling particular species, the aim of our sampling scheme was to provide exhaustive coverage of the Orobian Alps. We sampled a string of ten populations covering the entire length of the northern ridge as well as two populations in the south of the area: Grignetta in the southwest and Pizzo Presolana in the southeast (Fig. 3.3B, Table 3.1). Additionally, samples from two nearby

Table 3.1: Sampling localities, sample sizes (N_{mt} = number of individuals sequenced for *Cox1* and *Cox2*, N_{nuc} = number of individuals sequenced for *PEPCK*) and species names *sensu* Focarile (1950).

Code	Population	Latitude	Longitude	Alt.	*species*	Nmt	Nnuc
outA	Passo Spluga	46°30’16"N	9°19’50"E	2115m	*T. schaumii*	11	1
1	Grignetta	45°55’21"N	9°23’22"E	2170m	*T. pygmaeus*	5	2
2	L. Rotondo	46°1’6"N	9°32’17"E	2256m	*T. brembanus*	11	2
3	L. Piazotti	46°1’19"N	9°33’32"E	2224m	*T. brembanus*	9	2
4	L. Ponteranica	46°1’26"N	9°35’36"E	2150m	*T. brembanus*	9	2
5	L. Verobbio	46°2’19"N	9°35’59"E	2026m	*T. brembanus*	12	2
6	Passo S. Marco	46°2’50"N	9°37’22"E	1985m	*T. brembanus*	11	2
7	L. Porcile	46°3’42"N	9°43’55"E	2095m	*T. intrusus*	10	2
8	L. Curiosi	46°0’51"N	9°52’31"E	2112m	*T. insubricus*	11	2
9	L. Diavolo	46°2’28"N	9°53’31"E	2141m	*T. insubricus*	11	2
10	L. Cocca	46°3’46"N	10°0’4"E	2108m	*T. insubricus*	11	2
11	L. Cerviera	46°3’34"N	10°3’47"E	2326m	*T. insubricus*	12	2
12	Pizzo Presolana	45°57’26"N	10°4’14"E	2521m	*T. barii*	12	3
12	Pizzo Presolana	45°57’26"N	10°4’14"E	2521m	*T. magistretti*	12	3
outB	Adamelllo (L. Avolo)	46°3’31"N	10°29’50"E	2393m	*T. tristiculus*	3	1

mountain ranges were included as geographic outgroups: *T. schaumii* from Passo di Spluga 40 km north of the Orobian Alps (outA), and *T. tristiculus* from the Adamello range 30 km to the west (outB) (Table 3.1, Fig. 3.3A). Adult specimens were collected by hand and stored in 98% ethanol.

3.1.2 Molecular work

A total of 150 individuals were sequenced for two mitochondrial DNA loci (Table 3.1). Whole genomic DNA was extracted using a simple Chelex protocol (Lopez-Vaamonde *et al.*, 2001; Nicholls *et al.*, 2010). Primers C1-J-2792a (Bogdanowitcz *et al.*, 1993) and C2B-605 (Simon *et al.*, 1994) were used to amplify a 773 bp fragment of mitochondrial DNA which includes 180 bp of cytochrome c oxidase I (*Cox1*), 531 bp of cytochrome c oxidase II (*Cox2*) and 62 bp of tRNA leucine (Contreras-Diaz *et al.*, 2007). PCR conditions followed Moya *et al.* (2004). Additionally, the non-overlapping 658 bp ‘barcode’ fragment of *Cox1* was amplified using primers HCO/LCO and standard PCR conditions (Folmer *et al.*, 1994).

A subset of 30 individuals was sequenced for a coding region of the nuclear locus Phosphoenolpyruvate carboxykinase (*PEPCK*; Table 3.1). This gene has no known paralogs and has proven useful for phylogeographic studies of carabid beetles (Sota & Vogler, 2001; Wild & Maddison, 2008). Primers Pepck19.5 and Pepck22.5, originally developed for bees (Leys *et al.*, 2002), amplified a PCR product in some individuals. This product was sequenced and used to design the following internal, *Trechus*-specific, primer pair in Primer3plus (Rozen & Skaletsky, 2000): PepckF (5’ CGATCAAAACGGTCAACTTCC

3') and PepckR (5' AGGTTTTGGGAACGGT TCTT 3'). We used PCR conditions given by Leys *et al.* (2002) with an increased annealing temperature of 57 °C. PCR products were sequenced in both directions on an ABI 3730 automated sequencer using BigDye v3.1 chemistry.

3.1.3 Phylogenetic analysis

Complementary ABI traces were aligned in SequenceNavigator (Parker, 1997) and checked by eye. Only unambiguous consensus sequences with an open reading frame were included in the analysis and all singleton mutations were double checked in the ABI traces. Final alignments were created using the Clustal W algorithm (Higgins & Sharp, 1988). Alignment of both mitochondrial genes and *PEPCK* was straightforward. Since the mitochondrion constitutes a single, non-recombining locus the two mitochondrial fragments were concatenated for all analyses. For simplicity, the tRNA leucine region, which only contained a single, uninformative polymorphic site, was excluded resulting in a final alignment of 1366 bases for 150 individuals from 12 populations. We tested for evidence of recombination in *PEPCK* by performing a four-gamete test (Hudson & Kaplan, 1985) in DnaSP v.4.1 (Rozas *et al.*, 2003).

Mitochondrial sequences and nuclear alignments were analyzed separately. Before implementing phylogeographic models we obtained a minimally parameterized mutation model through successive model simplification and Bayes factor comparisons in BEAST v.1.5.3 (Suchard *et al.*, 2001; Drummond & Rambaut, 2007; Stone *et al.*, 2009). We began by considering the most complex models of sequence evolution possible given the sequence diversity present in each sampled locus. These were HKY+I+G for a combined partition of 1st and 2nd codon positions and GTR+I+G for 3rd codon positions within the mitochondrial data and GTR+I+G without partitioning for *PEPCK*. Standard demographic models implemented in BEAST either assume panmixia (e.g. exponential growth) or complete isolation (birth-death), both of which do not apply to structured populations. To avoid any errors resulting from model misspecification, we used a Bayesian skyline plot, which indirectly incorporates the effects of population structure by allowing for arbitrary variation in effective population size. We also tested the support for a constant versus relaxed mutation rate model (Table 3.2). We applied a *Trechus*-specific mitochondrial mutation rate estimate of a mean of 0.0152 substitutions per site per MY (equivalent to 3.04% divergence/MY) calculated for Canary Island *Trechus* species using island ages (Contreras-Diaz *et al.*, 2007). Mitochondrial analyses were run for 30 million generations with a burn-in of 20 million, repeated using different random number seeds and checked for convergence using Tracer v1.4 (Rambaut & Drummond, 2007), while PEPCK analyses were run for 3 million generations with a burn-in of 2 million generations.

3.1.4 Bayesian inference of relationships among populations

We used a recently developed Bayesian framework implemented in BEAST v 1.5.3, described in detail by Lemey *et al.* (2009), to reconstruct the colonization history of *Trechus* in the Orobian Alps. This approach models geographic locations as discrete character states 'evolving' along a rooted, time-measured phylogeny. Rates for location state changes (LSC) and ancestral location states in the gene tree are estimated simultaneously with phylogenetic model parameters using Markov chain Monte-Carlo (MCMC) sampling. In contrast to maximum parsimony, this method incorporates branch length information as well as uncertainty in gene tree topology (Pagel *et al.*, 2004; Ronquist, 2004). While the original implementation is limited to reversible LSCs, a recent extension allows the modelling of non-reversible, i.e. directional LSCs (Ceiridwen *et al.*, 2010). Location states were modeled for both data sets but only the mitochondrial DNA data contained enough information to infer a putative sequence of extreme founder events with any confidence.

For a sample of *n* locations, there are *n(n-1)* possible directional LSCs. In practice, however, many of these may not occur in a particular gene tree and the full model is drastically over-parameterized. Lemey *et al.* (2009) proposed the use of Bayesian stochastic search variable selection (BSSVS) to find a minimal set of LSC parameters. This approach has been introduced in regression problems as a way of finding a subset of potential predictors that optimally explains the variance in an multi-dimensional outcome variable (Kuo & Mallick, 1998), as deterministic model search strategies tend not to find the optimal solution unless all possible subsets are explored which is generally computationally impractical. BSSVS achieves model selection by assigning a binary indicator variable to each parameter. Each LSC parameter has an equal prior probability of being zero, which is given by a prior distribution on the total number of nonzero rates. Following Lemey *et al.* (2009), we used a truncated Poisson prior for the number of nonzero rates initially with a mean of *ln2* and an offset corresponding to the minimal possible number of rates ($n - 1 = 13$). This puts 50% prior probability on the minimal rate configuration, i.e. the model strongly favours reduced parameterisation. To assess the influence of this prior choice, we performed a sensitivity analysis by rerunning the BSSVS for larger prior means (Table 3.3).

We used Bayes factors constructed as posterior over prior odds ratios of indicators (Kass & Raftery, 1995) to assess the support for individual LSC parameters retained in the BSSVS (Lemey *et al.*, 2009) and infer the most parsimonious sequence of putative founder events using a cut-off of 3 to indicate positive support. The prior odds ratio for each LSC parameter is given by the total number of possible directional LSCs, *n(n-1)*; the posterior odds ratio is simply the proportion of generations of the MCMC during which the associated binary indicator is 1, i.e. the LSC parameter is 'switched on'. Importantly,

Table 3.2: Summary of models of sequence evolution evaluated for *Cox1/Cox2* and *PEPCK* using BEAST. The models with the highest logarithm of the harmonic mean of sampled likelihoods are indicated with an asterisk.

Cox1/Cox2			
1st and 2nd	3rd	Clock	ln(HML)
HKY+I+G	GTR+I+G	strict	-3248.10
HKY+I	GTR+I+G	strict	-3194.98
HKY+G	GTR+I+G	strict	-3207.80
HKY+I+G	GTR+I+G	relaxed	-3256.52
HKY+I	GTR+I+G	relaxed	-3184.67 *
HKY+G	GTR+I+G	relaxed	-3190.98
PEPCK			
all partitions		Clock	ln(HML)
GTR+I+G		strict	-945.54
GTR+G		strict	-941.70
GTR+I		strict	-940.93
GTR		strict	-940.56*
GTR+I		relaxed	-942.49

by assigning equal prior probability to each LSC parameter, BSSVS avoids making any assumptions about the genetic relationship of populations based on their location. Instead, genealogical relationships can be used to make inferences about likely founder events. For example, if colonisation occurs in a stepping stone fashion, most posterior probability mass in the BSSVS should be on LSC parameters between neighbouring populations and the most basal population in the set corresponds to the origin of the colonisation process. We chose the population associated with the highest posterior indicator in the BSSVS as the most likely founder of each population. The resulting set was taken as the most likely series of putative founder events for further analyses.

3.1.5 Testing topological constraints

One benefit of focusing on the two extreme histories of prolonged nunatak survival and the founder event model is that deviations from the monophyly and paraphyly criteria implicit in these models can be easily tested. Another advantage is that in both cases the time to the most recent common ancestor (T_{MRCA}) of each population gives a lower estimate of the relevant population genetic event (divergence and founder event respectively). Using BEAST we compared a topologically unconstrained model with models enforcing, i) reciprocal monophyly for all populations (prolonged nunatak survival), and ii) all paraphyly constraints given the putative sequence of founder events inferred by BSSVS (recolonisation out of a massif de refuge). We also performed constrained analyses for each population individually. In these cases we either only constrained a particular population sample to be monophyletic, or imposed monophyly of a population and all populations founded from it under the founder event model. The harmonic mean of the model likelihood (HML) was taken as an estimate of the marginal likelihood and used to compare topologically constrained models with the unconstrained model. We used a more conservative cut-off than Kass & Raftery (1995) of 2ΔlnHML= -20 to indicate strong evidence against a particular constraint.

3.2 Results

3.2.1 Phylogenetic analysis

The concatenated mitochondrial alignment (*Cox1* and *Cox2*) contained 139 polymorphic sites, 121 of which were parsimony informative. The best model of sequence evolution was HKY+I for 1st and 2nd codon positions combined and GTR+I+G for 3rd positions with a relaxed rate mutation model (Table 3.2). The alignment of *PEPCK* contained 25 polymorphic sites, 17 of which were parsimony informative. Two individuals were heterozygous at a single site and four individuals were heterozygous at three sites. In all cases it was possible to infer the haplotype phase from the different homozygotes present in the data. We found no evidence for recombination in *PEPCK*. The best model of sequence evolution for *PEPCK* was GTR for all sites and a constant mutation rate (Table 3.2).

3.2.2 Bayesian inference of relationships among populations

Using a prior mean on the number of non-zero rates of *ln2,* BSSVS on the mitochondrial data identified a set of 16 LSC parameters with a Bayes Factor > 3 among the 12 populations in the Orobian Alps (Fig. 3.4A, Fig. 3.5). Thirteen of these were between adjacent population pairs or those with only one intervening population. Given that all LSC parameters were assigned equal prior probability, that is, the prior did not incorporate information about geographic distance, this provides support for a stepping stone model of colonisation. However, BSSVS revealed a marked phylogeographic divide across the sampled area with two clusters of populations at the western and eastern end of the mountain range (Fig. 3.5). Although the number of LSC parameters supported by the data exceeded the minimum of 11 required to connect all ingroup populations, a cluster of three populations (L. Verrobio (pop. 5), Passo S. Marco (pop. 6) and L. Porcile (pop. 7)) in the centre of the northern ridge remain without a putative founder. To identify a minimal, connected set of founder events, we chose the population with the largest value

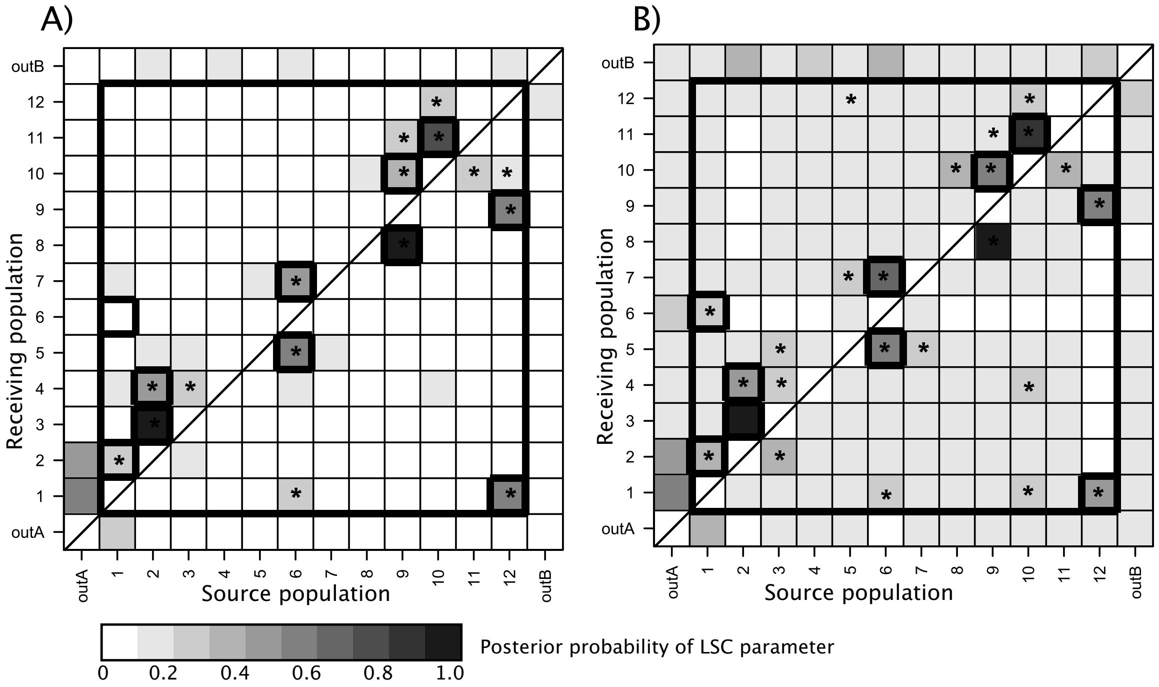

Figure 3.4: Matrices indicating posterior support for directional location state change (LSC) parameters among 12 *Trechus* populations in the Orobian Alps. Support for each parameter was assessed using Bayesian stochastic search variable selection (BSVSS) on the combined mitochondrial data *Cox1/Cox2*. Posterior support for each possible LSC parameter is indicated by the strength of the shading in the matrix (locations are ordered from west to east). LSC parameters with a Bayes Factor > 3 are indicated by an asterisk. Most posterior probability is on LSC parameters between adjacent populations (cells just below or above the diagonal). This is true regardless of whether the number of non-zero rates in the BSSVS is assumed to be close to the minimum of $n-1$ by choosing a prior mean of ln2 (A), or allowing for a much larger number of rates using a prior mean of 15 (B). Similarly, the minimal set of putative founder events (cells with thick borders) is insensitive to this parameter.

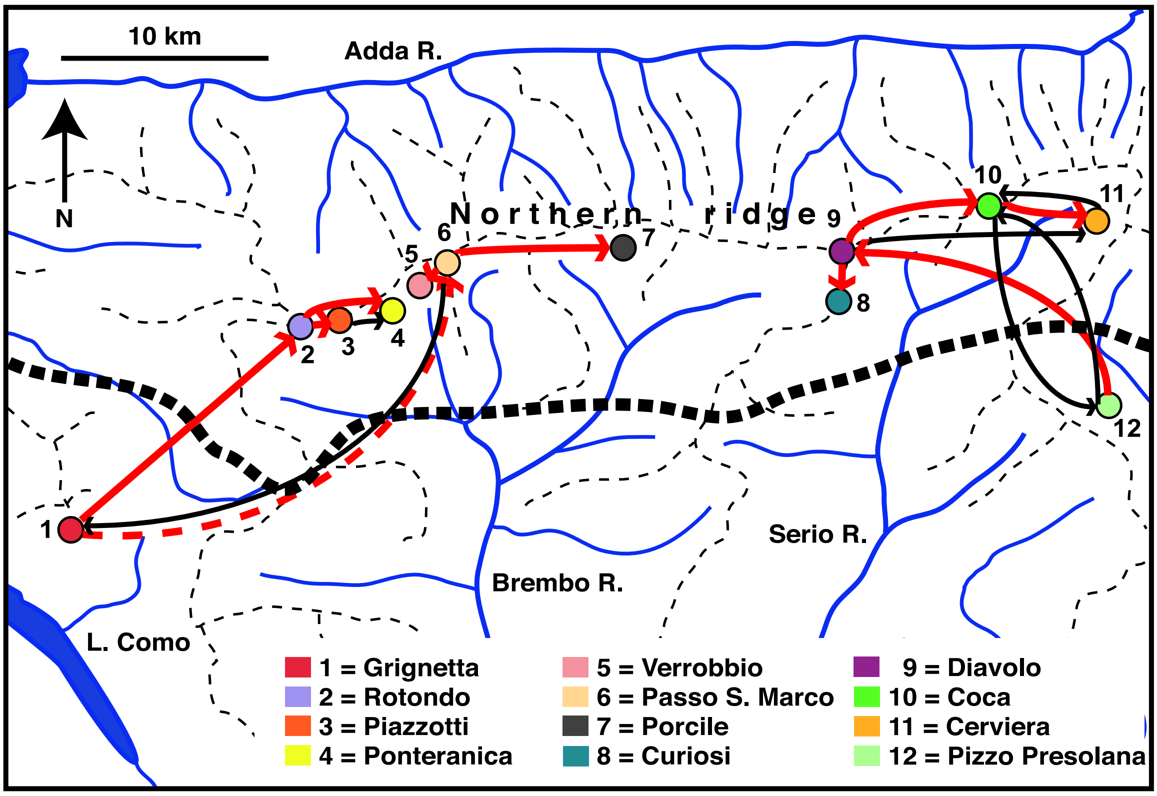

Figure 3.5: Location state change (LSC) parameters among 12 *Trechus* populations in the Orobian Alps inferred from mt genes *Cox1/Cox2* using BSSVS represented as arrows. Shown in red is the most parsimonious minimal set, which can be interpreted as a minimal model of phylogeography of sequential founder events. The putative founder event connecting populations 1 and 6 (dotted arrow) was the only LSC parameter with a Bayes Factor < 3. Additional LSC parameters with high support (BF > 3) in the BSSVS are shown as black arrows. Note that the most basal founder event (pop. 12 to 1) connecting the two clusters in the East and West of the Orobian Alps is not shown for clarity. The southern limit of the last glaciation (Jäckli, 1970) is indicated as a thick dashed line.

Table 3.3: A sensitivity analysis to investigate the impact of the prior on the number of nonzero rates in the BSSVS was carried out for the mitochondrial alignment (*Cox1/Cox2*).

Cox1/Cox2			
Prior mean	Post Median	BCI	ln(HML)
ln2	15	14,18	-3283.55
1	16	13,18	-3270.81
5	21	16,25	-3274.31
10	26	20,32	-3279.82
15	31	24,38	-3276.36*

in the corresponding row of the matrix of posterior indicators (Fig.3.5) as the most likely source of each population (red arrows in figure 4). Enforcing this reverses the connection between Grignetta (pop. 1) to Passo S. Marco (pop. 6). All other LSC parameters in the minimal set of the founder event model had a Bayes Factor > 3.

Rerunning the BSSVS with larger prior means on the number of nonzero rates either had no effect on our estimate of the marginal likelihood (ln(HML)) or increased it, with changes in the posterior median value mirroring increases in the prior (Table 3.3). We interpret this as a reconstruction of the prior resulting from the limited topological information in the data. This is confirmed by inspection of the matrix of posterior means of indicators (Fig. 3.4), as increasing the prior mean on the number of nonzero rates uniformly increased the posterior support for all LSCs, reflected by the darker background in figure 3.4B. However, the set of putative founder events inferred for the 12 ingroup populations was not affected by the prior.

3.2.3 Testing topological constraints

Constraining all populations to be monophyletic (prolonged isolation on nunataks) resulted in a drastic reduction in marginal likelihood for the mitochondrial data (2ΔlnHML= -220, Table 3.4). Similarly, imposing the full set of paraphyly constraints inferred under the founder event model also decreased the overall likelihood (2ΔlnHML= -115, Table 3.4). This indicates that neither a strict nunatak model nor an extreme founder event model is supported by the data. Evaluating constraints for individual populations, we found strong evidence against monophyly for five populations and moderate evidence for one population (Table 3.4). In contrast, only two populations (L. Piazotti (pop. 3), L. Curiosi (pop. 9)) and one population (L. Rotondo (pop. 2)) showed strong or moderate evidence respectively against the paraphyly constraints of the founder event model. This suggests that i) the extreme founder event model provides a better fit to the data; and ii) some incomplete lineage sorting and/or migration are required to fully explain genealogical relationships.

Although our approach removes the need for qualitative interpretations of gene tree topologies, many but not all of the putative founder events inferred using BSSVS are easily confirmed by visual inspection of the gene trees. Inferred ancestral location states were summarized by computing the maximum clade credibility tree for each locus (Fig. 3.6). Both mitochondrial genes and *PEPCK* unambiguously separated the sample, including the geographic outgroups, into deep western and eastern clades, which is in agreement with the founding of Grignetta (pop. 1) from Pizzo Presolana (pop. 12) being the most basal founder event. Similarly, several of the inferred putative founder events for the eastern populations clearly correspond to single transitions in ancestral location state at well supported clades in the *Cox1/Cox2* maximum clade credibility tree (e.g. from pop. 12 to 9 and from 10 to 11). In contrast, the series of putative founder events inferred by BSSVS for the western populations, in particular the basal status of Grignetta (pop. 1), are less obvious from this tree. Finally, the three populations that violated the paraphyly criterion under the founder event model were polyphyletic in the *Cox1/Cox2* maximum clade credibility tree, as expected. Samples from L. Curiosi (pop. 9) and L. Rotondo (pop. 2) occurred in multiple deeply divergent clades, which most likely reflects incomplete lineage sorting. In contrast, only a single individual from the population at L. Piazotti (pop. 3) was placed away from the majority of samples from this population into a clade of L. Rotondo (pop. 2) sequences (Fig. 3.6). Given the close proximity of the two locations (L. Rotondo is situated just 300m uphill from L. Piazotti) this may reflect a recent migration event into the L. Piazottti population. We expect LSCs that occur multiple times in the gene tree to be associated with higher posterior mean indicator values in the BSSVS, and this was indeed the case (see LSCs from pop. 9 to 8 and from 2 to 3 in Fig. 3.4, Fig. 3.6).

Under a model of extreme founder events, the T_{MRCA} of each resulting clade can be taken as a lower estimate of the time of the founder event itself. Median estimates for the T_{MRCA} of the seven northern ridge populations that were compatible with the implicit paraphyly criterion ranged from 36 KY (17 - 80 KY 95% highest posterior density) at L. Ponteranica (pop. 4) to 569KY (238 - 1,087 KY 95% highest posterior density) at L. Diavolo (pop. 8) (Table 3.4). In all cases, the lower 95% highest posterior density bound predates the onset of deglaciation at the end of the last iceage 14.5-15 KY ago, suggesting that *Trechus* were present on the northern ridge for at least part of the last ice age, if not before.

3.3 Discussion

We used a parameter-rich Bayesian approach to infer the phylogeographic history of a local radiation of high alpine ground beetles. We have deliberately focused on two extreme models of population history,

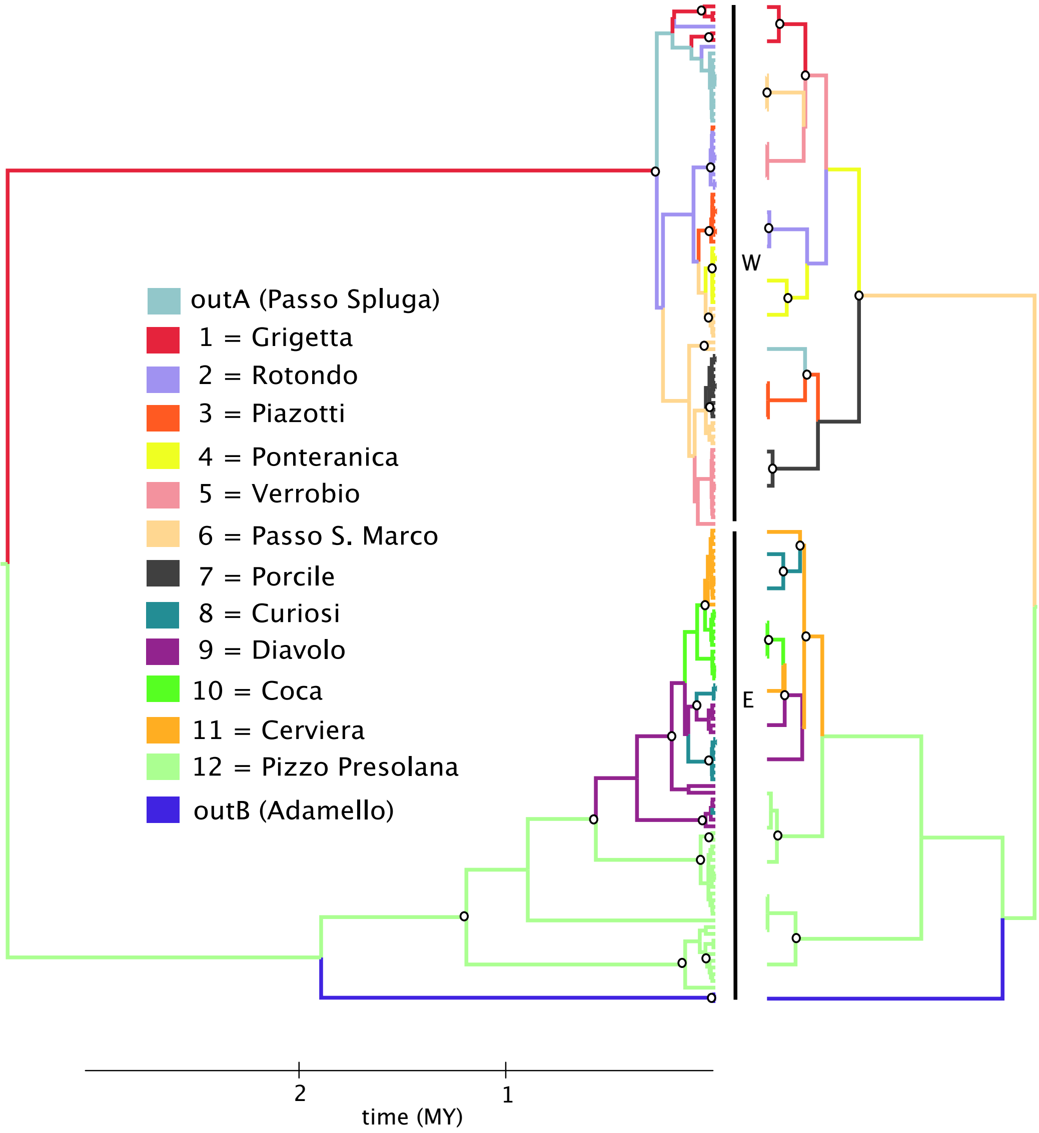

Figure 3.6: Maximum clade credibility trees for mt genes *Cox1/Cox2* (left) and the nuclear locus *PEPCK* (right). Branches are coloured by location state. Nodes with posterior support >80% are marked as white dots. Note that both mt and nuclear genetrees show a deep phylogeographic break between populations in the west (W) and east (E) of the Orobian Alps.

Table 3.4: Estimates of T_{MRCA} (median and lower and upper 95% posterior density in KY) for *Trechus* populations in the Orobian Alps under a model of sequential founder events (see red arrows in Fig. 3.5). We tested for reciprocal monophyly under the nunatak model and the monophyly implicated by the founder event model for all populations (last row) and each population separately in BEAST. Given are 2ΔlnHML (relative to the unconstrained model) combined from the two constrained runs for each populations (* indicates moderate support, ** strong support against the respective mono or paraphyly constraint) and the T_{MRCA} (median and highest posterior density (HPD) intervals) of each population obtained without imposing constraints.

Population	2ΔlnHML founder event	2ΔlnHML nunatak	median T_{MRCA}	lower 95% HPD	upper 95%HP HP
Passo Spluga (outA)	n/a	-4.6	73	19	232
Grignetta (1)	3.1	-6.9	249	58	580
L. Rotondo(2)	-19.4*	-41.8**	205	73	394
L. Piazotti (3)	-27.8**	-27.8**	123	31	285
L. Ponteranica (4)	-0.5	-0.5	36	17	80
L. Verobbio (5)	0.3	0.3	111	29	274
Passo S. Marco (6)	5.3	-27.4**	208	65	404
L. Porcile (7)	-6.7	-6.7	41	16	88
L. Diavolo (8)	-7.8	-116.4**	569	238	1,087
L. Curiosi (9)	-133.6**	-133.6**	569	238	1,087
L. Cocca (10)	-4.5	-18.8*	103	39	207
L. Cerviera (11)	- 1.2	- 1.2	47	19	91
Pizzo Presolana (12)	root	-7.3	1,245	626	2,293
L. Avolo	n/a	-2.1	17	< 1	45
all	- 115.0**	-220.4**	n/a	n/a	n/a

prolonged isolation on nunataks and extreme founder events originating from a massif de refuge. While these are admittedly simplistic cartoons of history, their advantage is that they make explicit predictions about the mono- or paraphyly relationships which can be tested for individual populations and gene trees. Our results suggest a mixture of nunatak and massif de refuge patterns. On the one hand, half of the Orobian populations are reciprocally monophyletic as expected after prolonged *in situ* survival on small nunataks and – more importantly – the ages of the corresponding mitochondrial clades would suggest that northern ridge populations diverged either before or during the last ice-age, but not afterwards. On the other hand, there are multiple lines of evidence for directional recolonisation originating from two separate massifs de refuge. Firstly, the data are incompatible with only three of the eleven paraphyly constraints under the founder event model. This suggests that, although genealogical relationships are complicated to some extent by incomplete lineage sorting and/or migration (see discussion below), the founder event model provides a reasonable fit to the data. Without information from additional loci it is impossible to tell whether polyphyly for a particular population in the mitochondrial tree is due to some process specific to these populations (e.g. large effective population size or migration) or simply due to the randomness of genetic drift. The fact that the only polyphyly observed in the PEPCK maximum clade credibility tree involves L. Cerviera (pop. 11), a population that is monophyletic in the mitochondrial tree (Fig. 3.6), points to the latter. Secondly, the data show a clear directional signal, the most likely founder of most populations being a directly adjacent population. Finally, we found a deep congruent break in the centre of the Orobian ridge in both mitochondrial and nuclear data. Pizzo Presolana (pop. 12), one of the populations in the unglaciated south, is ancestral both in the inferred sequence of founder events (Fig. 3.5) and in the eastern clade of the two gene trees (Fig. 3.6) and thus constitutes a likely massif de refuge. In contrast, the ancestral location of the western clade is less well resolved. While Grignetta (pop. 1) is ancestral both in the inferred series of founder events and the western clade of the mitochondrial tree, Passo S. Marco (pop. 6) is the ancestral location in the *PEPCK* tree (Fig. 3.6).

Taken together, these findings suggests that a model of stepping-stone type recolonisation originating from two putative massifs de refuge, although not supported for all populations, provides at least a useful approximation to the history of Orobian *Trechus*.

How can this apparent signature of directional recolonisation be reconciled with the estimates of the T_{MRCA} of the clades on the Northern ridge all of which are older than the current interglacial (Table 3.4)? Since the founder event must predate the corresponding T_{MRCA}, the results would be compatible with recolonisation during a previous interglacial (0.130 - 0.115 MYA). Such prolonged persistence of populations in isolation could potentially result in adaptation to local environments which in turn has implication for the conservation status of *Trechus* populations. Alternatively, our molecular clock calibration may be wrong. Recently Ho *et al.* (2005) have shown that estimates of molecular rates are time-dependent and have attributed this effect to purifying selection, sequencing error and saturation. Consequently, calibrations based on old events such as the age of the Canary Islands in *Trechus* may lead to considerable overestimate of recent node ages. However, in the present case the short-term substitution rate would have to be an order of magnitude higher to affect our conclusion that Northern ridge populations were seeded before the current interglacial. Another potential cause for acceleration in substitution rates is positive selection on mitochondria. Although bacterial endosymbionts such as *Wobachia* have been shown to cause selective sweeps in mitochondria in many arthropods (Hurst & Jiggins, 2005), they are not known from Carabid beetles. However, there may be other selection pressures, in particular the need to adapt to changing temperatures (Dowling *et al.*, 2008) acting on mitochondrial genes. Without more informative data from nuclear loci and mutation rate estimates for them, we cannot rule out this possibility for high alpine *Trechus*.

3.3.1 Patterns and causes of phylogeographic structure

The extent of phylogeographic structure on this small scale is in stark contrast to the complete lack of structure in more dispersive, winged insects over similar or greater scales (e.g. Nicholls *et al.*, 2010; Stone & Sunnucks, 1993). It also contrasts with mitochondrial genealogies of other high alpine radiations in which incomplete lineage sorting appears to be much more widespread (Knowles, 2001). Similar levels of genetic structure over scales of 50 km or less have to date only been found in giant springtails (Garrick *et al.*, 2009) suggesting that high alpine *Trechus* represent an extreme case of dispersal limitation and/or small population sizes.

An unexpected finding of this study was the deep phylogeographic break in the centre of the Orobian Alps supported by both mitochondrial DNA and nuclear gene trees. Simulation studies have shown that in one-dimensional habitats, such as mountain ranges, phylogeographic breaks can arise by chance without barriers to dispersal (Irwin, 2002). Furthermore, such breaks are more likely to occur in the centre of the range as is the case for Orobian *Trechus* populations. However, given the number of sampled individuals and populations it is improbable for a random phylogeographic break to occur congruently in two independently segregating loci (Kuo & Avise, 2005). Thus the east/west break in Orobian *Trechus* most likely reflects a true historic barrier to gene flow. Interestingly, the break coincides both with morphological species delimitations (*T. brembanus* and *T. intrusus* in the west and *T. insubricus* in the east) and the watershed between the two main rivers draining the Orobian Alps, the Serio and Brembo (Fig. 3.3). We therefore hypothesise that glacial range shifts and colonisation of the northern ridge proceeded along those watercourses and ultimately originated from two distinct southern refugia. This seems plausible, given the strong preference of high alpine *Trechus* for moist, glacial lake microhabitats. Moreover, genetic structure congruent with water catchment areas has previously been found in other dispersal-limited taxa (Garrick *et al.*, 2009). While passive dispersal of high alpine *Trechus* over large distances is frequently observed during flooding of alpine streams (Reitter, 1908), the present analysis suggests that active movement upstream and along mountain chains is slow. It is interesting that mitochondrial dates of most populations on the northern ridge are compatible with colonisation during or before the last ice age. In all cases, the lower 95% posterior density bound predates the onset of the current interglacial. Note that applying a mutation rate estimate from a temperate species is conservative, since one would expect high alpine specialists to have longer generation times and thus slower mutation rates than their temperate relatives, which, if anything, would push back inferred node ages.

The deep phylogeographic break observed in Orobian *Trechus* is in stark contrast to the large-scale refugia identified in plants (Schönswetter *et al.*, 2005) and suggests that patterns of vicariance and Pleistocene range shifts in alpine organisms may be highly dependent on dispersal ability and life history. It highlights the value of studying dispersal-limited alpine taxa, which are likely to preserve a signature of processes operating over local scales.

3.3.2 Locations as states in gene trees

Treating locations as discrete states in gene trees avoids many of the problems of fully parameterized population genetic models of divergence and population structure (Hey & Machado, 2003; Knowles, 2004; Wakeley, 2004b). The method is computationally tractable and LSCs inferred from gene trees can be superimposed onto the geographical map much more readily than the gene trees themselves (Fig. 3.5). The obvious drawback is that the method, if used on its own, lacks a population genetic basis and thus cannot distinguish between different processes acting at the population level. This is clearly not a problem when studying asexual organisms such as viruses whose histories can be described by a single phylogeny or - if there is reassortment - a small set of phylogenies (Lemey *et al.*, 2009). In this context directional LSC parameters can be straightforwardly interpreted as migration rates in real time (Lemey *et al.*, 2009). However, in sexual organisms gene trees and species/population trees are clearly different entities (Tajima, 1983; Pamilo & Nei, 1988) and studying the history of individual genes is only indirectly useful for making inferences about the underlying species history (Hey & Machado, 2003; Knowles, 2004). Crucially, different features of the species tree may lead to a LSC in the gene tree. For example, a particular LSC may either be associated with i) *in situ* divergence of populations, ii) the sorting of ancestral polymorphism resulting from such divergence or iii) migration of individuals between them. It is therefore problematic to equate LSC parameters as inferred by BSSVS with any one of the above processes without further testing despite multiple studies doing so (Nepokroeff *et al.*, 2003; Allan *et al.*, 2004; Lamm & Redelings, 2009). While a clear correspondence between LSCs and migration rates in the population genetic sense has been established for the symmetric island model (Slatkin & Maddison, 1998), the relationship between population genetic parameters and LSCs remains to be evaluated for more realistic, non-equilibrium models of structure.

How then can we use estimates of LSCs in gene trees to study population histories? Although desirable, analysing fully specified models is currently feasible only for small numbers of populations/species, and methods incorporating migration and incomplete lineage sorting are often restricted to pairs of populations (Hey & Nielsen, 2004; Becquet & Przeworski, 2007). The alternative is to use summary statistics and simulations to distinguish between at least some extreme alternative scenarios (DeChaine & Martin, 2006; Knowles, 2001). However, this requires making difficult choices about the range of models and parameters to be evaluated and may result in a considerable loss of information. For example, the summary statistic S, the total number of LSCs in the consensus tree (Slatkin & Maddison, 1998), which has been used to compare phylogeographic models (Knowles, 2001; DeChaine & Martin, 2006), is not informative about which changes have actually occurred. In other words, any information about the directionality of colonisation or migration is lost.

Given that a major challenge in statistical phylogeography is to identify a set of relevant models of history in the first place (Carstens *et al.*, 2009; Knowles, 2009), BSSVS should be a useful tool for reconstructing plausible population relationships that can serve as a starting point for further, model-based evaluation. It formalizes many of the qualitative inferences that researchers commonly make from ’eyeballing’ gene trees and potentially also provides a way of averaging phylogeographic signal across multiple loci. Given that BSSVS is highly sensitive to LSCs that occur multiple times in the gene tree (as would be expected from lineage sorting or migration), the approach is conservative when used to infer a putative sequence of extreme founder events, which correspond to unique LSCs in the gene tree. The downside of this is an increased sensitivity to topological uncertainty. For instance, the low power to infer a founder for the cluster of populations 5-7 and the western population in general is most likely a result of topological uncertainty. Likewise, the power to assess monophyly or paraphyly decreases with topological uncertainty. Thus despite the use of BSSVS to provide a statistical basis for phylogeographic inference, resolving population histories in detail ideally requires additional data from multiple, independent loci and more realistic population genetic models. Given the surprising extent of phylogeographic structure within a single mountain range revealed by this study and the potential insights about the effect of Pleistocene climate history on alpine diversity, further development of loci and models would be a worthwhile endeavour for *Trechus* and other high alpine specialists.

Chapter 4

Developing EPIC primers for chalcid Hymenoptera from EST and genomic data

Submitted as: Lohse, K., Sharanowski, B., Blaxter, M., Stone, N.G. (2010). *Molecular Ecology Resources*

Despite the increasing realisation that multilocus data are required to adequately resolve histories at or below the species level (Zhang & Hewitt, 2003; Jennings & Edwards, 2005; Carstens & Knowles, 2007b), the majority of phylogeographic analyses of non-model organisms are still primarily based on mitochondrial DNA. Rather than being analysed jointly in a model-based framework, nuclear data are often presented as an add-on used to 'corroborate' qualitative inferences made from mitochondrial genealogies. One reason for the relatively slow uptake of model-based approaches by phylogeographers is that obtaining a sufficient number of informative loci is a considerable effort for non-model organisms. A recent study using multiple loci to estimate divergence and migration across a phylogeographic barrier (Lee *et al.*, 2009) in a quantitative framework (Nielsen & Wakeley, 2001; Hey & Nielsen, 2004) found that stable parameter estimation requires a minimum of five nuclear loci. The general challenge is to identify enough loci that have a mutation rate high enough to generate a detectable signal of population level processes, whose evolution is at least approximately clock-like, and for which phylogeographic signal has not been overwritten by the effects of recombination. Additionally, on a practical level, amplification across related taxa is desirable both to reduce the cost of primer development and to facilitate comparisons across multiple species. In many ways this contradicts the requirement of high levels of intraspecific variation. For example, most of the loci commonly used in phylogenetic analyses or for DNA barcoding (Folmer *et al.*, 1994), such as the D2 region of the 28S ribosomal RNA gene, amplify readily across a wide range of insects (Cook *et al.*, 2002; Rokas *et al.*, 2002; Stone *et al.*, 2009), but show little or no genetic diversity below the species level (Stone *et al.*, 2007). Conversely, anonymous loci generally provide good resolution in the target species but generally do not cross-amplify well at all (Jennings & Edwards, 2005; Carstens & Knowles, 2007a; Lee *et al.*, 2009).

Introns in single-copy nuclear genes offer a potential escape from this conundrum (Creer, 2007). They evolve faster than coding regions and so are likely to contain sufficient intraspecific diversity to reconstruct genealogies, but are flanked by conserved exons (hence the term EPIC - exon-primed, intron-crossing - for such loci), which can be used as priming sites ensuring amplification across a reasonable taxonomic range (Lessa, 1992; Palumbi & S., 1994; Creer, 2007). Although intron sequences have been used in phylogeographic analyses of vertebrates (Gifford & Larson, 2008; Peters *et al.*, 2008; Lee *et al.*, 2009) and fruit flies (Wilder & Hollocher, 2003; Das *et al.*, 2004), their use in non-model taxa is still rare and their potential for comparative multispecies studies remains to be explored.

Here we develop EPIC loci for phylogeographic inference in chalcidoid parasitoid wasps (Hymenoptera: Chalcidoidea), species-rich components in most terrestrial communities and dominant natural enemies of many insect herbivores (Askew, 1980; Godfray, 1994; Bailey *et al.*, 2009). The complications of length variation in introns, which in diploid organisms often necessitates a time-consuming cloning step, can be avoided in Hymenoptera simply by using haploid males for which sequences can be obtained directly. Our aim was to identify loci that provide resolution at and below the species level whilst amplifying across a taxonomically diverse set of Chalcidoid taxa, allowing multilocus, multispecies analyses of natural parasitoid communities. To avoid having to design and optimize primers for each species individually, we took a large scale, genomic approach. The strategy was to develop primers for a large number of highly conserved genes using alignments of expressed sequence tags (ESTs) and publicly available genomic data from Hymenoptera (including the Chalcidoid, *Nasonia vitripennis)* and other insects. If transcripts have abundant conserved sites across a wide range of taxa, there should be ample non-degenerate priming sites to amplify from disparate taxa.

Amplification success of candidate loci was assessed in two diverse and well-studied, natural communities; herbivorous gall wasps (Hymenoptera; Cynipidae) on oak (*Quercus*) (Hayward & Stone, 2005) and fig wasps (Hymenoptera; Aagonidae) (Weiblen, 2002; Machado *et al.*, 2005). Primers were screened at increasing taxonomic distance from *Nasonia* (Pteromalidae); i) in different genera of Pteromalidae, ii) in different families of Chalcidoidae (Eulophidae, Eupelmidae, Eurytomidae, Ormyridae, Torymidae) and iii) for a selection of host taxa in both systems (Cynipidae and Aagonidae respectively). In total this screening set encompasses a diverse set of taxa including both pest species (Aebi *et al.*, 2006) as well as groups frequently used as biological control agents (Sha *et al.*, 2007; Mena-Correa *et al.*, 2009).

The rationale of having a large set of nuclear loci, which at least partially co-amplify across these assemblages, is to maximise overlap of loci used in future multispecies comparisons and to minimise potential ascertainment bias that species-specific choices of loci may introduce. To assess the potential of these loci for phylogeographic inference, we measured genetic diversity between major Palearctic refugia for two widespread Pteromalid parasitoids of oak galls, *Cecidostiba fungosa* and *Mesopolobus amaenus*.

4.1 Methods

4.1.1 Choice of nuclear loci and EST libraries

Putative orthologous gene alignments, developed for a separate phylogenomic study of Hymenoptera (Sharanowski *et al.*, 2010) were used to develop primers. EST alignments were constructed from cDNA libraries for six hymenopteran taxa: *Neodiprion sertifer* (Diprionidae), *Campoletis sonorensis* (Ichneumonidae), *Pelecinus polyturator* (Pelecinidae), *Pristaulacus strangliae* (Aulacidae), an unidentified ceraphronid (Ceraphronidae), and an unidentified eucoiliine (Figitidae). Sequences were also obtained from public databases (NCBI) from the following taxa: *Nasonia vitripennis* (Hymenoptera: Pteromalidae), *Solenopsis invicta* (Hymenoptera: Formicidae), *Lysiphlebus testacipes* (Hymenoptera: Braconidae), *Tribolium castaneum* (Coleoptera: Tenebrionidae), *Myzus persicae* (Hemiptera: Aphididae), *Acyrthosiphon pisum*(Hemiptera: Aphididae), and *Locusta migratoria* (Orthoptera: Acrididae). All sequences were compared against three annotated model genomes; *Drosophila melanogaster* (Diptera: Drosophilidae), *Bombyx mori* (Lepidoptera: Bombycidae), and *Apis mellifera* (Hymenoptera: Apidae). For details on cDNA library construction, contig assemblies, orthology determination, and alignment protocols, see methods in Sharanowski *et al.* (2010).

EST alignments for 76 genes meeting the orthology criterion (Sharanowski *et al.*, 2010) were filtered to include at least four hymenopteran taxa. Additionally, only alignments with less than 25% average difference at non-synonymous sites across all hymenopterans were utilized. Although this is an arbitrary cut-off, restricting the number of non-synonymous changes was intended to aid primer design by decreasing the amount of degeneracy required to achieve amplification across a broad range of taxa. The average numbers of non-synonymous sites for alignments were calculated using the Nei-Gojobori method (Nei & Gojobori, 1986) in MEGA 4 (Tamura *et al.*, 2007).

Of the 40 EST alignments meeting the above criteria, 27 were ribosomal proteins (RPs). We focused primarily on introns in ribosomal protein (RP) genes for three reasons: (i) RP genes are typically conserved across eukaryotes; (ii) most RP genes do generally not occur in multiple copies; and (iii) there is no evidence to suggest genetic linkage. We also designed primers spanning introns in 13 conserved regulatory genes that met the above criteria: *RACK1, SUI, Tctp, Mp20, myofilin, NIp ran, bellwether, AntSesB, nAcRbeta, magonashi, sansfille, pros25* (Table 4.1).

4.1.2 Primer design

EST and *Drosophila* genomic sequences were aligned in BioEdit using ClustalW (Thompson *et al.*, 1994) and checked by eye. Primers were anchored in coding exon regions flanking known introns in *D. melanogaster*. We chose priming sites that were conserved across Hymenoptera and, whenever possible, across other insect sequences in the alignment. Starting with the priming sequence for *N. vitripennis*, the only Chalcid in the set, primer degeneracy incorporating observed nucleotide substitutions at increasing taxonomic distance was built in by eye to increase amplification success. We set an upper limit of 54-fold degeneracy and attempted to choose priming sites for which all substitutions observed in the alignment could be built into the degeneracy. If this was not possible, we prioritised on degeneracy in positions near the 3' end. Sequences from the braconid wasp *L. testacipes* frequently proved too diverged to be included in the primer degeneracy. If possible multiple, often nested primers were designed for each locus (Table 4.1).

Standard primer characteristics (annealing temperature, scores for dimer formation, self annealing and 3' stability) were checked in FastPCR (Kalendar *et al.*, 2009) and Primer3 (Untergasser *et al.*, 2007) using default settings.

4.1.3 Screening amplification success

Whole genomic DNA was extracted from specimens stored in 98% ethanol in 50 μl of extraction buffer containing 5% Chelex$^{\mathrm{TM}}$100 resin (Bio-Rad, Hercules, CA). Primers were tested on three species of Pteromalid parasitoids associated with oak galls (*C. fungosa, M. amaenus, Caenacis lauta)* and three non-pollinating, parasitic Pteromalid figwasps (*Sycoscapter sp., Philotrypesis, Walkerella sp.*). *N. vitripennis*, the only chalcidoid sequence included in the EST alignments, was used as a positive control. We also tested all primers on one species from each of the remaining five Chalcidoid families parasitising oak galls in the Palearctic: *Torymus affinis* (Torymidae), *Omyrus nitidulus* (Ormyridae), *Eupelmus annulatus* (Eupelmidae), *Baryscapus pallidae* (Eulophidae) and *Eurytoma brunniventris* (Eurytomidae) (Table 4.2). However, these families are associated with foodwebs centred on many insect herbivores (Askew, 1980). Finally, primers were tested on six species of gall wasps (Cynipidae) and three species of pollinating fig wasp hosts (Aagonidae) (Table 4.2).

Polymerase chain reactions (PCR) were performed in 20 μl reactions using the following mix for all primer combinations: 2.0 ml 10x Bioline PCR buffer, 2.0 ul bovine serum albumin (10 mg/ml), 0.8 ul MgCl2 (50 mM), 0.16 ul dNTPs (25 mM each), 0.1 ul Taq Polymerase (5 U/ul, Bioline), 0.2 ul of each primer (20 uM) and 1 ul DNA template.

A generic touchdown PCR protocol was used for all loci: 94 °C for 3 min, followed by cycles of 94 °C for 15 s, an annealing step of 40 s, 72 °C for 3 min and a final step at 72 °C for 10 min. The annealing temperature was varied as follows: The first 10 cycles decreased in 1 °C increments from 65 °C to 55 °C, followed by 30 cycles each with an annealing step at 55 °C.

4.1.4 Divergence, diversity and information content

To assess the utility of the new EPIC loci for intraspecific studies we obtained sequences for two Pteromalid taxa (*C. fungosa* and *M. amaenus*). In each species three male individuals, one each from different Pleistocene refugium in southern Europe (Iberia, the Balkans and Asia Minor), were sequenced for all loci that amplified in the initial screen. Sequences were also obtained from a single male of *C. lauta*, a species closely related to *C. fungosa*. PCR products were sequenced directly in both directions using ABI BigDye chemistry (Perkin Elmer Biosystems, Waltham, MA) on ABI 3700 and 3730 sequencers in the GenePool Edinburgh. Chromatograms were checked by eye and complimentary reads aligned using Sequencher v. 4.8. *C. fungosa*, *M. amaenus* and *C. lauta* sequences were aligned in ClustalW and checked by eye. Exonic regions were assigned by comparison with *D. melanogaster* protein sequences and checked for an open reading frame. To allow comparison with a frequently used mitochondrial locus, we sequenced a 689 bp region of the cytochrome *c* subunit 1 gene (*Cox1*) for the above samples using primers COI_pF2 and COI_2413d, a modified version of C1-J-2441 (Simon *et al.*, 1994) (Table 4.1). These primers amplify a fragment largely overlapping the LCO/HCO region of *Cox1* (Folmer *et al.*, 1994), but excluding a poly-T repeat at its 5’ end present in Chalcidoidea which causes slippage during PCR resulting in uninterpretable sequence.

The final dataset for *C. fungosa* and *M. amaenus* consisted of alignments for all loci that amplified in those species. For each locus average pairwise diversity (π) in Europe (both in *C. fungosa* and *M. amaenus*) and divergence (K) between *C. fungosa* and the closely related outgroup *C. lauta* was computed

Table 4.1: Primer sequence, CG identifier, annealing temperature (C°) for 26 nuclear loci which amplified a product in at least one of the focal taxa (Table 4.2) and *Cox1*. Degeneracy codes used are standard: N = A, G, C or T; R = A or G; Y = C or T; M = A or C; S =ÊG or C; W = A or T; K = G or T; V = Not T, D = Not C, H = Not G, B = Not A.

Locus	**primer**	**CG**	**Forw**	**C°**	**Rev**	**C°**
Ant_sesB	40Fb/Rb	16944	GCCAAYGTYATCMGDTACTTC	61.8	TACKGTRTCRAAKGGATAGGA	61.7
bellwether	33Fb/Rb	3612	GAAGAGGAAGTWYGARTTRGGWC	57.5	TTCRTACCAYTGBCTGAADGG	57.9
magonashi	38F/R	9401	CTACGTCGGHCACAARGGHAART	61.5	TCTTGAACDAGRTARTAAAARCATC	60.2
nAcRbeta	39F/R	11348	GAGACBGACATCACBTTCTACAT	59.5	AGNAGATAYTTGGCRATGAGY	61.8
nAcRbeta	39Fb/Rb	11348	ATYATGAARTCRAACGTHTGG	60.1	ATGTAGAAVGTGATGTCVGTCTC	59.5
NIp	31F/R	7917	CTYTTRGGWCCAGARGCYAA	59.4	GTDSCAAGDAGATKGTGTCC	60.5
pros25	26F/R	5266	GAATATGCYTTRGCHGCNGT	60.2	GTAKGCDCCVGADGGATCAC	62.6
RACK1	18Fb/Rb	7111	GATGGGTYACBCAAATYG	61.9	ATACCTTGACDACNCGRTCC	60
ran	32F/R	1404	TAYATTCARGGMCARTGYGC	61.2	GGRTCCATTGTRACTTCTGG	60.4
RpL10ab	19F/R	7283	TAYGATCCVCARAAGGACAARC	62.5	AGGAGHCCAGGRAATTTRCCR	61.5
RpL12	10F/R	7939	GTGTACAGRCCDAMRATCGT	60	AADCCAGTTGGNARCATRTG	61
RpL13a	6F/R	1475	ATGACKGGCTTCAGYRAWAAG	57.1	GACATRAACTTYADCTTGTTCCTG	59.4
RpL15	2F/R	17420	GGGTGCNACTTAYGGHAARC	62.8	GCGMAGYTCACGRTGYTTDTG	62.8
RpL27a	28Fb/R	15442	CAAYTTYGACAARTACCATCCWG	58.7	CCYTTKCCYARRAGTTTGTA	60
RpL37	27F/R	9091	GAARGGTACNTCVAGYTTTGG	60.1	GACCRGTDCCRGTRGTCTTCCT	59.5
RpL37a	36F/R	5827	CGHACVAAGAAGGTTGGAATCAC	59.9	GTYCTYTTGCAYCGYTTGC	62.1
RpL39	16F/R	3997	ATGTCGGCHCAYAARACKTT	61.8	CTTBARCTTGGTTCKYCTCCA	58.6
RpS12	23F/R	11271	ATGGATGTSAAYACMGCMCTS	58.6	AGGGGTHTCHTCACCRAART	60
RpS15	20Fb/R	8332	GAYCARCTYCTDGAYATGC	61.9	CKACCRTGYTTWACAGGYTT	62.5
RpS17	34Fb/Rb	3922	CGCTATYATTCCWASCAARC	60.9	CAATRATRTCRTGYTCCARAGC	61.9
RpS18	22F/R	8900	GTYATGTTYGCYATGACNGC	60.1	KRAGRCCCCAGTARTGWCG	62.3
RpS23	21F/R	8415	ACVMGVTGGAAGGCYAATCC	58.2	ATGACCYTTACGHCCRAATCC	58.9
RpS4	11F/R	11276	BAARGCATGGATGTTRGACA	62.9	GGTCWGGRTADCGRATRGT	59.6
RpS8	5F/R	7808	GAAGAGGAAGTWYGARTTRGGWC	57.5	TTCRTACCAYTGBCTGAADGG	57.9
sansfille	35F/R	4528	CHWTVAAAATGCGTGGWCAAG	60.8	CDGGGAAYTGATTRAACARCAT	61.2
SUI	24F/R	17737	CCTTTGCWGATGCAATCAAG	59.4	CCGTGVACCTTSAGYTGDTC	60.5
Tctp	25F/R	4800	AYGAGATGTTCTCNGAYAC	60.1	GATRTCCATDGATTCNCCRGT	58.8
Cox1	pF2/ 2413d	n/a	ACCIGTDATRATRGGDGGITTYGGDAA		GCTADYCAICTAAAAATYTTRATW CCD GT	n/a

using DNAsp (Rozas and Rozas 1995) (Table 4.3, for *C. fungosa* individuals E1, C1, and W1 were used). These summaries were calculated separately for each locus and for intron (K_{in}, π_{in}) and synonymous exon (K_s, π_s) sites.

When choosing loci for intraspecific studies it is crucial to avoid ascertainment bias. Selecting loci based on their diversity in the focal taxon potentially confounds coalescence variance with differences in mutation rate between loci. Thus, to obtain a measure of information content based on divergence, we computed the number of divergent sites between *C. fungosa* and *C. lauta* at each locus normalized by the mean across loci. Both *bellwether* and *SUI* failed to amplify in *C. lauta*, leaving 18 loci for which divergence and information content could be computed.

4.2 Results

4.2.1 Screening amplification success

Of the 40 loci tested, 32 successfully amplified a product in the positive control, *N. vitripennis* and of these 26 yielded a PCR product in at least one other chalcid taxon (Table 4.2). Amplification success differed markedly both between different Pteromalid taxa and between Chalcidoid families. For example, fewer loci amplified in the fig-associated Pteromalids compared with the species attacking oak galls. Similarly, only 13 loci amplified in *E. annulatus* (Eupelmidae), whereas amplification success in *E. brunniventris* (Eurytomidae) (24 loci) and *Ormyrus nitidulus* (Ormyridae) (22 loci) was comparable to that in the three oak gall associated Pteromalid species. Only nine loci (*AntSesB, bellwether, RACK1, ran, RpL15, RpL37, RpL37a, RpS23, RpS4*) cross-amplified a product in all six Chalcidoid families associated with oak galls. Amplification success was considerably lower both in the Cynipidae and Aagonidae compared to any of the Chalcidoid parasitoids, which is expected given that the former are taxonomically much more distantly related to *Nasonia* (Table 4.2).

Product length varied widely both between Chalcidoid species with some combinations of primer pairs and taxa yielding PCR products in excess of 1000 bp, too long for direct sequencing (Table 4.2). Similarly, some fragments (*AntSesB* and *SUI*) were consistently larger in Cynipids than in Chalcids. Whether this variation is random or reflects genome wide differences in intron length or indeed genome size itself between hymenopteran taxa remains to be explored. The fact that the majority of the loci which amplified in *T. affinis* were longer in this species and in other Torymid species (not shown) than in any of the other 5 Chalcidoids, does suggest some general genome-wide difference between chalcid families.

Table 4.2: Amplification success and product sizes of primers developed from hymenopteran EST libraries tested on Hymenopteran taxa from two natural communities. Locus names are from FLYBASE according to the *D. melanogaster* genomic region used in the alignment for primer design. Only primer pairs that amplified in at least one of the test species are shown. Primer pairs that failed to amplified a PCR product in a particular species are indicated by 0; combinations resulting in multiple bands by D. Sequencing was only attempted in Chalcidoidae associated with oak galls (first three species in Ptermoalidae and second column). If the exact product size could not be determined due to messy sequence at the ends, only the length of the readable sequence is shown (in bold). Product sizes in the other taxa were estimated on Agarose gels.

		Pteromalidae						Chalcidoidae					Cynipidae						Aagonidae		
LOCUS	primers	*Cecidostiba fungosa*	*Caenacis lauta*	*Mesopolobus amaenus*	*Sycoscapter sp.*	*Philotrypesis sp.*	*Walkerella sp.*	*Eurytoma brunniventris*	*Baryscapus pallidae*	*Torymus affinis*	*Eupelmus annulatus*	*Ormyrus nitidulus*	*Andricus quercusramuli*	*Dryocosmus kuriphilus*	*Andricus dentimitratus*	*Plagiotrochus quercusilicis*	*Pediaspis aceris*	*Diplolepis rosae*	*Plestodontes frogatti*	*Ceratosolen appendiculatus*	*Platyneura sp.*
AntSesB	40Fb/Rb	728	612	**592**	700	700	750	622	639	**1257**	**910**	754	850	1500	1500	1500	0	750	0	0	850
bellwether	33Fb/Rb	576	D	595	D	600	D	D	444	D	D	593	D	D	0	0		D	0	D	D
magonashi	38F/R	350	-	-	-	-	-	309	0	0	0	1500	-	-	-	-	-	-	-	-	-
nAcRbeta	39F/R	289	279	279	350	350	350	546	0	0	283	618	0	0	0	0	0	0	1500	0	350
nAcRbeta	39Fb/Rb	488	485	-	600	600	600	502	800	0	0	944	0	0	0	0	0	0	0	1100	850
Nlp	31F/R	0	-	-	-	-	-	0	499	544	372	400	-	-	-	-	-	-	-	-	-
pros25	26F/R	470	472	0	500	500	550	658	445	0	0	500	0	1000	D	0	0	0	0	0	0
Rack1	18Fb/Rb	**862**	**566**	**825**	850	0	850	**1086**	882	**907**	950	892	0	0	0	0	900	0	0	0	0
Ran	32F/R	499	499	469	600	600	600	485	491	546	515	573	550	1000	500	900	600	1000	600	450	650
RpL10ab	19F/R	**968**	**1028**	**987**	1000	1000	1000	473	972	**1025**	0	**930**	1000	1000	0	1000	1000	1000	0	0	0
RpL12	10F/R	D	-	0	0	0	0	404	D	0	0	D	750	0	0	0	0	0	0	0	0
RpL13a	6F/R	864	933	0	0	0	0	**962**	0	0	0	0	0	0	0	0	0	0	0	0	0
RpL15	2Fb/Rb	652	628	642	700	700	0	527	488	511	525	488	0	700	700	0	700	850	700	0	0
RpL27a	28Fb/R	609	554	583	800	650	700	603	588	0	0	593	0	0	0	1500	800	800	0	0	0
RpL37	27F/R	**903**	952	628	650	650	650	D	613	**942**	546	504	350	400	400	400	D	600	900	650	600
RpL37a	36F/R	220	222	232	250	250	250	211	226	223	222	250	750	0	0	0	0	0	250	250	250
RpL39	16F/R	585	564	592	0	0	0	663	589	685	0	625	0	600	0	0	0	0	600	0	0
RpS12	23F/R	800	-	-	-	-	-	0	800	0	0	765	-	-	-	-	-	-	-	-	-
RpS15	20Fb/R	761	765	800	0	650	800	514	0	0	0	0	0	0	0	0	650	800	0	0	0
RpS17	34Fb/Rb	861	900	0	900	600	800	616	0	0	0	0	0	1000	0	1500	0	650	0	0	0
RpS18	22F/R	819	843	836	900	1000	1000	1000	1000	0	1500	0	900	D	1000	1000	1000	1500	1500	0	0
RpS23	21F/R	268	268	268	300	300	300	264	260	303	263	229	350	0	D	0	0	300	300	300	300
RpS4	11F/R	782	769	761	800	800	800	806	764	817	844	D	800	0	0	0	0	0	0	800	0
RpS8	5F/R	**446**	454	460	550	550	0	492	472	477	466	0	700	700	0	700	800	0	550	0	0
sansfille	35F/R	447	450	434	0	0	0	0	0	0	0	472	0	0	0	0	0	0	0	0	0
SUI	24F/R	887	0	831	0	900	0	825	797	884	0	821	1500	1500	0	1500	1500	1500	0	900	900
Tctp	25F/R	494	498	462	0	0	0	611	507	0	0	500	1000	0	0	0	0	0	0	0	0
Total(*)		26	22	19	17	18	16	24	21	14	13	22	13	12	7	9	10	12	9	8	9

Table 4.3: Sampling and rearing information of individuals used for sequencing.

	species	code	sex	*host*	country	locality	col. date	*oak host*
out	*Caenacis lauta*	Clau32	m	*Cynips korsakovi*	Iran	Azerbaijan, Ardabil For.	7/10/04	*Q. macranthera*
C1	*Cecidostiba fungosa*	Cfun0070	m	*Callirhytis glandium*	Hungary	Szentkut	3/5/02	*Q. cerris*
C2		Cfun0071	m	*Andric caputmedusae*	Hungary	Matrafured	30/6/02	*Q. pubescens*
C3		Cfun0079	f	*Andricus burgundus*	Hungary	Godollo	8/7/01	*Q. cerris*
W1		Cfun0139	m	*Andricus quercustozae*	Spain	vila, Puerto de Villatoro	2/3/06	*Q. pyrenaica*
W2		Cfun0140	f	*Andricus quercustozae*	Spain	vila, Puerto de Villatoro	2/3/06	*Q. pyrenaica*
W3		Cfun0144	f	*Andricus quercustozae*	Spain	vila, Puerto de Villatoro	2/3/06	*Q. pyrenaica*
E1		Cfun0088	m	*Andricus lucidus*	Iran	Lorestan, Piran Shahr	Oct-04	*Q. infectoria*
E2		Cfun3510	m	*Andricus polycerus*	Iran	Kordestan, Marivan	2005	*Q. infectoria*
E3		Cfun3511	m	*Andricus insana*	Iran	Kordestan, Marivan	2005	*Q. infectoria*
	Mesopolobus amaenus	Mama50	m	*Pseudoneuroterus macropterus*	Iran	Mazandaran	Oct-04	*Q. castaneifolii*
		Mama51	m	*Andricus grossulariae*	Hungary	Vitnyéd	10/05/08	*C. cerris*
		Mama55	m	*Andricus burgundus*	Spain	Caldes de Malavella	7/6	*Q. suber*

4.2.2 Divergence, diversity and information content

Taken across loci, mean per site divergence between *C. fungosa* and *C. lauta* was higher at synonymous exon sites (K_s =12.4%) than in introns (K_{in}= 6.7%). In contrast, average per site diversity was similar between synonymous sites (π_s = 0.9%, 1.1%) and introns (π_{in} = 1.0%, 1.0%) in *C. fungosa* and *M. amaenus* respectively (Table 4.4). Loci differed considerably in their overall information content (Table 4.4). In *C. fungosa* the most informative loci include *RpL37, nAcRbeta*, *RpL13a, RpS15*. Perhaps not surprisingly, those also tended to have rather high diversity in the introns (π_{in}), which in some cases was comparable to synonymous site diversity in *Cox1*. Conversely, the two loci with the lowest diversity in either *C. fungosa* (*RpL39*, *RpL37a*) or *M. amaenus* (*RpS23* and *RpS8*) had low or average information content (Table 4.4). Generally, average K_s was about three times lower for nuclear loci than *Cox1* and levels of intraspecific diversity both in *C. fungosa* and *M. amaenus* were much lower than synonymous diversity in *Cox1*. Levels of diversity observed at individual loci differed considerably between *C. fungosa* and *M. amaenus* triplets, despite the fact that the mean values were similar for the two species. For example *RpL39*, which is monomorphic in *C. fungosa*, had above average diversity ($S = 7$) in *M. amaenus* and — on a similar spatial scale — has proven to be informative in the Torymid *Megastigmus stigmatizans* (Nicholls *et al.*, 2010). This is expected because genetic diversity at a particular locus is not only determined by its mutation rate but also has a large stochastic component, due to genetic drift.

4.3 Discussion

We have shown that EPIC markers can be developed relatively easily for non-model organisms using publicly available EST and genomic data. Our strategy of testing a large number of degenerate primers on a set of focal taxa avoids time-consuming, species-specific PCR optimization, and efficiently identified a set of loci of likely value across six families of chalcidoid parasitoids and beyond. We emphasize that numbers of loci available in candidate species within these families could probably be increased by further taxon-specific PCR optimization or an additional cloning step. Although nuclear mutation rates are on average lower than those of mitochondria, this and previous studies (Lee *et al.*, 2009) show that, because of coalescent and mutational variance, the same does not necessarily hold for levels of diversity observed at individual loci. We also do not find the dramatic difference between mitochondrial and nuclear divergence which has been reported for *Nasonia* sister species and attributed to *Wolbachia*-induced sweeps (Oliveira *et al.*, 2008). Thus, despite their lower per site mutation rate, multiple EPIC loci such as the ones developed here, if analyzed jointly, should be far more informative about within-species phylogeographic history than mitochondrial data (see chapter 5).

Table 4.4: Basic properties of nuclear loci in *C. fungosa* and *M. amaenus*. Length values exclude indels in the *C. fungosa alignment*. Diversity across three major Pleistocene refugia and divergence between *C. fungosa* and *C. lauta* were calculated for introns (π_{in}, K_{in}) and synonymous exon sites (π_s, K_s) separately. Also shown are the number of introns (#In), the total number of polymorphic sites (S) in the single triplets and, for *C. fungosa*/*C. lauta*, the relative mutation rate μ and information content (Info). Loci for which larger samples for *C. fungosa* were obtained for the Bayesian analyses presented in chapter 5 are shown in bold.

		Length			*C. fungosa / C. lauta*				Diversity (*C. fungosa*)			Diversity (*M. amaenus*)		
LOCUS	primers	#In	Total	Intron	K_s	K_{in}	μ	Info	π_s	π_{in}	S	π_s	π_{in}	S
AntSesB	40fb, 40rb	2	728	156	0.076	0.148	0.984	0.981	0.000	0.008	2	0.000	0.024	7
bellwether	33fb, 33rb	1	550	216	n/a	n/a	n/a	n/a	0.000	0.003	2	n/a	n/a	n/a
nAcRbeta-64B	39f, 39r, 39fb, 39rb	2	728	113	0.371	0.227	1.703	2.039	0.004	0.000	1	0.000	0.044	10
Rack1	18fb, 18rb	2	560	304	0.087	0.052	0.627	0.578	0.000	0.007	3	0.021	0.010	10
Ran	32f, 32r	1	496	202	0.090	0.091	0.802	0.659	0.011	0.003	2	0.000	0.009	3
RpL10ab	19f, 19r	2	955	807	0.072	0.043	0.641	1.001	0.000	0.003	3	0.044	0.006	9
RpL13a	6f, 6r	2	851	720	0.000	0.097	1.414	1.975	0.000	0.019	21	n/a	n/a	n/a
RpL15	2fb, 2rb	2	617	412	0.233	0.056	1.047	1.065	0.000	0.002	2	0.000	0.011	7
RpL27a	28fb, 28r	2	549	338	0.155	0.101	1.309	1.078	0.017	0.030	16	0.000	0.007	4
RpL37	27f, 27r	1	869	788	0.017	0.123	1.882	2.681	0.033	0.020	24	0.000	0.016	13
RpL37a	36f, 36r	1	220	91	0.408	0.069	1.203	0.436	0.000	0.000	0	0.000	0.013	2
RpL39	16f, 16r	1	465	444	0.000	0.086	1.386	1.055	0.000	0.000	0	0.000	0.009	7
RpS15	20fb, 20rb	1	756	475	0.073	0.091	1.076	1.308	0.058	0.035	30	n/a	n/a	n/a
RpS18	22f, 22r	2	813	562	0.072	0.052	0.757	1.011	0.020	0.005	6	n/a	n/a	n/a
RpS23	21f, 21r	1	247	79	0.119	0.127	0.926	0.408	0.016	0.042	6	0.016	0	1
RpS4	11f, 11r	2	745	431	0.094	0.083	1.040	1.290	0.000	0.000	1	0.000	0.008	7
RpS8	5f, 5r	1	458	242	0.060	0.034	0.447	0.311	0.029	0.008	6	0.000	0.003	1
sans_fille	35f, 35r	1	446	84	0.140	0.037	0.501	0.367	0.017	0.000	2	0.017	0.000	2
SUI	24f. 24r	1	823	636	n/a	n/a	n/a	n/a	0.000	0.006	6	0.000	0.006	6
Tctp	25f, 25r	2	493	148	0.134	0.088	0.826	0.670	0.000	0.014	3	0.040	0.018	8
Total		30	12232	7249							136			97
MEAN			**611.6**	**362.9**	**0.139**	**0.073**			**0.009**	**0.010**	**6.8**	**0.011**	**0.010**	**6.1**
COI		n/a	698	n/a	0.353				0.090		24	0.209		54

If patterns of divergence across loci in *C. fungosa* and *C. lauta* are at all representative, the most informative loci for within-species historical inferences in Chalcidoids are likely to include *RpL37*, *nAcRbeta*, *RpL13a*, *RpS15*, *RpS4* and *AntSesB*. If, as recent power analyses suggest, between five and a dozen loci are sufficient to reliably infer ancestral population parameters in divergence models (Jennings & Edwards, 2005), these EPIC loci should allow multilocus phylogeographic analysis across a broad taxonomic range of chalcidoid parasitoids, and in turn, facilitate comparative phylogeographic analysis of natural chalcidoid assemblages. The observed variation in amplification success between families would suggest that it may be impossible to use a standard set of loci across taxa even if this may be desirable to avoid confounding true differences in species histories with locus-specific effects. However, as long as enough loci per species are sampled to capture the variance in genealogical history and outgroup comparisons are used to account for heterogeneity in mutation rates across loci, there is no *a priori* reason against using only partially overlapping sets of loci in multi-species comparisons. Given that the primers developed here are anchored in highly conserved coding regions and at least partially amplify across a large taxonomic range, they may also prove useful as genomic tools more broadly in the Hymenoptera and other Insects. For example, some of the loci employed in this study (e.g. *RpL15, RpL27a, ran*) have previously been used as markers for QTL mapping in Lepidoptera (Papanicolaou *et al.*, 2005).

An important question is to what extent introns in highly conserved genes evolve neutrally. Generally, our finding of lower levels of divergence in introns compared to synonymous sites in *C. fungosa* is consistent with previous results from genome wide studies in *Drosophila* suggesting that introns are under purifying selection, which may be particularly strong in highly conserved genes (Haddrill *et al.*, 2005; Halligan & Keightley, 2006). Similarly, negative correlations between intron length and divergence have been interpreted as evidence for selective constraints on regulatory elements present in long introns (Halligan & Keightley, 2006). We tested for this in *C. fungosa* and found a negative but non-significant trend between intron length and K_{in} (*r* = -0.265, *p* = 0.189). This suggests that any correlation between intron length and selective constraint, if present in *C. fungosa*, is likely to be weak. Thus, it may be difficult to avoid potential biases arising from selective constraints by selecting short introns. On the contrary, since information content is a function of both intron length and K_{in}, the most informative loci in the present set are those containing long introns (Table 4.4). However, while selective constraints on introns or linked exons should not lead to systematic biases in estimates of ancestral population parameters, they may result in lower information content than that expected in selectively neutral regions. This has been demonstrated previously in a study on birds (Lee *et al.*, 2009) which found per site diversity in anonymous loci, presumably intergenic DNA, to exceed those in introns. On the other hand, using EPIC loci with known orthology and function for phylogeographic inference can be viewed as an improvement over anonymous loci for which orthology and function are generally unknown (Jennings & Edwards, 2005). In general, with the increasing volumes of publicly available genome data making primer development for non-model organisms straightforward, multilocus nuclear sequence data will surely become the standard in studies of population history and phylogeography rather than the exception.

Chapter 5

Quantifying the Pleistocene history of the oak gall parasitoid *Cecidostiba fungosa* using twenty intron loci

Published in slightly modified form as: Lohse, K., Sharanowski, B., Stone, N.G. (2010). *Evolution, in press*

Many western palaearctic taxa have their current centres of genetic diversity to the east of Europe, suggesting that refugial populations around the Mediterranean basin are ultimately derived from a more eastern source (Din *et al.*, 1996; Rokas *et al.*, 2003; Juste *et al.*, 2004; Michaux *et al.*, 2004; Culling *et al.*, 2006; Koch *et al.*, 2006; Challis *et al.*, 2007; Stone *et al.*, 2007). Westwards dispersal of such taxa into southern European refugia is often thought to have occurred in the early Pleistocene, if not before (e.g. Taberlet *et al.*, 1998; Rokas *et al.*, 2003; Juste *et al.*, 2004; Culling *et al.*, 2006; Challis *et al.*, 2007) and of necessity must predate the well-documented latitudinal range shifts associated with the last iceage (Taberlet *et al.*, 1998; Hewitt, 1999) by at least one glacial cycle. However, the few studies that have attempted to estimate the age of this older longitudinal dispersal are largely qualitative, being based on a small set of (primarily mitochondrial) gene trees (e.g. Taberlet *et al.*, 1998; Hewitt, 1999; Rokas *et al.*, 2003; Juste *et al.*, 2004; Culling *et al.*, 2006; Challis *et al.*, 2007). It has been noted that species differ considerably in their mitochondrial divergence between refugia and this has been attributed to species-specific responses to Pleistocene climate cycles (Taberlet *et al.*, 1998). However, an obvious alternative explanation for the observed lack of interspecific temporal congruence is that mitochondrial gene trees are dominated by incomplete lineage sorting, the extent of which may be large in general and/or different between species (Nichols, 2001).

Because polymorphism within ancestral populations must originate before daughter populations diverge, branches of gene trees are necessarily longer than those of population trees and a naïve interpretation of node ages may severely overestimate population divergence (Pamilo & Nei, 1988; Maddison, 1997). Similarly, gene tree topologies may be incongruent with the order of population divergence (Tajima, 1983; Pamilo & Nei, 1988; Rosenberg, 2002). Since the magnitude of both these effects depends on the size and stability of the ancestral populations (Tajima, 1983; Maddison, 1997; Nichols, 2001), they are likely to be exaggerated when resolving the origins of - and relationships among - refugial populations, which are stable by their very nature (Hewitt, 1999). Thus, assessing the generality of an ‘Out of the East’ pattern ideally requires replication both at the level of species and loci.

Assemblages of parasitoids associated with oak cynipid galls offer unmatched replication at the species level. In the Western Palaearctic, an estimated 120 species of chalcidoid wasps are obligate natural enemies of the inhabitants of oak cynipid galls (Csóka *et al.*, 2005; Hayward & Stone, 2005). Phylogeographic studies on Western Palaearctic oak gallwasps show their populations to be divided into three major refugial areas: the Iberian Peninsula in the west, Central Europe and the Balkans in the center, and Asia Minor and Iran in the east (Rokas *et al.*, 2001, 2003; Stone *et al.*, 2001; Challis *et al.*, 2007; Stone *et al.*, 2008), broadly paralleling patterns seen in oak phylogeography (Dumolin-Lapegue *et al.*, 1997). In the gallwasps, allele frequency data for multiple nuclear markers support the conclusion that there has been very little subsequent gene flow between these regions (Rokas *et al.*, 2001; Stone *et al.*, 2001; Rokas *et al.*, 2003; Challis *et al.*, 2007; Stone *et al.*, 2008). Oak gallwasps are thought to have diversified in regions to the east of Europe prior to the Pleistocene (Stone *et al.*, 2009), and pre-Pleistocene or early Pleistocene westwards range expansion across Europe has been suggested by patterns of genetic variation in several widespread species (Rokas *et al.*, 2001, 2003; Challis *et al.*, 2007). An obvious question is whether gall-associated parasitoids have pursued their hosts from the east. At least two of them, the torymids *Megastigmus stigmatizans* and *M. dorsalis*, appear to have done so (Hayward & Stone, 2006; Nicholls *et al.*, 2010). The challenge now is to reconstruct longitudinal colonisation processes in the Western Palaearctic for a broader taxonomic spread of oak gall-associated parasitoids, to assess the generality of an ‘Out of the East’ pattern, and to determine whether parasitoids dispersed over a similar timescale to their hosts, or after a delay – so allowing their hosts a measure of ‘enemy free space’ (Hayward & Stone, 2006). One reason for caring which of these scenarios is true is that close phylogeographic concordance

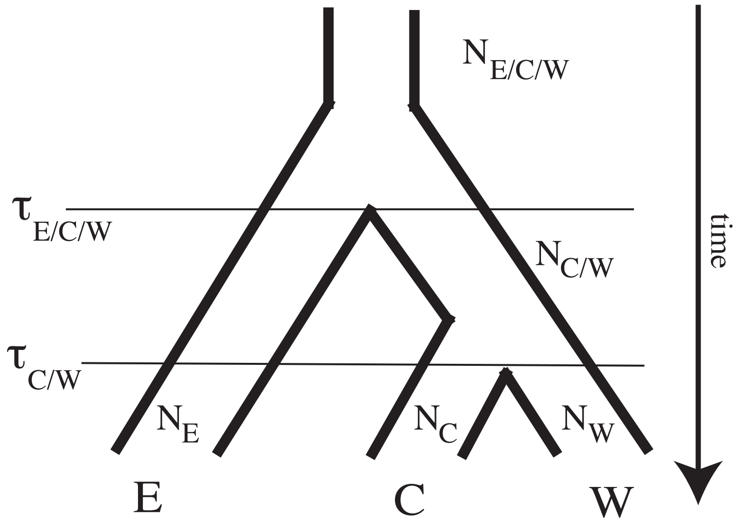

Figure 5.1: Model of successive population divergence between major Palearctic refugia from East to West: Asia Minor and Iran (E) Balkans and Central Europe (C), Iberia (W). With minimal sampling of one individual per population, topological probabilities of gene trees are determined by only 2 model parameters, the time between population divergences ($\tau_{\text{E/C/W}}$-$\tau_{\text{C/W}}$) and the effective sizes of the ancestral population during this time ($N_{\text{C/W}}$).

increases the potential for coevolution among community members, and such communities are inherently sensitive to disturbance by species gain (Stone & Sunnucks, 1993; Schönrogge *et al.*, 1996a, 1998) or loss (Lennartsson, 2002; Pauw, 2007).

Here, we use sequence data from 20 intronic loci to study the history of refugial populations in the pteromalid parasitoid *Cecidostiba fungosa*, a widespread species in oak gall communities (Askew, 1961; Schönrogge *et al.*, 1996a; Bailey *et al.*, 2009). The three-refuge phylogeographic pattern of oak gallwasp communities allows us to compare two analytical methods - a maximum likelihood (ML) approach (Yang, 2002), and an analogous, Bayesian approach (Rannala & Yang, 2003). Both estimate ancestral population parameters (population sizes and divergence times) directly from patterns of polymorphism in sequence data (rather than from gene trees inferred for each locus) and assume a model of divergence between three populations (Fig. 5.1). The order of population divergence or the topology of the population tree can be viewed as an additional model parameter and the likelihoods in both methods can be used to compare statistical support for different topologies. We address the following, specific questions:

i) Do data for *C. fungosa* support an ‘Out of the East’ population history, such that refugial populations in the centre and west of Europe are derived from a shared ancestral population in the centre which in turn is derived from a common ancestral population further east (Fig. 5.1)?

ii) When did refugial populations split from each other, and how large were their ancestral populations?

iii) How different are multilocus estimates of population divergence times from gene divergence times (both nuclear and mitochondrial)?

A strategy of sampling many loci from a single individual per taxon has been used extensively to study divergence between closely related species, in particular the Great Apes (Yang, 2002; Jennings & Edwards, 2005; Patterson *et al.*, 2006). There are two reasons why such minimal sampling is of interest. Firstly, going backwards in time, only lineages that persist into the ancestral species/population contribute to estimates of ancestral population parameters. Coalescent theory shows that samples taken from the same species or population quickly coalesce down to a small number of lineages (Griffiths, 1981; Tavaré, 1984; Nordborg, 1998) (Fig. 5.2). This means that even if divergence is relatively recent, i.e. less than N_e generations ago, the power gained by increasing within population sampling levels off relatively rapidly. In contrast, each additional sampled locus provides an independent replicate of the coalescent process in the ancestral population irrespective of the divergence time (Wakeley, 2004b). So if the total cost of sampling is number of loci x number of sampled individuals, the optimal sampling scheme is one of few individuals sequenced at a large number of loci. Secondly, minimal sampling is currently the only sampling scheme for which a statistically optimal likelihood method allowing parameter estimation directly from site patterns exists (Yang, 2002). In contrast, Bayesian approaches (Rannala & Yang, 2003) or gene tree - species tree methods (Degnan & Salter, 1995; Degnan & Rosenberg, 2009; Maddison & Knowles, 2006; Liu & Pearl, 2007; Kubatko *et al.*, 2009) have the advantage that they can deal with arbitrary sample sizes and numbers of populations. However, this comes at the potential cost of prior assumptions and/or difficulty in integration over topological uncertainty in the gene trees.

These issues are relevant in selecting an appropriate study design in systems where there is a trade off between sampling multiple individuals and generating data for multiple loci or species. Ability to obtain informative population parameters from small numbers of individuals is likely to be particularly important in comparative studies of communities, such as the oak gall system, in which some taxa are rare enough that increasing sample size is not an option. It is therefore useful to ask how much information about ancestral population parameters over phylogeographic timescales can be obtained with minimal sampling. To investigate the influence of sample size, we compared minimal sampling of a single individual per population with an extended sample of three individuals per population. We then use theoretical expec-

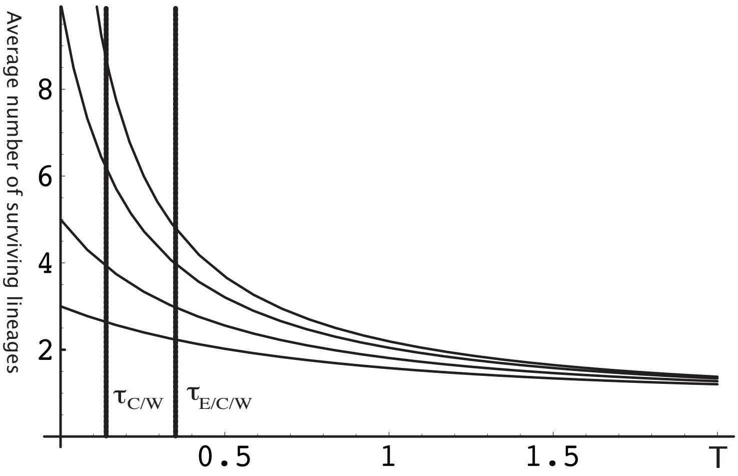

Figure 5.2: The expected mean number of lineages surviving coalescence into an ancestral population (Tavaré, 1984, equation 5.5) plotted against divergence time (T) in coalescence units ($2N_e$ generations) for 4 different sample sizes (from top to bottom, n = 20, 10, 5, 3). Since only surviving lineages contribute to the estimation of ancestral parameters and their number decreases rapidly, the expected gain in power from increasing sample size is limited even if divergence is relatively recent (T < 0.5). The solid lines show the divergence time estimates (scaled by twice the mean of population sizes N_{E}, N_{C} and N_{W} obtained for *C. fungosa* in this study (priors a).

tations for the number of surviving lineages given the estimated divergence history (Fig. 5.2) to consider the likely gain in power for larger sample sizes in our Discussion.

5.1 Methods

5.1.1 Choice of loci

We obtained sequences for 20 newly developed intronic loci for *C. fungosa* and the closely related species *Caenacis lauta*, which was used as an outgroup in some analyses. These loci included twelve ribosomal protein genes (*RpL10ab, RpL13a, RpL15, RpL27a, RpL37, RpL37a, RpL39, RpS15, RpS18, RpS23, RpS4, RpS8*) and eight regulatory genes (*AntSesB, bellwether, nAcRbeta-64B, Rack1, Ran, sansfille, SUI, Tctp*) (Table 4.1, 4.4), all of which are thought to be single copy genes with no known paralogs in insects. Primer development and testing is described in detail in chapter 4. No or little polymorphism at a particular locus may arise either as a result of a low mutation rate (so limiting signal), or a recent coalescent event (and so important to demographic inference), or both. Excluding loci that are invariant in *C. fungosa* results in an upward bias in estimates of population divergence time. To avoid such bias, we used all nuclear loci available for *C. fungosa* (Table 4.4) and tested whether accounting for differences in mutation rate between loci influenced our estimates. To allow comparison of information content in the nuclear loci with a frequently used mitochondrial locus, we sequenced a 689 bp region of the cytochrome *c* subunit 1 gene (*Cox1*) using primers COI_pF2 and COI_2413d (Table 4.1).

5.1.2 Molecular methods

Whole genomic DNA was extracted from specimens stored in 98% ethanol in 50 μl of extraction buffer containing 5% Chelex$^{\text{TM}}$100 resin (Bio-Rad, Hercules, CA). To allow for direct sequencing of PCR products without the need to discriminate between haplotypes in heterozygotes, we used males, which are haploid in Hymenoptera, whenever possible. The exceptions were three female *C. fungosa*, for which haplotypes were distinguished by cloning of PCR products as necessary (see below). PCR mixes and conditions used were as described in chapter 4.

All PCR products showing single amplified bands were sequenced directly in both directions using ABI BigDye chemistry (Perkin Elmer Biosystems, Waltham, MA) on ABI 3700 and 3730 sequencers in the GenePool Edinburgh. Chromatograms were checked by eye and complimentary reads aligned using Sequencher v. 4.8.

For five loci (*RpS4, RpL27a, RpL37, RpL15b, nAcRbeta*) sequences from female individuals of *C. fungosa* contained putative heterozygous sites or were not readable due to indels. These PCR products were cloned using a mini-Prep kit (Qiagen, Valencia, CA). Five clones were sequenced per locus and individual, one of which was chosen at random for subsequent coalescent analyses. In one case (sample C3, locus *RpS4*) none of the sequenced clones matched the expected product. This sample was excluded from the analysis.

5.1.3 Model of population divergence and population sampling strategies

We consider a simple model of divergence between three putative refugial populations of *C. fungosa*: Asia Minor and Iran (east, E), Balkans and Central Europe (centre, C), and Iberia (west, W).This is analogous to a model of divergence between three species (Takahata *et al.*, 1995; Yang, 2002) that has been used to estimate divergence times and ancestral population sizes in Great Apes (Rannala & Yang, 2003; Patterson *et al.*, 2006), fruit flies (Villablanca *et al.*, 1998; Li *et al.*, 1999), birds (Jennings & Edwards, 2005) and plants (Zhou *et al.*, 2007). The model makes the standard population genetics assumptions of random mating within each population, fixed population sizes between divergence events, and no migration after divergence. The first and last assumptions at least are supported by multilocus allele frequency data for the gallwasp hosts in this system (Stone & Sunnucks, 1993; Rokas *et al.*, 2003; Stone *et al.*, 2008).

Following recent studies on Hominids and model organisms (Takahata *et al.*, 1995; Li *et al.*, 1999; Chen & Li, 2001; Rannala & Yang, 2003; Jennings & Edwards, 2005; Patterson *et al.*, 2006), we initially adopted a sampling scheme that maximises the number of loci available by using only a single haploid male from each of the three refugial populations listed above (Table 4.1). To examine the impact of increased sampling within populations, we generated an extended dataset, comprising three haploid sequences per population for 13 loci (names shown in bold in Table 4.4) and a single sequence per population for the remaining seven loci as before. Impacts of further increases in sample size will be considered based on the theoretical expectation of the number of surviving lineages (Fig. 5.2).

We used ML (Yang, 2002) and Bayesian approaches (Rannala & Yang, 2003) (described below) to

i) test whether the most likely order of population divergence is compatible with an 'Out of the East' scenario and

ii) estimate divergence times and ancestral population sizes under this scenario using the single individual per population sampling.

To investigate the impact of sample size on parameter estimation, Bayesian analyses were repeated using the extended dataset as defined above.

5.1.4 Alignment and mutation rate

C. fungosa and *C. lauta* sequences were aligned in ClustalW and checked by eye (Genbank accession nos. HM208872-HM209026). Exonic regions were assigned by comparison with *D. melanogaster* protein sequences and checked for an open reading frame. Indels in the alignment were treated as missing data.

In the ML and Bayesian analyses all model parameters are scaled by the per site mutation rate, μ. Conversion of the scaled time between divergence events (γ) into real times (τ), and of the scaled mutation rate (θ) into effective population sizes (N_e), therefore requires an estimate of μ and its incorporation into the relationships $\gamma = \tau\mu$ and $\theta = 4N_e\mu g$, where g is the average generation time in years. Note that for haplodiploids $N_{e\_hd} = (9N_f N_m)/(2N_f + N_m)$, where N_f and N_m are the number of males and females respectively in a randomly mating population. Assuming equal sex ratio and variance in fitness between sexes, N_{e_hd} is 0.75 N_{e_d} (Hedrick & Parker, 2003).

To calculate a mean estimate of μ for our loci we first estimated a synonymous genome-wide mutation rate for the closely related pteromalid wasp genus *Nasonia*, using a divergence time of 0.4 MYA between *N. giraulti* and *N. longicornis* (Campbell *et al.*, 1993; Oliveira *et al.*, 2008; Raychoudhury *et al.*, 2009) and a nuclear genome-wide distance at synonymous sites (K_s) of 0.011 between these species (Oliveira *et al.*, 2008). With $\mu = K_s/2t$ these values give 1.375×10^{-8} b/yr. The *Nasonia* divergence time was derived by applying estimates of bacterial silent sites substitution rates (Ochman & Wilson, 1987) to *Wolbachia* symbionts infecting the two *Nasonia* species(Raychoudhury *et al.*, 2009). Although such estimates may have a substantial error (Ho *et al.*, 2005; Pulquério & Nicholls, 2007), it should be noted that the resulting nuclear substitution rate for *Nasonia* is roughly similar not only to the few other molecular clock calibrations that exist for insects, e.g. 1.11×10^{-8} b/yr for Hawaiian Drosophilids (calibrated from island ages (Tamura *et al.*, 2004)), but also agrees with rate estimates derived from mutation accumulation experiments by order of magnitude (Keightley *et al.*, 2009).

To apply the *Nasonia* mutation rate to our intron-rich (and so partially non-coding) sequences, we scaled it by the ratio of the observed average divergence between *C. fungosa* and *C. lauta* at synonymous sites, K_s over the average divergence across all sites K_{Total}. This yields a factor of 0.478, so the total average substitution rate for our loci is $\mu = 1.375 \times 10^{-8} \times 0.478 = 6.27 \times 10^{-9}$ b/yr. Note that since this is an average across all sites, it is lower than the substitution rate for synonymous coding sites. This calculation incorporates any mutational constraints on introns and coding sites in *C. fungosa* without making *a priori* assumptions about intron evolution. We estimated a relative mutation rate for each locus as the observed K_{Total} at each locus over the average K_{Total} (Chen & Li, 2001; Yang, 2002; Jennings & Edwards, 2005), shown in Table 4.4.

To calculate ancestral effective population sizes we assumed an average generation time of $g = 0.5$ years for *Nasonia* and *C. fungosa*. This is reasonable for *C. fungosa*, which attacks both sexual spring galls and asexual autumn galls (Askew, 1961; Schönrogge *et al.*, 1995, 1996a) (as synonyms *C. adana* and *C. hilaris*), and for temperate populations of *Nasonia*. For comparison with mitochondrial node ages we calculated a mutation rate for *Cox1* using the Jukes-Cantor corrected distance between *N. giraulti* and *N. longicornis* at this locus and a divergence time of 0.4 MYA as before. This gives 22.3% (Oliveira *et al.*, 2008) divergence per site and per million years. We compared this locally calibrated clock with estimates obtained in previous studies using the commonly assumed arthropod mitochondrial clock of 2.3 % per site and per million years (Brower, 1994). Despite the obvious shortcomings of the ‘Brower clock’, comparison of relative node ages in this way is valid as long as the same calibration is used across taxa, and a molecular clock assumption is tested and supported in each taxon, as here.

5.1.5 Recombination tests and gene tree reconstruction

Both phylogenetic reconstruction and the coalescent analyses described below make the crucial assumption of no recombination within loci. We determined the minimum number of recombination events using a four-gamete test in DNAsp (Rozas & Rozas, 1995) on the largest alignment of each locus. Three loci (*RpS4*, *RpS18*, *RpL15*) showed evidence for recombination and were trimmed to the largest non-recombining block (Galtier *et al.*, 2000; Jennings & Edwards, 2005). Alignments for these loci were shortened by 117, 16 and 132 bases respectively as a result.

Although both the ML and Bayesian approaches described below use site patterns directly and do not rely on estimated gene trees, we reconstructed trees to visualize the data and to test the molecular clock hypothesis which is implicit in both approaches. ML trees were reconstructed for each locus in PAUP* (Swofford, 2001). For single individual alignments (triplets) this was done using exact searches, while for the three individual per population alignments branch and bound searches were used. Loci varied considerably in relative intron length and hence in base composition. We therefore assumed a single substitution rate but unequal base frequencies (Felsenstein, 1981). To test the support for internal nodes in each triplet gene tree, 1000 bootstrap replicates were performed taking a bootstrap value of 70% to indicate strong nodal support (Hillis & Bull, 1993). We compared rooting with a strict molecular clock to rooting with *C. lauta* for the triplet gene trees (Jennings & Edwards, 2005). To further test the validity of the molecular clock assumption, we performed Tajima's 1-degree of freedom test on each triplet (Tajima, 1993; Jennings & Edwards, 2005; Tamura *et al.*, 2007). This nonparametric test is designed for triplet samples given a known species topology and is simpler and more powerful than similar model-based tests (Tajima, 1993; Nei & Kumar, 2000; Jennings & Edwards, 2005).

5.1.6 Maximum Likelihood analysis

For minimal sampling, only four parameters in the three-population divergence model matter: the two divergence times $\tau_{\mathrm{C/W}}$ and $\tau_{\mathrm{E/C/W}}$ and the sizes of the two ancestral populations $N_{\mathrm{C/W}}$ and $N_{\mathrm{E/C/W}}$ (Fig. 5.1) and an exact likelihood approach to inference is possible. The program Ne3sML numerically maximises the likelihood for a given population/species topology (Yang, 2002). By default the method assumes an infinite sites mutation model and a molecular clock. Given the level of polymorphism observed in *C. fungosa* (Table 4.4), this simple model of sequence evolution seems appropriate. For example, if diversity at silent sites (synonymous exon sites and introns) is 0.01(Table 4.4), the chance of a back mutation is 10^{-4} per site. Since we are analysing slightly fewer than 10^{4} silent sites in total, we expect to see at most a single backmutation in the entire dataset and can safely ignore more complicated mutation models.

The likelihood approach of Yang 2002 differs crucially from methods which estimate a species tree conditional on a set of reconstructed gene trees(Degnan & Salter, 1995; Degnan & Rosenberg, 2009; Maddison & Knowles, 2006; Carstens & Knowles, 2007a; Liu & Pearl, 2007; Kubatko *et al.*, 2009) in that it uses the site information directly. The method integrates over all possible gene tree topologies and branch lengths at each locus and computes the joint log likelihood for a given population history (topology and parameter estimates) as the sum over the log likelihoods of individual loci (Yang, 2002; Rannala & Yang, 2003). The advantage of this is that in contrast to gene tree species tree approaches (Liu & Pearl, 2007; Degnan & Rosenberg, 2009; Kubatko *et al.*, 2009), information from unresolved or poorly resolved loci is incorporated automatically. This is particularly important in recently diverged populations. For example, a monomorphic locus resulting from a recent coalescence event would be excluded from analyses conditional on gene tree reconstruction as uninformative, resulting in upwardly biased estimates of divergence time.

We first compared the likelihood of all three possible population tree topologies. Although assessing the statistical significance of non-nested models is difficult in a likelihood setting, models may be ranked by their likelihood (Carstens *et al.*, 2009). Under the 'Out of the East' scenario, central and western populations are derived from a shared ancestral population in the centre, which in turn split from a common ancestral population in the east, i.e. the population tree topology is (E, (C, W)) (Fig. 5.1). The two alternative topologies are (W, (C, E)), which corresponds to an 'Out of the West' scenario, and (C, (E, W), which is difficult to interpret in the geographic context of *C. fungosa* populations, because it is unclear where the two ancestral populations would be located.

ML analyses under the most likely population history were performed for two different mutational models. The simplest model assumes a single mutation rate across all loci. We reran this analysis using the relative rates calculated for each locus as described above (Table 4.4), thereby accounting for possible rate heterogeneity (Table 5.2).

5.1.7 Bayesian estimation of divergence times and ancestral population sizes

MCMCcoal (Rannala & Yang, 2003) is the Bayesian equivalent of the ML approach described above. The program uses Markov chain Monte-Carlo sampling (MCMC) to estimate posterior probabilities for all model parameters conditional on prior distributions. If multiple individuals per population are sampled the three population sizes between the present and the most recent divergence event (i.e. N_{E}, N_{C}, N_{W}) (Fig. 5.1) are modelled as additional parameters. Note that the parameterization in MCMCcoal differs slightly from Ne3sML, as the former uses divergence times rather than internode intervals.

In a Bayesian framework support for alternative but non-nested models can be compared using Bayes factors (Kass & Raftery, 1995). Natural logarithms (ln) of the harmonic mean of sampled likelihoods (HML) were used to estimate the marginal likelihood of each population tree topology (using prior means in analysis *a* described below) and to test support for the 'Out of the East' scenario. Following Kass and Raftery 1995, values of twice the difference in lnHML(2ΔlnHML) of 2-6, 6-10 and >10 represent respectively positive, strong and very strong support for the model with higher marginal likelihood.

Since in the case of *C. fungosa* we have no prior knowledge of the model parameters, we used exponentially distributed priors (shape parameter $\alpha = 1$) for all parameters (Jennings & Edwards, 2005). To check how sensitive posterior estimates are to prior settings, all analyses were performed twice using different prior means, by adjusting β, the scale parameter of the gamma distribution (Table 5.3). In the first analysis (*a*), we set prior means to 0.150 MYA and 0.050 MYA for $\tau_{\text{E/C/W}}$ and $\tau_{\text{C/W}}$ respectively (β = 380) and 215,000 for both ancestral population sizes (β = 1520). In the second analysis (*b*), the prior means for all parameters were increased by an order of magnitude (i.e. changing β to 38 and 152) (Table 5.3). Although the individual parameter values are arbitrary these two sets of priors should be different enough to assess the robustness of the Bayesian estimation (Jennings & Edwards, 2005). Given that incorporating relative mutation rates did not improve estimation using the ML method (see Results), for simplicity all Bayesian analyses were performed assuming a single mutation rate across all loci. Runs were continued for 10^6 generations with a burn-in of 10^5 and repeated using different random number seeds to check for convergence.

5.2 Results

5.2.1 Gene trees

When only a single individual was sampled from each refugial population, phylogenetic reconstructions for eight of the 18 polymorphic nuclear loci supported the ‘Out of the East’ topology (E, (C, W)) (Fig. 5.3A), as did the mitochondrial locus *Cox1* (Fig. 5.2D). Of the remaining loci, two supported each of the two incongruent topologies (Fig. 5.2B, C) and six showed an unresolved topology (*RpL15, RACK1, ran, Tctp, sansfille, SUI*). Clock-rooted and outgroup-rooted topologies agreed for all resolved loci, but bootstrap support was generally weaker for outgroup rooting (Fig. 5.3). Though this is not a formal test, the majority of resolved gene trees thus support the ‘Out of the East’ hypothesis (Fig. 5.1). Tajima’s 1-D test rejected a strict molecular clock for only two out of 20 loci (*RpS15, RpL 37*). Thus the majority of loci meet the clock assumption implicit in the ML and Bayesian approaches used here.

Increasing sample size to three individuals from each refugial population resulted in increased variation in gene tree topology (Fig. 5.4). Despite the many unresolved nodes in some trees, figure 5.4 reveals extensive incomplete lineage sorting between *C. fungosa* populations, resulting in a ‘forest’ of largely incongruent gene trees.

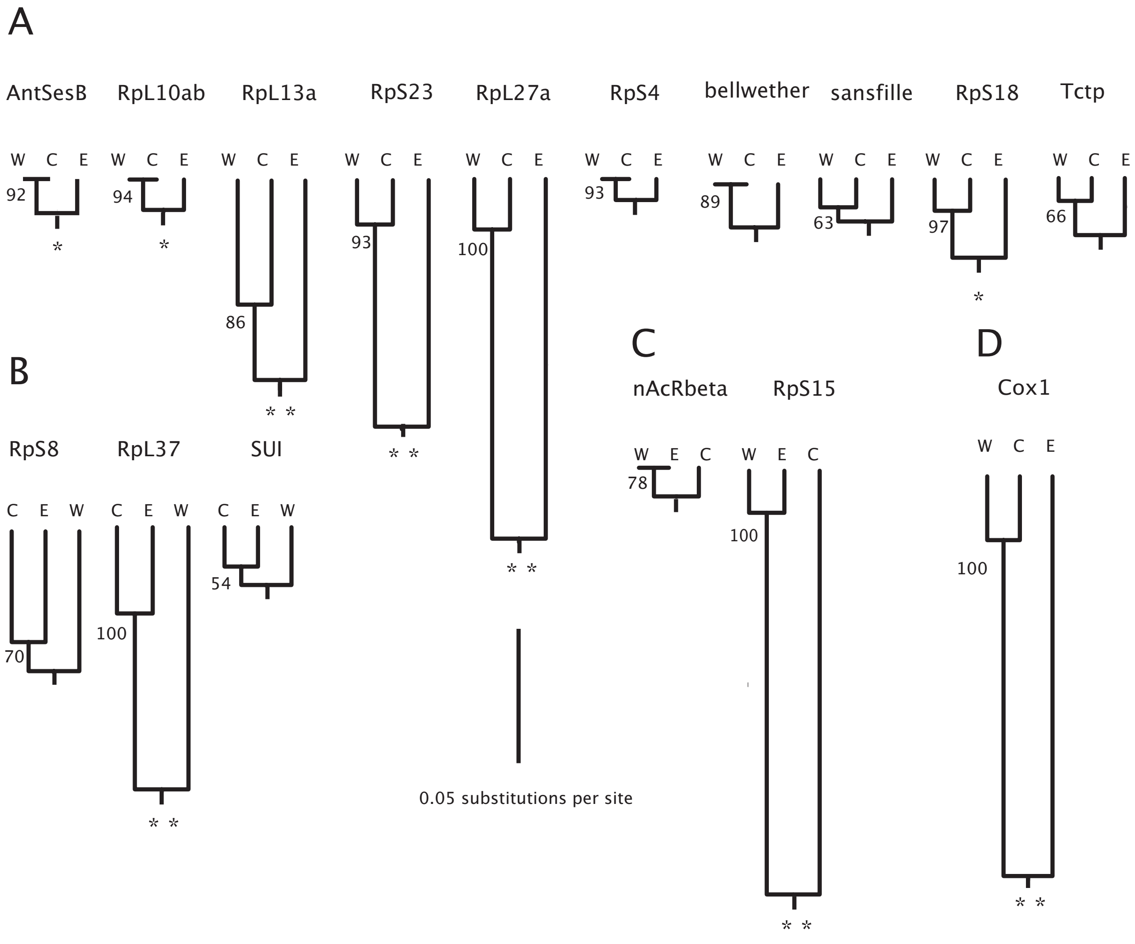

Figure 5.3: ML trees reconstructed for nuclear loci and *Cox1* assuming a strict molecular clock. Bootstrap proportions for the internal node are shown next to each tree. Loci with unresolved topologies (< 50 % bootstrap support) are not shown. Eight loci have a topology congruent with the 'Out of the East' hypothesis (E, (C, W)) (A), two each have topology (W, (C, E)) (B) and (C, (E, W)) (C). The mitochondrial locus *Cox1* is also congruent with 'Out of the East' (D). Bootstrap support using rooting with *C. lauta* is indicated with asterisks (* > 50%, ** >70%) below each tree.

Table 5.1: Comparison of support for alternative population tree topologies, using the lnL of the maximum likelihood estimation (NeML3s) and the harmonic mean likelihood (lnHML) in the Bayesian analyses. In each case the 'Out of the East' topology has the highest likelihood (in bold). Values in parentheses show the ln Bayes factor (2ΔlnHML) of the 'Out of the East' hypothesis relative to alternatives. Topologies which fit significantly worse than the 'Out of the East' hypothesis are indicated with asterisks, using a ln Bayes factor (lnBF) of 2-6 to indicate positive support (*), 6-10 to indicate strong support (**), and >10 to indicate very strong support (***), following Kass & Raftery (1995).

	Out-of-the-East	Out-of-the-West	
	(E, (C, W))	(W, (C, E))	(C, (E, W))
NeML3s (single triplet) lnL	-796.94	-799.06	-799.05
MCMCcoal (single triplet) lnHML(lnBF)	-19100.69	-19103.82 (6.25)**	-19103.06 (4.73)*
MCMCcoal (extd. triplet) lnHML(lnBF)	-19558.24	-19563.90 (11.324)***	-19559.00(0.76)

5.2.2 Maximum likelihood analyses

The population tree topology (E, (C, W)) had a higher likelihood than either of the two alternative topologies (C, (E, W)) and (W, (C, E)), consistent with the 'Out of the East' hypothesis (Table 5.1). The maximum likelihood estimates (MLEs) of model parameters are broadly consistent between the variable rate (18 loci) and single rate mutational models (using the same 18 loci). However, because the variable rates model has a lower log likelihood, the simpler single rate model was used in all subsequent analyses including the Bayesian runs. This also allowed the loci *SUI* and *bellwether*, for which no outgroup sequences could be obtained, to be included in the analyses, giving a total of 20 loci.

Under the 'Out of the East' topology (E, (C, W)), the MLE for the older population splitting time between the Iranian population and the ancestor of Hungary and Spain, $\tau_{\mathrm{E/C/W}}$, is estimated as 0.110 million years ago (MYA; Table 5.2). The MLE for $\theta_{\mathrm{E/C/W}}$ corresponds to an ancestral population with an effective size of 614,000 before this first split. However, both the MLE for the time between the two population splits, $\tau_{\mathrm{E/C/W}}$ - $\tau_{\mathrm{C/W}}$ and the population size during that time, $N_{\mathrm{C/W}}$ are close to zero, suggesting that Iberian and Hungarian populations may have split almost immediately after the initial divergence from the ancestral eastern population (Table 5.2).

5.2.3 Bayesian estimation of divergence times and ancestral population sizes

Minimal sampling

Bayes factor comparison of lnHML (Table 5.1) shows that the 'Out of the East' model fits the data significantly better then either of the alternative population tree topologies. The contrasting sets of priors *a* and *b* had little impact on posterior estimates of three of the four model parameters (Table 5.3, Fig.

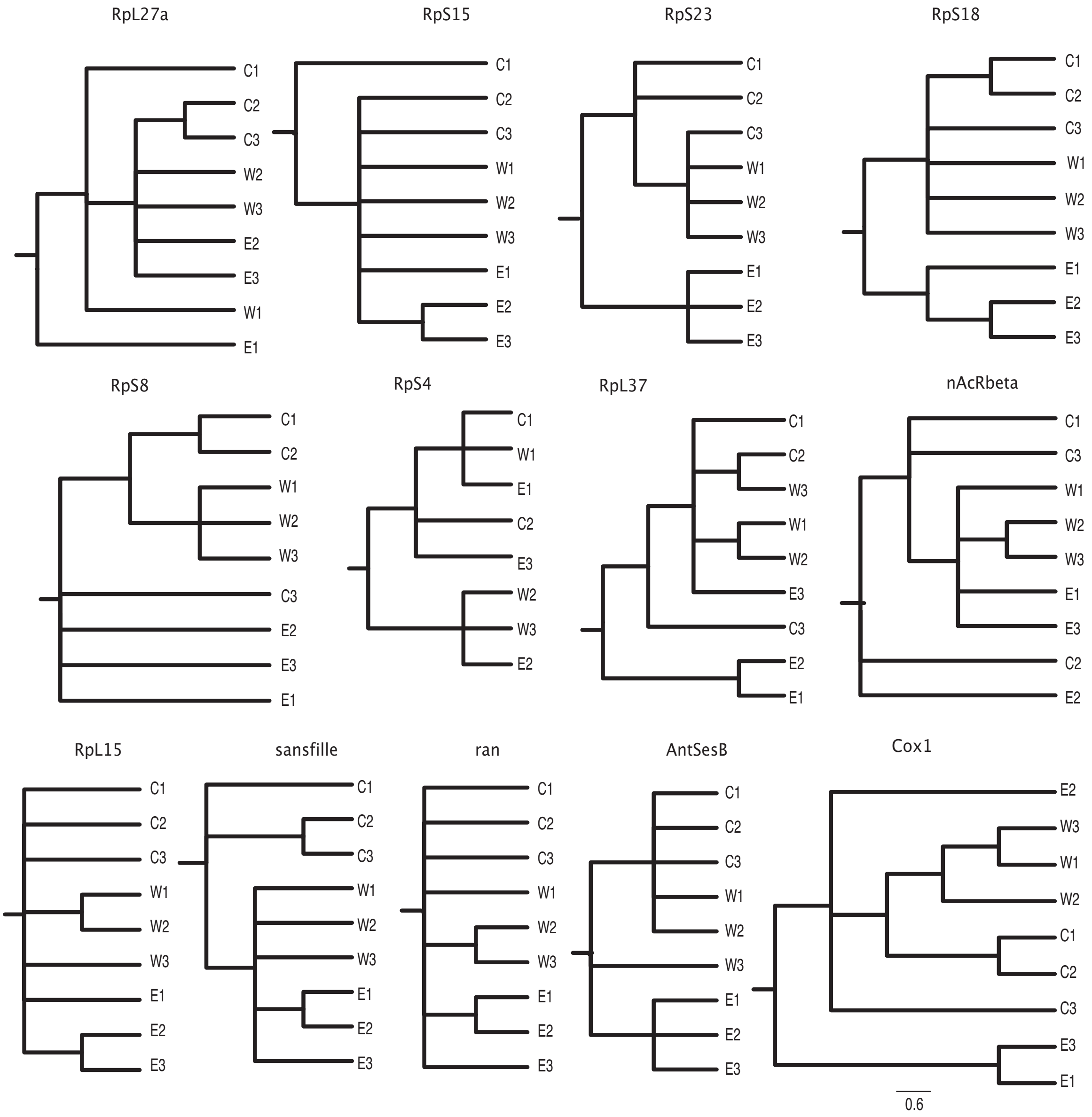

Figure 5.4: ML trees for the extended sampling of three individuals (labeled 1-3) per population for twelve nuclear loci and *Cox1* rooted using *C. lauta*. *RpL37a* is monomorphic and not shown. Although on average, samples from the same population are more closely related than those from different populations, there is extensive lineages sorting, resulting in a 'forest' of partially incongruent gene trees.

Table 5.2: Maximum Likelihood estimates (MLEs) of ancestral population sizes and population divergence times for refugial populations of *C. fungosa* assuming a population tree topology (E, (C, W)). Corresponding N_e and τ values are shown in brackets. The simplest mutational model assumes a single rate for all loci. In the variable rates analysis a relative mutation rate was computed for each locus from divergence to *C. lauta*.

	single rate (20 loci)	single rate (18 loci)	variable rates (18 loci)
$\theta_{\mathrm{E/C/W}}$ ($\mathbf{N}_{\mathrm{E/C/W}}$)	0.0076979 (**614,000**)	0.007995 (**637,000**)	0.008933 (**712,000**)
$\theta_{\mathrm{C/W}}$ ($\mathbf{N}_{\mathrm{C/W}}$)	0.000008 (**1000**)	0.000002 (**1000**)	0.000003 (**1000**)
$\gamma_{\mathrm{E/C/W}}$-$\gamma_{\mathrm{C/W}}$ ($\boldsymbol{\tau_{\mathrm{E/C/W}}}$-$\boldsymbol{\tau_{\mathrm{C/W}}}$)	0.0000032 (**0.001**)	0.000001 (**0.001**)	0.000001 (**0.001**)
$\gamma_{\mathrm{C/W}}$ ($\boldsymbol{\tau_{\mathrm{C/W}}}$)	0.0006924 (**0.110**)	0.000712 (**0.114**)	0.000756 (**0.121**)
lnL	-853.486	-794.948	-796.913

5.5A, B and D). Posterior mean ages for the split between eastern populations and the common ancestor of central and western populations $\tau_{\mathrm{E/C/W}}$ were 0.118 MYA and 0.134 MYA in analyses *a* and *b* respectively, with values of 0.043 MYA and 0.046 MYA for the divide between central and western populations $\tau_{\mathrm{C/W}}$ (Table 5.3). This comparatively long interval between the two divergence times ($\tau_{\mathrm{E/C/W}}$ - $\tau_{\mathrm{C/W}}$) is in apparent contrast to the results of the ML analysis. However, the 95% credibility intervals for the two divergence times overlap in both prior settings *a* and *b*, such that the lower confidence interval for $\tau_{\mathrm{E/C/W}}$ - $\tau_{\mathrm{C/W}}$ includes zero, compatible with divergence between western and central populations occurring immediately after the initial split from the ancestral eastern population. Likewise, the posterior estimate for the effective size of the population ancestral to all three refugial populations ($N_{\mathrm{E/C/W}}$) was minimally influenced by the prior (Table 5.3, Fig. 5.5D) (551,000 for *a* and 585,000 for *b*).

In contrast, posterior distributions for the effective size of the population ancestral to central and western populations, $N_{\mathrm{C/W}}$, differed considerably between prior settings *a* and *b* (197,000 and 698,000) (Table 5.3, Fig. 5.5C). $N_{\mathrm{C/W}}$ was also the parameter with the largest variance, the 95% credibility interval spanning two orders of magnitude (priors *b*, Table 5.3). Notably, with both prior settings, posterior distributions of $N_{\mathrm{C/W}}$ peak at the origin (Fig. 5.5C). This suggests that there is little information about $N_{\mathrm{C/W}}$ in the data, with posterior distributions largely reconstructing the prior.

To investigate to what extent $N_{\mathrm{C/W}}$ and the interval between population splits ($\tau_{\mathrm{E/C/W}}$ - $\tau_{\mathrm{C/W}}$) are confounded and whether this could account for the apparent difference in ML and Bayesian estimates of these parameters, we carried out a third MCMCcoal run (Table 5.3, priors *c*). When the prior mean for $N_{\mathrm{C/W}}$ is set to a very low value (2100), the posterior distribution for $\tau_{\mathrm{C/W}}$ shifts markedly towards the right (Fig. 5.5A) such that the two divergence events are estimated to have happened in close succession (0.091 and 0.089 MYA) in agreement with the ML results (Table 5.2).

Table 5.3: Prior and posterior means and 95% credibility intervals (CI) for divergence times and ancestral population sizes in Bayesian analyses using minimal sampling of a single individual per population and assuming an ‘Out of the East’ population tree topology (E, (C, W)). Corresponding N_e and τ values are shown in bold below. All analyses (*a*-*c*) assumed exponentially distributed priors (α = 1), but differed in their prior means .The population size inbetween the two divergence events, $N_{\mathrm{C/W}}$ is the parameter most sensitive to prior choice and has the widest confidence interval.

Parameter	(α, β)	Prior Mean (95% CI)	Posterior Mean (95% CI)
		priors *a*	
$\theta_{\mathrm{E/C/W}}$	(1, 380)	0.00271 (0.00011, 0.00968)	0.00691 (0.00239, 0.01830)
$\mathbf{N_{E/C/W}}$		**216,000 (10,000, 772,000)**	**551,000 (190,000, 1,459,000)**
$\theta_{\mathbf{C/W}}$	(1, 380)	0.00267 (0.00009, 0.00982)	0.002477 (0.00033, 0.00727)
$\mathbf{N_{C/W}}$		**213,000 (8,000, 783,000)**	**197,000 (26,000, 580,000)**
$\gamma_{\mathrm{E/C/W}}$	(1, 1519)	0.00095 (0.00012, 0.00276)	0.00074 (0.00019, 0.00139)
$\tau_{\mathbf{E/C/W}}$		**0.151 my (0.019 my, 0.440 my)**	**0.118 my, (0.030 my, 0.221 my)**
$\gamma_{\mathrm{C/W}}$	(1, 1519)	0.000329 (0.00001, 0.00119)	0.00027 (0.00001, 0.00076)
$\tau_{\mathbf{C/W}}$		**0.052 my, (0.002 my, 0.189 my)**	**0.043 my, (0.002 my, 0.121 my)**
		priors *b*	
$\theta_{\mathrm{E/C/W}}$	(1, 38)	0.02664 (0.00083, 0.09691)	0.00734 (0.00464, 0.01121)
$\mathbf{N_{E/C/W}}$		**2,124,000, (66,000, 7,726,000)**	**585,000 (370,000, 894,000)**
$\theta_{\mathbf{C/W}}$	(1, 38)	0.02639 (0.00064, 0.09669)	0.00875 (0.00050, 0.05260)
$\mathbf{N_{C/W}}$		**2,104,000 (51,000, 7,709,000)**	**698,000 (40,000, 4,141,000)**
$\gamma_{\mathrm{E/C/W}}$	(1, 152)	0.00980 (0.00113, 0.02918)	0.00084 (0.00023, 0.00156)
$\tau_{\mathbf{E/C/W}}$		**1.563 (0.180, 4.653) my**	**0.134 (0.037, 0.249) my**
$\gamma_{\mathrm{C/W}}$	(1, 152)	0.00326 (0.00008, 0.01198)	0.00029 (0.00001, 0.00084)
$\tau_{\mathbf{C/W}}$		**0.520 (0.131, 1.910) my**	**0.046 (0.002, 0.134) my**
		priors *c*	
$\theta_{\mathrm{E/C/W}}$	(1, 380)	0.00257 (0.00004, 0.00961)	0.00741 (0.00485, 0.01088)
$\mathbf{N_{E/C/W}}$		**205,000 (3,000, 766,000)**	**591,000, (387,000, 868,000)**
$\theta_{\mathbf{C/W}}$	(1, 38000)	0.00003 (0.00001, 0.00009	0.00005 (0.00001, 0.00015)
$\mathbf{N_{C/W}}$		**2,100 (1000, 7,000)**	**5,000, (1,000, 13,000)**
$\gamma_{\mathrm{E/C/W}}$	(1, 1519)	0.00096 (0.00011, 0.00277)	0.00057 (0.00011, 0.00111)
$\tau_{\mathbf{E/C/W}}$		**0.153 (0.017, 0.442) my**	**0.091 (0.018, 0.177) my**
$\gamma_{\mathrm{C/W}}$	(1, 1519)	0.00033 (0.00001, 0.00122)	0.00056 (0.00011, 0.00108)
$\tau_{\mathbf{C/W}}$		**0.053 (0.013, 0.195) my**	**0.089 (0.018, 0.172) my**

Table 5.4: Prior and posterior means and 95% credibility intervals (CI) for divergence times and ancestral population sizes in Bayesian analyses of extended sampling (20 loci, 13 sampled for three individuals per population) assuming an 'Out of the East' population tree topology (E, (C, W)). All analyses (a-c) assumed exponentially distributed priors (α =1), but differed in their prior means.

Parameter	(α, β)	Prior Mean (95% CI)	Posterior Mean (95% CI)
		priors *a*	
$\theta_{\mathrm{E/C/W}}$	(1, 380)	0.00263 (0.00007, 0.00968)	0.00793 (0.00552, 0.01089)
$\mathbf{N_{E/C/W}}$		**210,000 (6,000, 772,000)**	**632,000 (440,000, 867,000)**
$\theta_{\mathbf{C/W}}$	(1, 380)	0.00260 (0.00008, 0.00963)	0.00688 (0.00113, 0.01520)
$\mathbf{N_{C/W}}$		**207,000 (7,000, 768,000)**	**579,000 (95,000, 1,280,000)**
$\gamma_{\mathrm{E/C/W}}$	(1, 1519)	0.00098 (0.00011, 0.00287)	0.00084 (0.00040, 0.00136)
$\tau_{\mathbf{E/C/W}}$		**0.156 (0.018, 0.458) my**	**0.134 (0.064, 0.217) my**
$\gamma_{\mathrm{C/W}}$	(1, 1519)	0.00032 (0.00001, 0.00117)	0.00035 (0.00003, 0.00085)
$\tau_{\mathbf{C/W}}$		**0.051 (0.002, 0.187) my**	**0.056 (0.005, 0.136) my**
θ_{E}	(1, 380)	0.00267 (0.00010, 0.00973)	0.00481 (0.00203, 0.00979)
$\mathbf{N_{E}}$		**213,000 (8,000, 776,000)**	**383,000 (162,000, 772,000)**
θ_{C}	(1, 380)	0.00265 (0.00008, 0.00966)	0.00597 (0.00138, 0.0140)
$\mathbf{N_{C}}$		**212,000 (7,000, 770,000)**	**476,000 (110,000, 1,116,000)**
θ_{W}	(1, 380)	0.00267 (0.00008, 0.00974)	0.00226 (0.00027, 0.00615)
$\mathbf{N_{W}}$		**2113,000 (7,000, 776,000)**	**180,000 (22,000, 490,000)**
		priors *b*	
$\theta_{\mathrm{E/C/W}}$	(1, 38)	0.02693 (0.00082, 0.09876)	0.00771 (0.00533, 0.01071)
$\mathbf{N_{E/C/W}}$		**2,148,000, 65,000, 7,875,000)**	**615,000 (425,000, 845,000)**
$\theta_{\mathbf{C/W}}$	(1, 38)	0.02652 (0.00071, 0.09822)	0.01766 (0.00233, 0.06712)
$\mathbf{N_{C/W}}$		**2,115,000 (56,000, 7,832,000)**	**1,416,000 (186,000, 5,351,000)**
$\gamma_{\mathrm{E/C/W}}$	(1, 152)	0.00974 (0.00101, 0.02864)	0.00101 (0.00054, 0.00154)
$\tau_{\mathbf{E/C/W}}$		**0.155 (0.016, 0.457) my**	**0.161 (0.086, 0.246) my**
$\gamma_{\mathrm{C/W}}$	(1, 152)	0.00332 (0.00008, 0.01208)	0.00044 (0.00006, 0.00096)
$\tau_{\mathbf{C/W}}$		**0.053 (0.013, 0.193) my**	**0.070 (0.009, 0.154) my**
θ_{E}	(1, 380)	0.02653 (0.00058, 0.09794)	0.00872 (0.00282, 0.02646)
$\mathbf{N_{E}}$		**2,115,000 (46,000, 7,810,000)**	**695,000 (225,000, 2,110,000)**
θ_{C}	(1, 380)	0.02588 (0.00059, 0.09664)	0.02986 (0.00394, 0.09791)
$\mathbf{N_{C}}$		**2,064,000 (47,000, 7,706,000)**	**2,380,000 (314,000, 7,530,000)**
θ_{W}	(1, 380)	0.02642 (0.00081, 0.09737)	0.00433 (0.00046, 0.01911)
$\mathbf{N_{W}}$		**2,107,000 (65,00, 7,765,000)**	**345,000 (37,000, 1,524,000)**
		priors *c*	
$\theta_{\mathrm{E/C/W}}$	(1, 380)	0.00257 (0.00004, 0.00961)	0.00832 (0.00592, 0.01129)
$\mathbf{N_{E/C/W}}$		**205,000 (3,000, 766,000)**	**663,000 (472,000, 901,000)**
$\theta_{\mathbf{C/W}}$	(1, 38000)	0.00003 (0.00001, 0.00009)	0.00005 (0.00001, 0.00015)
$\mathbf{N_{C/W}}$		**2,000 (1000, 8,000)**	**4,000, (1000, 12,000)**
$\gamma_{\mathrm{E/C/W}}$	(1, 1519)	0.00096 (0.00011, 0.00277)	0.00069 (0.00032, 0.00110)
$\tau_{\mathbf{E/C/W}}$		**0.153 (0.017, 0.442) my**	**0.110 (0.051, 0.175) my**
$\gamma_{\mathrm{C/W}}$	(1, 1519)	0.00033 (0.00001, 0.00122)	0.00068 (0.00032, 0.00109)
$\tau_{\mathbf{C/W}}$		**0.053 (0.013, 0.195) my**	**0.108 (0.051, 0.174) my**
θ_{E}	(1, 380)	0.00266 (0.00009, 0.00975)	0.00444 (0.00179, 0.00939)
$\mathbf{N_{E}}$		**212,000 (8,000, 778,000)**	**354,000 (142,000, 749,000)**
θ_{C}	(1, 380)	0.00265 (0.00008, 0.00969)	0.00739 (0.00286, 0.01552)
$\mathbf{N_{C}}$		**212,000 (7,000, 773,000)**	**590,000 (228,000, 1,237,000)**
θ_{W}	(1, 380)	0.00267 (0.00008, 0.00971)	0.00343 (0.00130, 0.00755)
$\mathbf{N_{W}}$		**213,000 (7,000, 774,000)**	**274,000 (78,000, 602,000)**

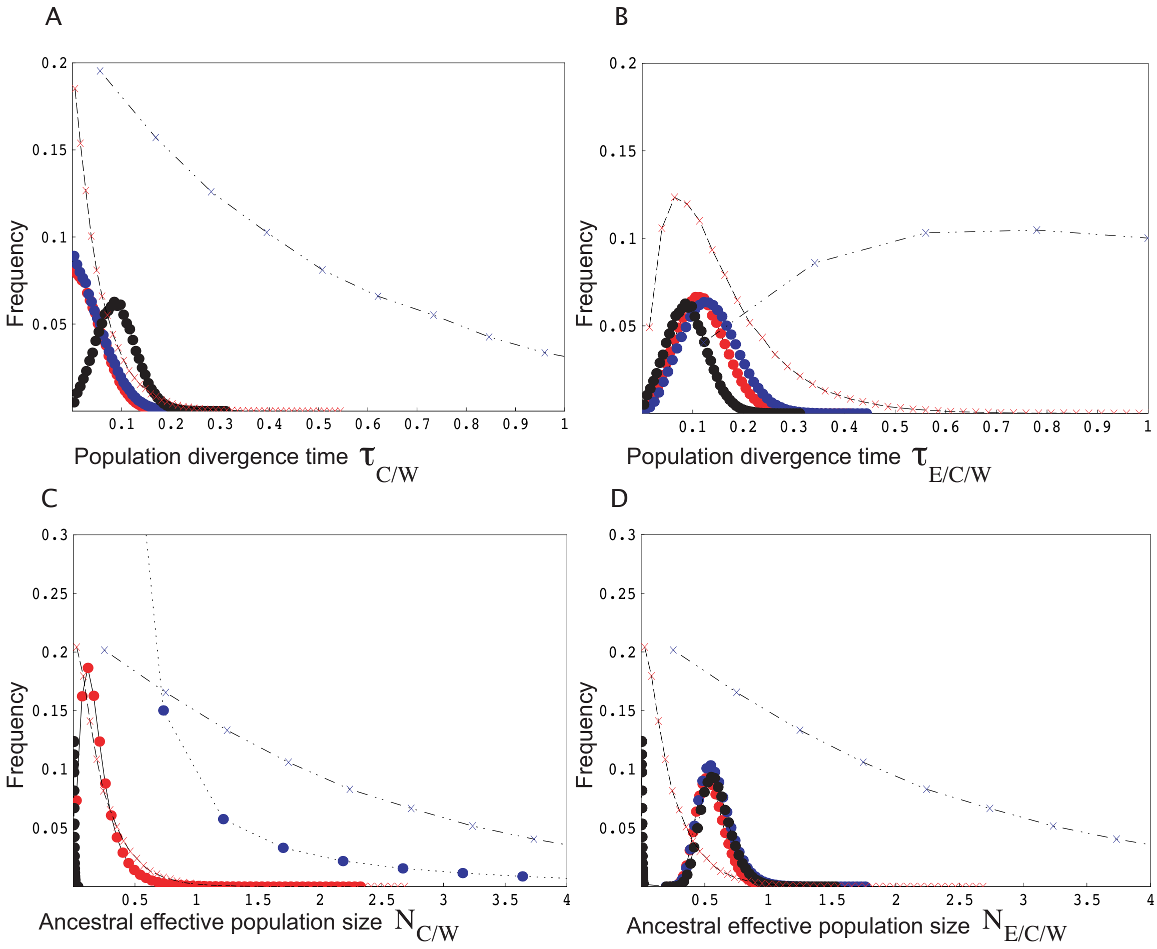

Figure 5.5: Prior and posterior distributions of parameters under the 'Out of the East' model of population divergence using minimal sampling of a single individual per population. Prior distributions for the first two MCMCcoal analyses are shown as dashed lines (a = mixed long and short dashes between blue symbols, b = long dashes between red symbols), posterior distributions for the single triplet analysis are in colour (a = red, b = blue and c = black). Whereas $\tau_{\text{E/C/W}}$ (B) and $N_{\text{E/C/W}}$ (D) are little influenced by the prior means, $N_{\text{C/W}}$ (C) is extremely sensitive. This parameter is also confounded with $\tau_{\text{C/W}}$. When setting a low prior mean for $N_{\text{C/W}}$ (analysis c) the posterior distribution for $\tau_{\text{C/W}}$ shifts markedly towards the right (see black line in A). Note that despite $\alpha = 1$ for all model parameters, the prior distribution for $\tau_{\text{E/C/W}}$ (B) is not exponential because of the constraint $\tau_{\text{E/C/W}} > \tau_{\text{C/W}}$.

Extended (three individual) sampling

MCMCcoal analyses of the extended (three individual per population) dataset again gave strongest support to the ‘Out of the East’ scenario (Table 5.1). While Bayes factor comparison strongly rejects the ‘Out of the West’ topology (W, (C, E)), the second alternative topology (C, (E, W)) does not provide a significantly worse fit to the data (Table 5.1).

Parameter estimates agree well with those obtained when only a single individual per population was sampled (Table 5.4 and Fig. 5.6). However, increased sampling does have some influence on parameter estimation. First, estimates of $N_{\mathrm{C/W}}$, are larger and less sensitive to prior settings when three individuals per population are sampled for both prior sets *a* and *b* (Supporting Information Table S3). Second, the posterior distributions for $\tau_{\mathrm{C/W}}$ are now unimodal, rather than L-shaped with a maximum at the origin (Fig. 5.6). However, this has little impact on the variance of the posterior. For example, the 95% credibility interval for $\tau_{\mathrm{C/W}}$ is 0.005 - 0.136 MYA (priors *a*) in the analysis of the extended samples of three individuals per population, compared with 0.002 - 0.121 MYA when sampling a single individual (Table 5.3).Taken together this suggests that increasing sample size per population to three haploid individuals adds some, but not much, power to the estimation of model parameters.

Sampling multiple individuals per population we can also estimate the effective sizes of the three sampled populations between the present and the first divergence events, N_{E}, N_{C}, N_{W}. (Table 5.4). Although estimates of these parameters had fairly wide confidence intervals and were sensitive to prior settings, their relative magnitude was consistent across analyses. N_{C} was always the largest followed by N_{E} and N_{W}. It is also noteworthy that all three estimates were smaller than those obtained for ancestral populations, paralleling the findings of Jennings and Edwards 2005 and previous results in Great Ape studies (Chen & Li, 2001; Yang, 2002; Patterson *et al.*, 2006).

5.2.4 Gene divergence times

Following Jennings & Edwards (2005), we calculated Jukes Cantor distances (D) to estimate coalescence times for each divergence event (D/2) and compared the average distance across loci with the estimated population divergence time and the mitochondrial (*Cox1*) node ages for both single and three individual samples. In both cases nuclear genes sampled from central and western populations diverged on average almost 0.4 million years (or three glacial periods) prior to the estimated population divergence (Fig. 5.7). Coalescence times estimated for *Cox1* depend on the assumed mutation rate. Applying the calibration by Oliveira *et al.* (2008) both coalescence times for *Cox1* (0.013 MY and 0.145 MY respectively) are younger than the average coalescence at nuclear genes but are well within the 95% credibility interval

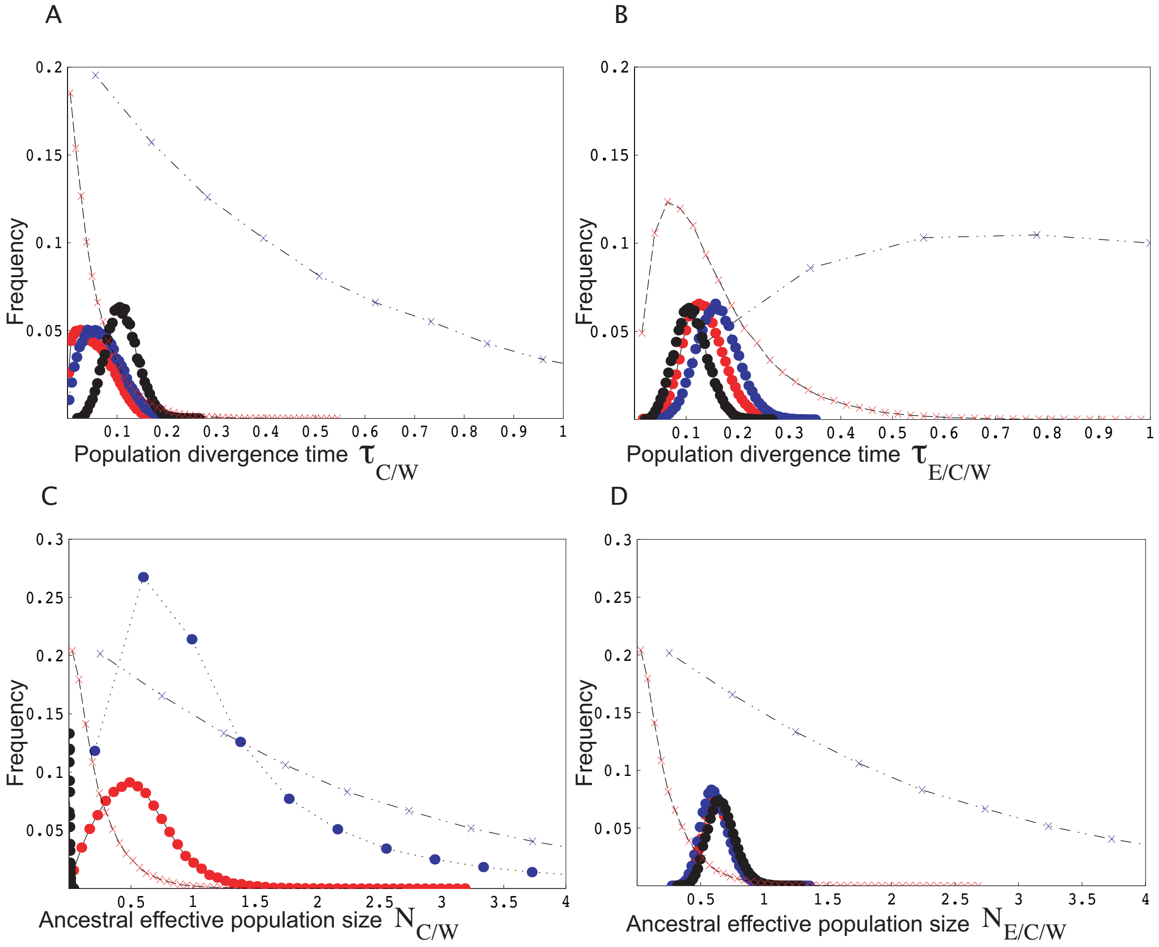

Figure 5.6: Prior and posterior distributions of model parameters under the 'Out of the East' scenario of population history obtained for the extended sampling (20 loci, 13 sampled for three individuals per population). Prior distributions a and b are shown as dashed lines (a = narrow, b = wide), posterior distributions are in colour (a = red, b = blue and c = black). Both $\tau_{\mathrm{C/W}}$ (A) and $\tau_{\mathrm{E/C/W}}$ (B) are little influenced by the prior means. Note that in comparison with figure 4, the maxima for the posterior distributions for $N_{\mathrm{C/W}}$ are > 0.

of the estimated population divergence (Table 5.3). Using Brower (1994), mitochondrial coalescence between the ancestor of central and western samples and the eastern sample (1.433 MYA) predates the average coalescence times for nuclear genes (0.714 MYA), whereas the mitochondrial coalescence time between central and western samples (0.125 MYA) is still more recent than that for nuclear genes (0.467 MYA) (Fig. 5.7).

5.3 Discussion

We analyzed a large multilocus dataset under the simplest possible model of divergence between three populations to make quantitative inferences about the longitudinal history of *C. fungosa*. Reconstructing the genealogical histories of individual loci leads to a 'forest' of largely incongruent and often poorly resolved gene trees (Fig. 5.4), which individually contain little information about the underlying population history. However, analyzing these data jointly in a coalescent framework, the relationship between major refugial populations of *C. fungosa* can be described as a quantified population tree, which includes relevant population genetic parameters (Fig. 5.8). This is a considerable improvement over previous phylogeographic studies in this system, which have largely been based on mitochondrial sequence data and allozymes (Rokas *et al.*, 2001, 2003; Stone *et al.*, 2001; Challis *et al.*, 2007; Stone *et al.*, 2009) and allows us to quantify important aspects of the phylogeographic history of *C. fungosa*.

First, both likelihood and Bayes factor comparisons of population tree topologies (Table 5.1) support the 'Out of the East' scenario for *C. fungosa*.

Second, both ML and Bayesian estimates for the time of the first population split between the eastern population and the common ancestral population of central and western populations $\tau_{\mathrm{E/C/W}}$ fall well within the late Pleistocene. Likewise, both methods suggest that the more recent divergence between central and western populations ($\tau_{\mathrm{C/W}}$) occurred either during the last interglacial or glacial period. However, since the MLE for the time between population splits ($\tau_{\mathrm{E/C/W}}$ -$\tau_{\mathrm{C/W}}$) is effectively zero and the 95% credibility intervals for the two divergence times overlap in all Bayesian analyses, we cannot exclude the possibility that the two population splits happened in close succession.

Finally, the present coalescent analyses provide information about the effective sizes of ancestral and present populations. Although our estimates of both ancestral population sizes, in particular $N_{\mathrm{C/W}}$, have large confidence intervals and, in the case of $N_{\mathrm{C/W}}$, are sensitive to prior settings (discussed below), they provide an important comparison with model organisms. For example the observed diversity in *C. fungosa* $\pi_s = 0.92\%$, Table 4.3) is comparable with that in non-African populations of *D. melanogaster* ($\pi_s = 1.33\%$) (e.g. Andolfatto, 2001, Table 3). Similarly, estimates for the effective population sizes

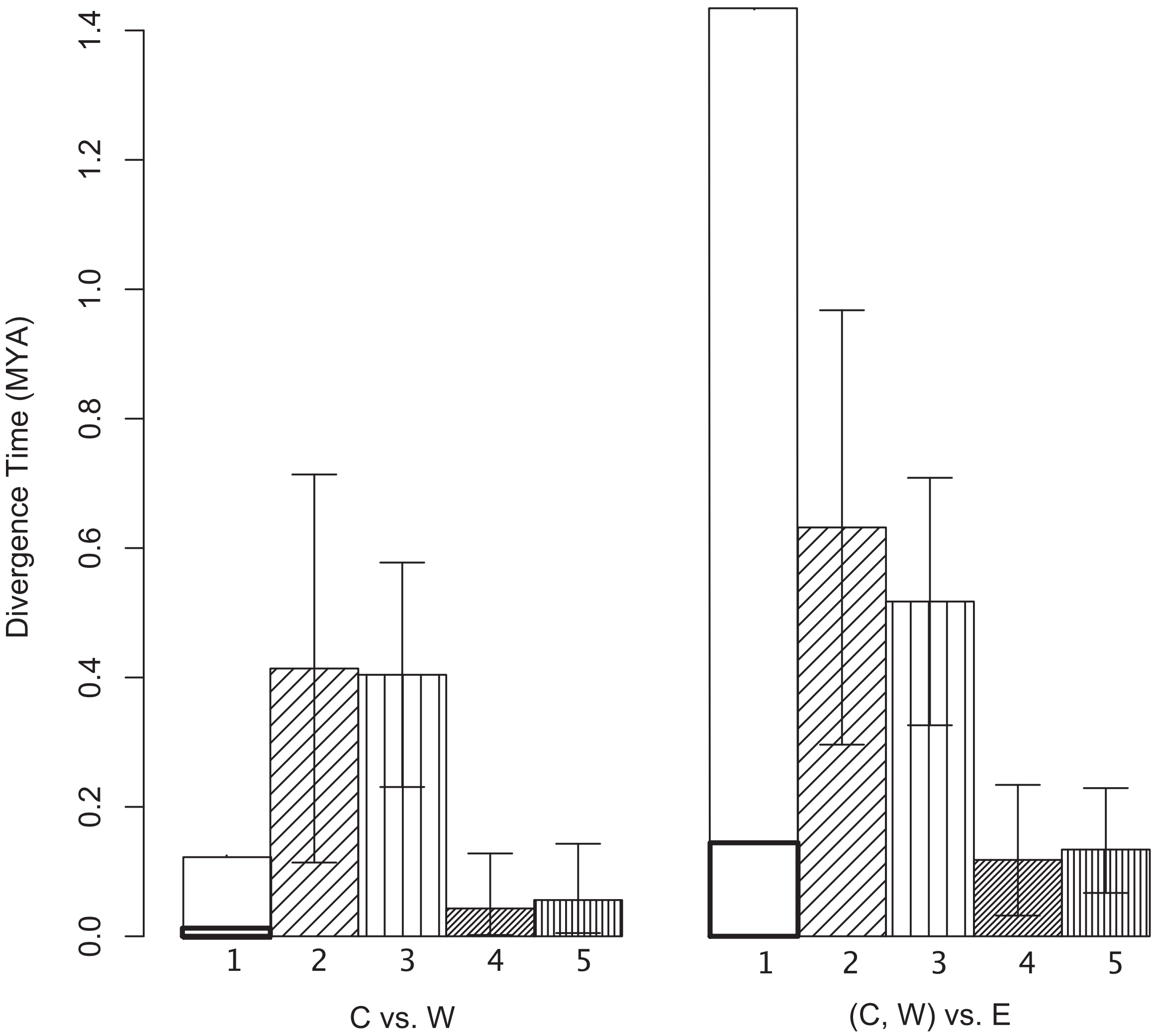

Figure 5.7: Divergence times for the two splits in the ‘Out of the East’ model (C vs. W left and (C,W) vs. E right). The figure shows that Bayesian estimates (prior settings a) of population divergence times for both single and extended triplet samples (columns 4 and 5 in each figure respectively) are more recent than the mean coalescence time across nuclear loci for both sampling schemes (columns 2 and 3 in each figure). Mitochondrial divergence (column 1) was calculated from node ages in the single triplet tree using both the rate of Oliveira *et al.* (2008) rate calibrated from Nasonia sister species (lower estimates, bold bars in column 1) and the widely applied rate estimate of Brower (1994) (higher estimates, column 1). Error bars show 95% confidence limits.

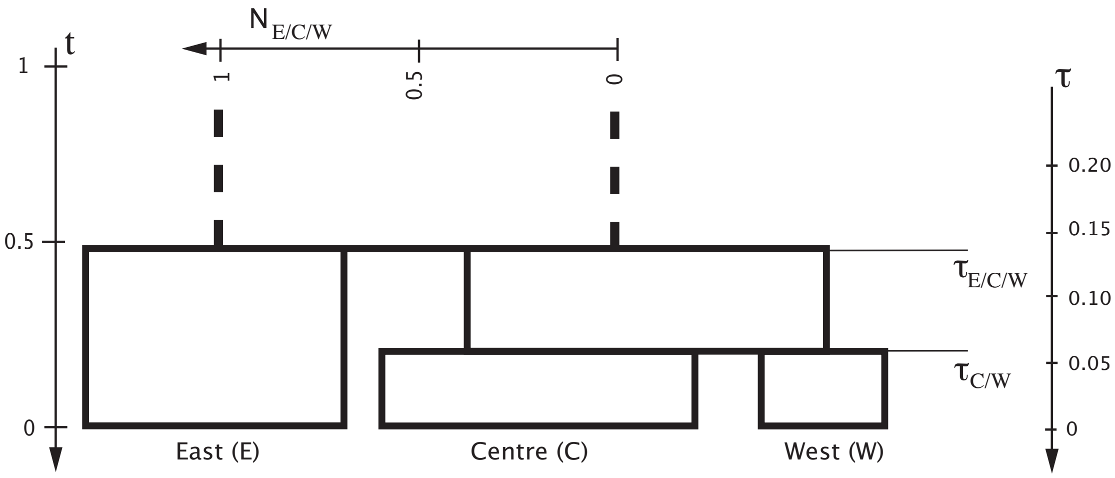

Figure 5.8: Population tree for Western Palearctic *C. fungosa* inferred from 20 genetrees. Means of posterior distributions of model parameters were obtained from the Bayesian analysis (priors a, extended sampling of three sequences per population,Table 5.4 and figure 5.6). The widths of blocks correspond to effective population sizes (scale at top). Divergence times are shown on two different scales: τ in MY (right hand scale), and $t = \tau/2N_{\text{E/C/W}}$ generations assuming two generations/yr, i.e. $g = 0.5$ (left hand scale). Note that all blocks have a greater width than height such that pairs of lineages sampled from the same population are more likely to coalesce in their ancestral population.

of *D. melanogaster* of 10^6 (Andolfatto & Przeworski, 2000) and for effective size of the ancestor of *D. melanogaster* and *D. simulans* of $Ng = 3.9 \times 10^5$ (Li *et al.*, 1999) agree with our results for *C. fungosa* in order of magnitude. If effective population sizes of 10^6 are the rule in insect parasitoids, their longitudinal histories will inevitably involve extensive incomplete lineage sorting, strengthening the case for multilocus approaches for meaningful phylogeographic inferences.

How do these results compare with those obtained from single gene trees both in *C. fungosa* and in other co-distributed oak gall parasitoids and their hosts? In *C. fungosa* the topology of the inferred population tree (Fig. 5.8) is congruent with both the majority of resolved nuclear gene trees as well as the mitochondrial gene tree when a single individual per refugial population was sampled (Fig. 5.3). More generally, the eastern origin of *C. fungosa* is consistent with the mitochondrial gene tree for another oak gall parasitoid, *Megastigmus stigmatizans* (Hayward & Stone, 2006), with mitochondrial and nuclear gene trees in the parasitoid *Megastigmus dorsalis* (Nicholls *et al.*, 2010) and three species of host gall wasps (Rokas *et al.*, 2003; Challis *et al.*, 2007; Stone *et al.*, 2007, 2009).

While by definition gene divergence must predate the divergence of populations, our results suggest that the magnitude of this difference is considerable in *C. fungosa* and very relevant for our interpretation of its Pleistocene history. It is noteworthy that the estimates for $\tau_{\text{E/C/W}}$ coincide with the last (Eemian) interglacial 0.130 - 0.115 MYA, which suggests that divergence between refugial populations is as recent as it possibly can be (given the definition of glacial refugia). We know from the fossil record that both oaks (Velichko *et al.*, 2005) and associated gall wasps species (van der Ham *et al.*, 2008; Stone *et al.*, 2008) known to be attacked by *Cecidostiba* expanded their range in Central and Northern Europe during this period. It is thus plausible for population divergences associated with westward range expansions of *C. fungosa* to have occurred over a similar timescale.

Although the unknown error in the mitochondrial clock and the large discrepancy between different calibrations (Brower, 1994; Oliveira *et al.*, 2008) make a direct comparison with mitochondrial dates problematic, it is nevertheless reassuring that the mitochondrial ages obtained for *C. fungosa* fall within the 95% credibility interval of (Oliveira *et al.*, 2008) or predate (Brower, 1994) the estimated time of population divergence (Fig. 5.7), as they should. A mitochondrial divergence more recent than that inferred for the population would be inconsistent with the assumed model, and require gene flow between populations. However, it is noteworthy that regardless of the mitochondrial mutation rate used, the *Cox1* divergence times are very different from the average divergence times at nuclear genes (Fig. 5.7). This demonstrates the extremely large variance in coalescence times and highlights the danger of over-interpreting node ages of single gene trees. An additional problem with mitochondrial mutation rate calibrations is that they are likely to be confounded by the selective dynamics of bacterial endosymbionts (Oliveira *et al.*, 2008), the prevalence of which is known to differ both between populations and closely related species of Pteromalids (Weinert *et al.*, 2009). It is therefore not clear to what extent the *Nasonia* rate applies to *C. fungosa*. In contrast, the nuclear estimates for *Nasonia* are broadly consistent with those obtained for other Insects.

The fact that divergence at a single locus can only provide an upper bound of the population divergence time may well explain why mitochondrial dates found in previous studies on other species of European gall parasitoids and their gall wasp hosts (Hayward & Stone, 2006) are considerably older than the population divergence estimates for *C. fungosa* obtained here. For instance, mitochondrial divergence between Central European and Iberian clades of the parasitoid *Megastigmus stigmatizans* has been estimated at 0.264 MYA (Hayward & Stone, 2006). Mitochondrial divergence estimates between Central Europe and Iberia for gall wasp host species are still older; e.g. 0.383 MYA in *Andricus kollari* (Hayward & Stone, 2006) and 1.6 MYA in *Andricus coriarius sensu stricto* (Challis *et al.*, 2007). Analyses of multilocus datasets are clearly required to provide better estimates of population divergence times in these species. As our results show, the fact that the variance in coalescence time is lower for mitochondrial loci given their smaller N_e may reduce but does not alleviate this problem. This underlines the possibility raised by Nichols (2001) that between-taxon variation in mtDNA-inferred dates of divergence between glacial refugia may well be attributable to coalescent variance rather than taxon-specific differences in post-glacial dispersal. Rigorous testing of the hypothesis of taxon-specific variation in divergence times requires broader application of multilocus approaches.

5.3.1 Ancestral N_e and sampling

The results of the Bayesian analyses show that estimates of $\tau_{\text{C/W}}$, or rather the time between the population splits ($\tau_{\text{E/C/W}}$-$\tau_{\text{C/W}}$) and the population size during that time, $N_{\text{C/W}}$, are confounded. Considering that it is the ratio of the two parameters which determines the chance of coalescence between population splits (Hudson, 1983; Saitou & Nei, 1986; Yang, 2002), this makes intuitive sense and may explain the poor ability to estimate $N_{\text{C/W}}$ independently. A large variance in ancestral N_e has also been reported by most earlier multilocus analyses of divergence models (Chen & Li, 2001; Yang, 2002; Rannala & Yang, 2003). In general, explanations for the low power to estimate this parameter fall into two categories: violations of the model assumptions; and limited signal in the data.

Ignoring within-locus recombination and mutational rate heterogeneity, for example, can in principle overestimate ancestral population sizes (Satta *et al.*, 2000; Yang, 2002; Wall, 2003). However, the few studies that have incorporated these factors suggest that they have little influence on estimates of ancestral N_e (Satta *et al.*, 2000; Yang, 2002; Wall, 2003). Similarly, the fact that our ML results for the variable mutation model are in agreement with those assuming a single rate despite large differences in relative mutation rates (Table 4.4) suggests that any impact of mutational heterogeneity between loci is greatly outweighed by coalescence and mutational variance and therefore an unlikely explanation for the low power to estimate $N_{\text{C/W}}$.

In general, there are two factors that determine statistical power to infer ancestral parameters; i) the number of lineages that contribute to the estimate (Fig. 5.2) and ii) the mutational information available to infer their relationships. Both clearly depend on the timescale of divergence. Relating the estimated population divergence times (scaled by the mean of current population sizes) for *C. fungosa* to the theoretical expectation for the number of surviving lineages, we can ask how much power could potentially be gained by further increasing sample sizes. For example, figure 5.2 shows that sampling three instead of a single individual per population roughly doubles the expected number of eastern lineages that survive into the common ancestral population, while 16 more individuals are required for a further twofold increase. For the more recent divergence at $\tau_{\text{C/W}}$, the increase in the number of surviving lineages from additional samples is of course more substantial (Fig. 5.2). However, if our analysis was limited by sample size, we would expect to see an improvement in parameter estimation proportional to the increase in the number of surviving lineages when sampling three individuals. The fact that this is not the case (i.e. the variance in the estimates of three of the four model parameters is little affected despite the doubling of surviving lineages) suggests that the power to infer ancestral parameters is largely limited by the mutational variation available rather than the sample size. However, our finding of a markedly higher posterior mean $N_{\mathrm{C/W}}$ for the three individual sampling suggests that the estimation of this parameter may indeed be sensitive to the sample size. This makes intuitive sense if we extend the 'number of surviving lineages' argument above and consider that only lineages which survive into $N_{\mathrm{C/W}}$ **and** coalesce before they reach $N_{\mathrm{E/C/W}}$ contribute to the estimate of $N_{\mathrm{C/W}}$. One would therefore expect increased power to estimate this parameter with increasing sample sizes both in *C. fungosa* and in the bird divergence studied by Jennings and Edwards 2005. Thorough investigation of the effect of sampling on statistical power in divergence models both theoretically and using empirical data is required to inform sample designs of future population genetic and phylogeographic studies. In particular disentangling the effects of mutational limitation and those of sample size (both the number of sampled loci and individuals) would be useful. If mutational information is not limiting, gene tree - species tree methods (Degnan & Salter, 1995; Degnan & Rosenberg, 2009; Maddison & Knowles, 2006; Liu & Pearl, 2007; Kubatko *et al.*, 2009) should converge to the same answer as the inference methods used here.

Another way to improve power may be to use outgroup information in the likelihood calculation. At present Ne3sML and MCMCcoal rely on clock rooting (Yang, 2002), which, given the small number of polymorphic sites in some loci, results in large topological uncertainty. Being able to distinguish between parsimony informative sites and singleton mutations by reference to an outgroup should in principle enhance the power of both approaches.

5.3.2 Assumptions and extensions of the model

Considering the large confidence intervals in parameter estimates, it is clear that quantitative inference of population history is a data-hungry problem, particularly if divergence is recent. It is therefore questionable how much scope there is to probe more realistic models without increasing the amount of data drastically. In general, inferences of ancestral population parameters are likely to be much more sensitive to violations of the divergence model than they are to violations of the model of sequence evolution. Since there are key population processes omitted from the present analyses that render population history less tree-like, one could argue that the notion of a 'population tree' as such is an unrealistic description of phylogeographic history.

Firstly, the model assumes that there is no migration after divergence. While at least in the host gallwasps, allele frequency data support this assumption (Rokas *et al.*, 2001, 2003; Stone *et al.*, 2001, 2008; Challis *et al.*, 2007), we cannot exclude the possibility of migration after divergence for *C. fungosa*. It would therefore be interesting to relax this assumption and IMa, which uses the algorithm of MCMCcoal, has recently been extended to estimate divergence with migration for more than two populations (Hey, 2010b). However, modelling migration explicitly in a three-population model introduces six additional parameters. Considering the low divergence between *C. fungosa* populations for our loci, there would appear to be little power in the data to distinguish between a divergence model with a very recent split as inferred here and more complicated models involving both divergence and subsequent gene flow. Clearly, much larger amounts of data are needed to successfully explore such models. An additional problem with analysing models of migration is that, in contrast to strict divergence models, they are sensitive to unsampled populations (Wilkinson-Herbots, 2008; Lohse, 2009). With the advent of nextgen sequencing technologies, the volumes of data required to explore divergence with gene flow on such recent timescales should soon be available.

Secondly, the model assumes constant population sizes between divergence events. Again, allowing for changes in population size opens up a myriad of possible historical scenarios and potentially increases the number of parameters dramatically. Fortunately however, the *C. fungosa* data allow us to at least exclude drastic demographic events. For instance, under a model of colonization through extreme founder events (without subsequent migration), widespread incongruence between gene trees and population trees would not be expected. Thus the mere presence of all possible gene tree topologies in our data allows us to reject this scenario for *C. fungosa*.

And finally, the model assumes panmixia within populations, which may be unrealistic over short timescales and large geographic areas. Recent theoretical work (Slatkin & Pollack, 2008) and simulations (Becquet & Przeworski, 2009) have demonstrated that subdivision in ancestral populations can lead to mis-inference under simple divergence models.

In general, any model-based analysis faces the challenge of choosing models that contain sufficient realism to capture key features in the data whilst being simple enough to be useful. We have shown that in the case of *C. fungosa* a simple divergence model between three populations can explain the observed genetree incongruence and be used to estimate both the origin and divergence time of refugial populations despite the recency of this history. We hope that this study motivates similar analyses of more realistic models.

5.3.3 Towards a multilocus approach to community phylogeography

The close ecological dependence of oak gall parasitoids on their hosts and the large number of species involved make this and similar host-parasitoid communities valuable systems in which to study the evolution of ecological interactions (Schönrogge *et al.*, 1995; Hayward & Stone, 2005). Unlike most organisms for which similar multilocus analyses have been conducted (Li *et al.*, 1999; Rannala & Yang, 2003; Jennings & Edwards, 2005), the ecology of chalcidoid parasitoids involves intricate interactions with co-distributed species at different trophic levels. Linking the extensive information on species composition and food web structure for these communities (Schönrogge *et al.*, 1995, 1996a; Bailey *et al.*, 2009) with population genetic and phylogeographic inferences at the species level opens up an exciting opportunity to address novel and general questions about co-evolution and assembly of communities. For instance, do particular lineages or guilds within trophic levels show earlier longitudinal range expansion than others? And if so, what are the ecological properties of such species? For example, are they generalists rather than specialists, and so less likely to go locally extinct (Hayward & Stone, 2006)? Further questions arise when considering multiple trophic levels. How correlated are phylogeographic histories between hosts and parasitoids? Is there a general lag between the arrival of gallwasp (or other herbivore) hosts and associated parasitoids such that herbivores experience periods of enemy-free space (Hayward & Stone, 2006)? We are currently working on obtaining multilocus data for co-distributed chalcidoid parasitoid species and their gallwasp hosts to address these questions in a quantitative framework. The rarity of many of the species involved (Schönrogge *et al.*, 1995, 1996a,b, 1998) means that we will have to make the most of small sample sizes.

Chapter 6

Topological probabilities in models of divergence with gene flow

It s well known that the topology of a neutral locus sampled from closely related populations or species may be incongruent with the order of divergence of those populations (Hudson, 1983; Tajima, 1983; Nichols, 2001; Pamilo & Nei, 1988). In the simplest case of divergence between three species or populations (A, B, C) with population tree topology (A(B,C), and divergence at τ_1 and $\tau_1 + \tau_0$ (i.e. the model analysed in the previous chapter), the genealogy of a triplet sample (i.e. a single individual taken from each population), may have three possible topologies $(a(bc))$, $(c(ab))$ and $(b(ac))$ (Fig. 6.1a). Their probabilities depend on the interval between population splits on the coalescence time scale (Hudson, 1983; Tajima, 1983; Takahata *et al.*, 1995), i.e. $T_0 = \tau_0/(2N_e)$. This is because incongruent topologies are only possible if the b and c lineage survive interval τ_0 without coalescence the chance of which is e^{-T_0}. Once all lineages find themselves in the common ancestral population, each topology has the same chance $1/3$. Therefore:

$$P_{(c(ab))} = P_{(b(ac))} = \frac{1}{3}e^{-T_0} \tag{6.1}$$

Full results for the joint probability of topologies and branch lengths under this simple divergence model have been derived by Yang (2002) and, assuming infinite sites mutations (Kimura, 1969), can be used to calculate the marginal likelihood of model parameters from patterns of sequence polymorphism in a set of loci (see chapter 5). Because this assumes free recombination between loci but lack of recombination within them and computation time increases linearly with the number of loci, analyses using this full theory are usually restricted to moderate numbers of loci of relatively short length. Alternatively, model parameters may be estimated from patterns of diversity on a genome wide scale. In particular, the product of the probability of a topology and its expected internal branch length leads to an expression for the expected number of shared derived mutations (or parsimony informative sites) corresponding to it which in turn can be used to compute point estimates of ancestral parameters from genomic triplet alignments. The idea of using genome wide site counts to estimate divergence times and ancestral population sizes was first put forward by Patterson *et al.* (2006) who studied the divergence between humans, chimpanzees and gorillas.

A key feature of the three population divergence model is the symmetry of the two incongruent histories. Because incongruences can only arise if lineages survive into the common ancestral population which is assumed to be panmictic, the two possible incongruent topologies must be equiprobable and hence their expected frequencies are the same (eq. 6.1). This symmetry is a consequence only of the assumed exchangeability of lineages in the ancestral populations and is independent of their effective population sizes. Furthermore, the symmetry extends to the branch length distributions, which also is the same for $(c(ab))$ and $(b(ac))$ genealogies. Thus, in polarized (outgroup rooted) triplet alignments, derived mutations shared by a and b (ab sites) and a and c (ac sites), i.e. those corresponding to the internal branches of the two incongruent topologies, have the same expected frequency.

Perhaps surprisingly, the two studies that have explicitly investigated genome wide frequencies of either site counts and/or gene tree topologies in interspecific triplets, have both found significant asymmetries (Patterson *et al.*, 2006; Pollard *et al.*, 2006). Patterson *et al.* (2006) have counted site-types in genomic data of human, chimpanzee and gorilla (rooted with orang-utan). While they observed no significant difference in the number of derived mutations shared by human and gorilla (HG) compared to chimpanzee-gorilla (CG) sites on the autosome, they found a slight excess of HG over CG sites (3,074 , 26.2% vs. 2544, 21.8%) in a 964 kb region on the X-chromosome. Similarly, Pollard *et al.* (2006) studied topologies of close to 10 000 genes in a triplet of closely related species of *Drosophila*: *D. melanoaster, D. erecta* and *D. yakuba* and found a significant excess of (Dmel,Dere) (23.5%) over (Dmel,Dyak) (18.7%) gene tree topologies. Moreover this asymmetry is particularly convincing, given that it is found in other character types, in particular indels, nucleotide and amino acid replacements (Pollard *et al.*, 2006, Fig. 2).

These findings beg the question how such asymmetries can arise. Firstly, it is important to note that assessing the significance of asymmetries in genome-wide data involves making assumptions about genetic linkage, which tends to increase the variance in topological frequency (Pollard *et al.*, 2006). Two possible causes have been suggested: i) Sequencing error (Burgess & Yang, 2008); and, perhaps biologically more interesting, ii) violations of the simple divergence model.

Slatkin & Pollack (2008) have shown that certain types of structure in the ancestral population can create asymmetries in the frequency of the two incongruent topologies. They propose a divergence model with a barrier that persists from the common ancestral population until the most recent population split and coincides with the diverging populations (Slatkin & Pollack, 2008, Fig. 1) and derive the topological probabilities under this model. Using results for the expected coalescence times for the two incongruent topologies, they show that the asymmetries in topological frequencies observed by Pollard *et al.* (2006) can be explained by a very weak barrier with migration at rate $2Nm = 9.4$ across it. However, an obvious alternative mechanism by which topological asymmetries can arise is migration between the populations themselves. The aim of this chapter is to investigate the effect such gene flow after divergence has on topological probabilities.

6.1 Model and derivation

In the following, the basic model of divergence between three populations is extended by allowing for gene flow (at rate m per generation) between the older population (A) and one of the more recently diverged populations (B and C) (it is intuitively clear that gene flow involving B and C cannot create asymmetries) (Fig. 6.1). Although this model arguably represents a special case, it is the simplest divergence model in which topological asymmetries due to gene flow can arise. It also applies to some datasets of particular interest. For example, on an intraspecific scale, unidirectional gene flow may be a realistic scenario for many European taxa such as *C. fungosa* which have colonised major Southern refugia in a process of longitudinal range expansions possibly followed by continued gene flow from their eastern centres of diversity (chapter 5). The model may also be used to describe hybridization between European *Homo sapiens* and Neanderthals in Europe and/or Asia (with African *H. sapiens* populations as the ingroup) (Green *et al.*, 2010).

The aim is to derive the probabilities of the three genealogical topologies $P_{(c(ab))}$, $P_{(b(ac))}$ and $P_{(a(bc))}$ under this model. To keep the number of parameters at a minimum, we will assume that population sizes are equal and constant at all times. Furthermore we make all the standard simplifying assumptions of the neutral coalescent, namely large effective population sizes and panmixis and focus on two simple scenarios i) migration in one direction (from A to B) only and ii) symmetric migration (Fig. 6.1b and c).

Following Slatkin & Pollack (2008), the ancestry of a sample between population divergence events can be described as a discrete time Markov chain with state transitions occurring either due to migration of lineages between populations (at rate m) or coalescence of pairs of lineages (at rate $\lambda = 1/2N_e$) per generation. The divergence of populations can be modeled as a sudden change in state space. An

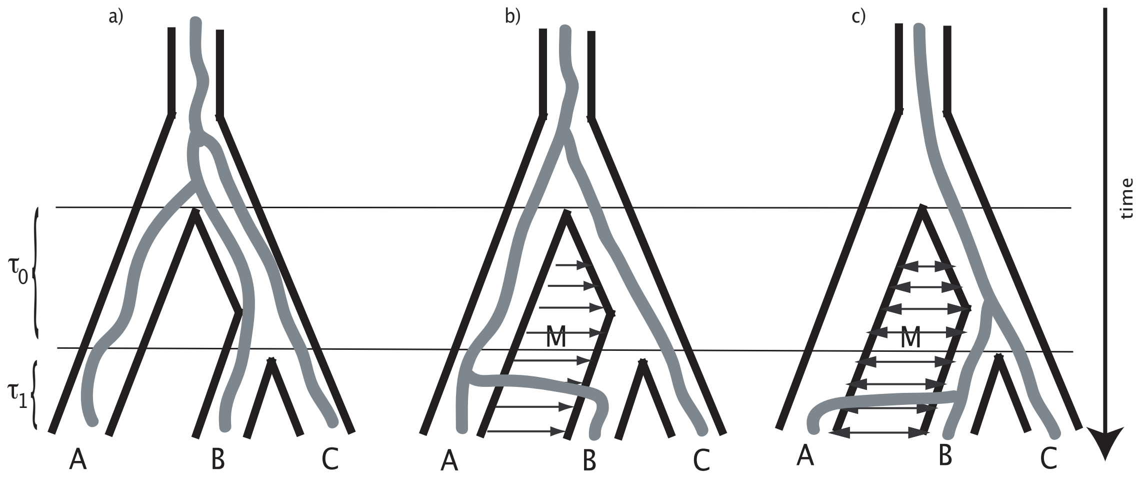

Figure 6.1: Under a simple model of divergence between three populations A, B, C with a topology $(A(BC))$ (a), incongruent genealogies (shown in grey) can arise due to incomplete lineage sorting in the common ancestral population prior to $(\tau_1 + \tau_0)$. With unidirectional (b) or symmetric (c) migration between A an B the probability of incongruent genealogies with topology $(c(ab))$ is increased. In both cases, the asymmetry in the probability of the two incongruent topologies arises as a result of a migration event during τ_1 only. Note that $(c(ab))$ genealogies are expected to be much shorter in c) than in b).

analogous approach has previously been used to find the probability of topologies for a pair of linked loci in a three population model without migration (Slatkin & Pollack, 2006).

6.1.1 Asymmetric migration

Below we consider the case of migration from A to B (Fig. 6.1b). Note that a model of migration in the opposite direction (from B to A) is slightly simpler, since any migration event prior to τ_1 brings all three lineages into the same deme. In the context of directional population founding (chapter 5), secondary gene flow from the ancestral to the derived populations (A into B or C) is more relevant than migration in the reverse direction. However, the basic results apply to both cases.

We need a notation that keeps track of both the origin and the locations of lineages. Denoting lineages by their sampling location (a, b, c) and keeping the order of populations fixed, the three possible states between the present and τ_1 are: $((a), (b), (c))$, $((ab), (), (c))$ and $coal_{ab}$, the latter corresponding to coalescence of the a and b lineage. Going backwards in time, the lineage in population C cannot migrate or coalesce during the first time interval, but there is a chance that the lineage sampled in B jumps to A during τ_1 (which corresponds to a migration event in the opposite direction forwards in time) and, if it does, that it coalesces with the resident lineage resulting in a gene tree topology $(c(ab))$. The starting configuration at the time of sampling is $P_{start} = (1, 0, 0)$. We only need to follow the process until the first coalescence event, i.e. the state, $coal_{ab}$ is absorbing.

The transition probabilities from the present to τ_1 are:

$$\mathbf{M}_1 = \begin{bmatrix} 1-m & 0 & 0 \\ m & 1-\lambda & 0 \\ 0 & \lambda & 1 \end{bmatrix}$$

The resulting state probabilities at time τ_1 are:

$$P_{\tau_1} = \mathbf{M}_1^{\tau_1}.P_{start} \tag{6.2}$$

Looking into the past, populations B and C merge instantaneously at time τ_1. During the following time interval τ_0, the ancestral process can again be described as a Markov chain which now has 7 states: $S1 = ((a),(bc))$, $S2 = ((ab),(c))$, $S3 = ((ac),(b))$, $S4 = coal_{ab}$, $S5 = coal_{ac}$, $S6 = coal_{bc}$ and $S7 = all$, the latter corresponding to the case where all 3 lineages find themselves in the same deme. As in the previous interval, coalescence events $(S4, S5, S6)$ are absorbing states. Because lineages are exchangeable and all topologies have the same probability of 1/3 once $(S7)$ is reached, this state is also absorbing. The matrix of transition probabilities is:

$$\mathbf{M}_0 = \begin{bmatrix} 1-2m-\lambda & 0 & 0 & 0 & 0 & 0 & 0 \\ m & 1-m-\lambda & 0 & 0 & 0 & 0 & 0 \\ m & 0 & 1-m-\lambda & 0 & 0 & 0 & 0 \\ 0 & \lambda & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 1 & 0 & 0 \\ \lambda & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & m & m & 0 & 0 & 0 & 1 \end{bmatrix}$$

As in the previous interval, the state probabilities at the end time $\tau_1 + \tau_0$ are given by taking the τ_0 th power of $\mathbf{M}_0$ and multiplying on the right with the state probabilities after the previous interval, P_{τ_1}.

$$P_{\tau_0} = \mathbf{M}_0^{\tau_0}.P_{\tau_1} \tag{6.3}$$

We can substitute 6.2 into the above and solve to get the total probabilities of the various states at time $\tau_1+\tau_0$, $P_{S1}, P_{S2}...P_{S7}$. It is straightforward to get from this to the topological probabilities. Consider first all states at which topologies are equiprobable. Because at $\tau_1 + \tau_0$ all remaining lineages automatically find themselves in the common ancestral deme, the total probability of reaching exchangeability, P_{eq} is:

$$P_{eq} = P_{S1} + P_{S2} + P_{S3} + P_{S7} \tag{6.4}$$

Evaluating the above using the matrix power function in *Mathematica* and simplifying gives:

$$\begin{aligned} P_{eq} \quad = \quad & \frac{1}{(m-\lambda)(m+\lambda)(2m+\lambda)}(2m^3(1-\lambda)^{\tau_1} \\ & -3(1-m)^{\tau_1}m^2\lambda + m^2(1-\lambda)^{\tau_1}\lambda - (1-m)^{\tau_1}m^2(1-2m-\lambda)^{\tau_0}\lambda \\ & +(1-m)^{\tau_1}(1-2m-\lambda)^{\tau_0}\lambda^3 + ((1-m)^{\tau_1}(m-2\lambda) + m(1-\lambda)^{\tau_1})(1-m-\lambda)^{\tau_0}\lambda(2m+\lambda)) \end{aligned}$$

We can scale model parameters by $2N_e$ so that coalescence happens at rate 1, i.e. $M = 2N_e m$, $T_i = \tau_i/2N_e$. Transforming to continuous time simplifies things slightly.

$$\begin{aligned} P_{eq} \quad = \quad & \frac{1}{(M-1)(M+1)(2M+1)}(2M^3e^{-T_1} - 3e^{-MT_1}M^2 + M^2e^{-T_1} - e^{-MT_1}M^2e^{-(2M+1)T_0} + \\ & e^{-MT_1}e^{-(2M+1)T_0} + (e^{-MT_1}(M-2) + Me^{-T_1})e^{-(M+1)T_0}(2M+1)) \end{aligned}$$

The probabilities of the three topologies are given as the sum of $1/3P_{eq}$ and the state corresponding to the respective coalescent event:

$$\begin{aligned} P_{(c(ab))} \quad = \quad & P_{S4} + 1/3P_{equ} = \\ & = \frac{1}{3(M-1)(M+1)(2M+1)}(6M^3 - 4M^3e^{-T_1} + 3M^2 - 2M^2e^{-MT_1}e^{-(1+2M)T_0} - 6M \\ & +3Me^{-MT_1} - 3 + 3e^{-MT_1} - 2e^{-MT_1}e^{-(1+2M)T_0} + e^{-(1+M)T_0}(2M+1)(-2Me^{-MT_1} \\ & +e^{-MT_1}(M+1))) \end{aligned}$$

$$\begin{aligned} P_{(b(ac))} \quad = \quad & P_{S5} + 1/3P_{equ} = \\ & = \frac{1}{3(M-1)(M+1)(2M+1)}(2M^3e^{-T_1} + M^2e^{-T_1} + 2M^2e^{-MT_1}e^{-(1+2M)T_0} \\ & -2e^{-MT_1}e^{-(1+2M)T_0} - 3Me^{-MT_1} + e^{-(M+1)T_0}(2M+1)(Me^{-T_1} + e^{-MT_1}(-2M+1))) \end{aligned}$$

$$\begin{aligned} P_{(a(bc))} &= P_{S6} + 1/3P_{equ} = \\ & \frac{1}{3(M-1)(M+1)(2M+1)}(2M^3e^{-T_1} + M^2e^{-T_1} - 4M^2e^{-T_1}e^{-(1+2M)T_0} - 3e^{-MT_1} \\ & +4e^{-T_1}e^{-(1+2M)T_0} + (e^{-MT_1}(M-2) + Me^{-T_1})e^{-(1+M)T_0(2M+1))} \end{aligned}$$

In the limit of $M \rightarrow 0$, these reduce to the results for the standard divergence model without gene flow (eq. 6.1). As in the case of structure in the ancestral population (Slatkin & Pollack, 2008), these analytical solutions are somewhat cumbersome and probably of limited use analytically. For instance, it is not possible to get an easy solution for M from the above expressions which could be used to estimate this parameter from observed topological frequencies. However, plotting the probabilities of the three topologies against the model parameters immediately gives a feeling for the properties of the model. As shown in figure 6.2a, $P_{(c(ab))}$ increases rapidly with M at the expense of $P_{(a(bc))}$ and $(c(ab))$ becomes the most likely topology for $M > 0.5$. T_1 has a similar effect on topological probabilities (Fig. 6.2b). $P_{(c(ab))}$ increases with larger values of T_1. However, in this case, both $P_{(a(bc))}$ and $P_{(b(ac))}$ go to 0. In contrast, the dependence of topological probabilities on T_0 is rather weak (Fig. 6.2c). It may therefore be useful to investigate topological probabilities in the limits of T_0.

$$\begin{aligned} T_0 \rightarrow \infty, P_{(c(ab))} &\rightarrow \frac{e^{-(1+M)T_1}3e^{T_1}(1+M) + e^{MT_1}(1+2M)(-2M^2 + 3e^{T_1}(M^2-1))}{3(M-1)(M+1)(2M+1)} \\ P_{(b(ac))} &\rightarrow \frac{e^{-(1+M)T_1}M(-3e^{T_1} + e^{MT_1}M(1+2M))}{3(M-1)(M+1)(2M+1)} \\ P_{(a(bc))} &\rightarrow \frac{e^{-(1+M)T_1}(-3e^{T_1} + e^{MT_1}M^2(1+2M))}{3(M-1)(M+1)(2M+1)} \end{aligned}$$

and,

$$\begin{aligned} T_0 \rightarrow 0, P_{(c(ab))} &\rightarrow \frac{3 - e^{-2MT_1} - e^{T_1}2M - 3M}{3-3M} \\ P_{(b(ac))} &\rightarrow \frac{e^{-MT_1} - e^{T_1}M}{3-3M} \\ P_{(a(bc))} &\rightarrow \frac{e^{-MT_1} - e^{T_1}M}{3-3M} \end{aligned}$$

Perhaps unsurprisingly, the upper limit agrees well with the exact solution in the parameter range where $(c(ab))$ is the most likely topology (Fig. 6.2a and b).

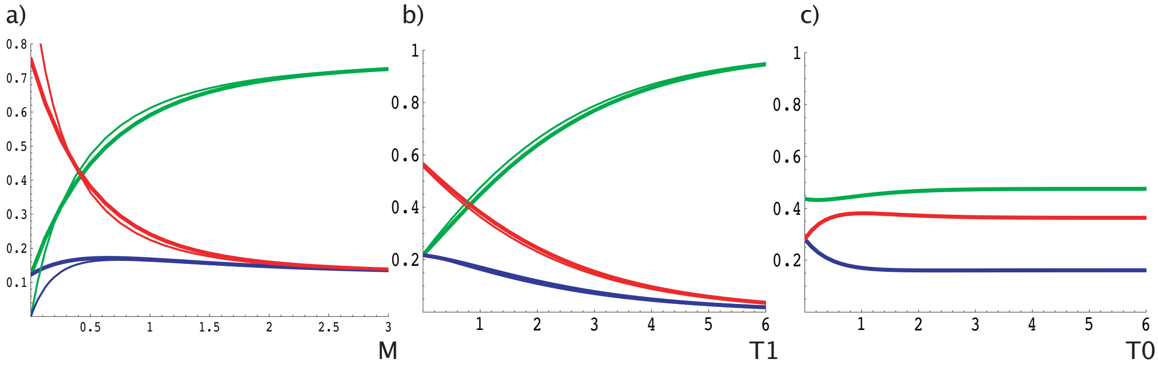

Figure 6.2: Analytical solutions (thick lines) for the probabilities of the three genealogical topologies (red $= (a(bc))$, green= $(c(ab))$, blue=$(b(ac))$) for the case of asymmetric migration from A to B (Fig. 6.2b) plotted against the three scaled model parameters: M (a), T_1 (b) and T_0 (c). In each case the other two model parameters respectively are held constant at $T_1 = 1$, $T_0 = 1$ and $M = 0.5$. For a and b the limits of $T_0 \to \infty$ are shown as thin lines.

6.1.2 Limit cases

Since we are primarily interested in the emergence of asymmetry rather than the topological probabilities as such, it may be illuminating to consider the difference between the probability of the two incongruent topologies, i.e. $D = P_{(c(ab))} - P_{(b(ac))}$.

Substituting and simplifying yields:

$$D = P_{S4} - P_{S5} = \frac{-M^2(e^{-T_1} - 1) + Me^{-(1+M)T_0}(e^{-MT_1} - e^{-T_1}) + (e^{-MT_1} - 1)}{M^2 - 1} \tag{6.5}$$

In the limit of $T_0 \to \infty$,

$$D \to \frac{1 - e^{-MT_1} - M^2 - e^{-T_1}M^2}{1 - M^2} \tag{6.6}$$

in the alternative limit of $T_0 \to 0$,

$$D \to \frac{1 - e^{-MT_1} - M - e^{-T_1}M}{1 - M} \tag{6.7}$$

As can be see from the equations above the two limits differ only in the way M enters (quadratically for $T_0 \to \infty$ and linearly for $T_0 \to 0$). In general, the upper limit agrees with the exact solution surprisingly well (Fig. 6.3).

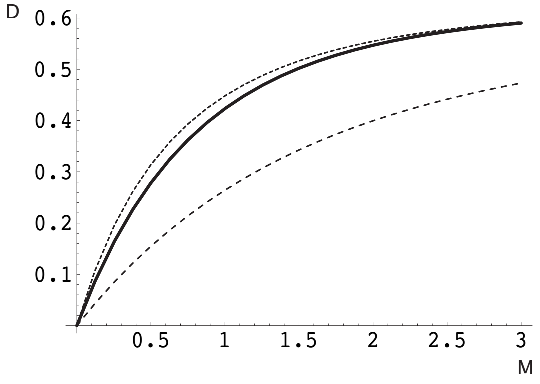

Figure 6.3: The difference in probability between the two incongruent topologies, $P_{(c(ab))} - P_{(b(ac))}$, plotted against the scaled migration rate, M for $T_1 = T_0 = 1$ (thick line). Also shown are solutions for the two limits $T_0 \to 0$ (wide dash below) and $T_0 \to \infty$ (narrow dash above).

6.1.3 Symmetric migration

Analogous derivations can be made for the case of symmetric migration between population A and B (Fig. 6.1). It is straightforeward to set up the matrices of transition probabilities for τ_1 and τ_0. During τ_1 the ancestral process is described by a 5x5 matrix. The states are $((a),(b),(c))$, $((b),(a),(c))$, $((ab),(),(c))$, $((),(ab),(c))$ and $coal_{ab}$ with starting state $P_{start} = (1,0,0,0,0)$

$$\mathbf{M}_1 = \begin{bmatrix} 1-2m & 0 & m & m & 0 \\ 0 & 1-2m & m & m & 0 \\ m & m & 1-2m-\lambda & 0 & 0 \\ m & m & 0 & 1-2m-\lambda & 0 \\ 0 & 0 & \lambda & \lambda & 1 \end{bmatrix} \tag{6.8}$$

Because migration is symmetric, we do not need to keep track of the locations of lineages during τ_0. All that matters is which pair of lineages finds itself in the same deme and coalesces first. Thus there are 7 possible states during τ_0, which can be denoted by a single bracket; $S1 = (bc)$, $S2 = (ab)$, $S3 = (ac)$, $S4 = coal_{ab}$, $S5 = coal_{ac}$, $S6 = coal_{bc}$ and $S7 = all$.

$$\mathbf{M}_0 = \begin{bmatrix} 1-3m-\lambda & m & m & 0 & 0 & 0 & 0 \\ m & 1-3m-\lambda & m & 0 & 0 & 0 & 0 \\ m & m & 1-3m-\lambda & 0 & 0 & 0 & 0 \\ 0 & \lambda & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 1 & 0 & 0 \\ \lambda & 0 & 0 & 0 & 0 & 1 & 0 \\ m & m & m & 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.9}$$

Unlike in the asymmetric case, $\mathbf{M}_0$ and $\mathbf{M}_1$ are not upper triangular matrices, so the resulting topological probabilities are more cumbersome. Substituting into equations 6.2, 6.3 and 6.4 and solving gives:

$$\begin{aligned} P_{(c(ab))} &= \frac{1}{3(4m+\lambda)a}(2^{-(2+\tau_1)}(64m^2(b^{\tau_1}-c^{\tau_1})+4m((d^{\tau_1}-9)\lambda(c^{\tau_1}-b^{\tau_1})-4a(b^{\tau_1}+c^{\tau_1}-32^{\tau_1})) \\ &\quad +\lambda(32^{(1+\tau_1)}(2-(1-2m)^{\tau_1}+(1-2m)^{\tau_1}d^{\tau_0})a-(3+d^{\tau_0})(\lambda(c^{\tau_1}-b^{\tau_1})+a(b^{\tau_1}+c^{\tau_1}))))) \end{aligned}$$

$$\begin{aligned} P_{(b(ac))} &= \frac{1}{3(4m+\lambda)a}(2^{-(2+\tau_1)}(4mab^{\tau_1}+d^{\tau_0}\lambda ab^{\tau_1}-(d^{\tau_0}\lambda(\lambda-4m)+(4m+3\lambda))b^{\tau_1} \\ &\quad +4mac^{\tau_1}+d^{\tau_0}\lambda ac^{\tau_1}+(d^{\tau_0}\lambda(\lambda-4m)+(4m+3\lambda))c^{\tau_1})) \end{aligned}$$

$$\begin{aligned} P_{(a(bc))} &= \frac{1}{3(4m+\lambda)a}(2^{-(1+\tau_1)}(-32m^2b^{\tau_1}+(d^{\tau_0}-3)\lambda^2(b^{\tau_1}-c^{\tau_1})+8m(ab^{\tau_1}+(4m+a)c^{\tau_1}) \\ &\quad +\lambda(-32^{(1+\tau_1)}(1-2m)^{\tau_1}(d^{\tau_0}-1)a+4m(3+d^{\tau_0})(c^{\tau_1}-b^{\tau_1})-(d^{\tau_0}-3)a(b^{\tau_1}+c^{\tau_1}))) \end{aligned}$$

where $a = \sqrt{16m^2+\lambda^2}, b = 2-4m-\lambda-a, c = 2-4m-\lambda+4$ and $d = (1-4m-\lambda)$.

A comparison between figures 6.4 and 6.2, shows that migration has a very similar qualitative effect to that in the simpler asymmetric migration model. $P_{(c(ab))}$ increases with M at the expense of the probability of the congruent topology $P_{(a(bc))}$ and, to a lesser extent, $P_{(b(ac))}$ (Fig. 6.4a). As before asymmetries can only arise during τ_1 and the dependency on τ_0 is weak (Fig. 6.4c), even more so than in the asymmetric migration scenario (Fig. 6.2c). Interestingly, for $M \to \infty$ the difference in the probability of the incongruent topologies D is approximately halved in the symmetric case. Although this may seem counterintuitive, since one would think that the increased possibility for migration also increases the

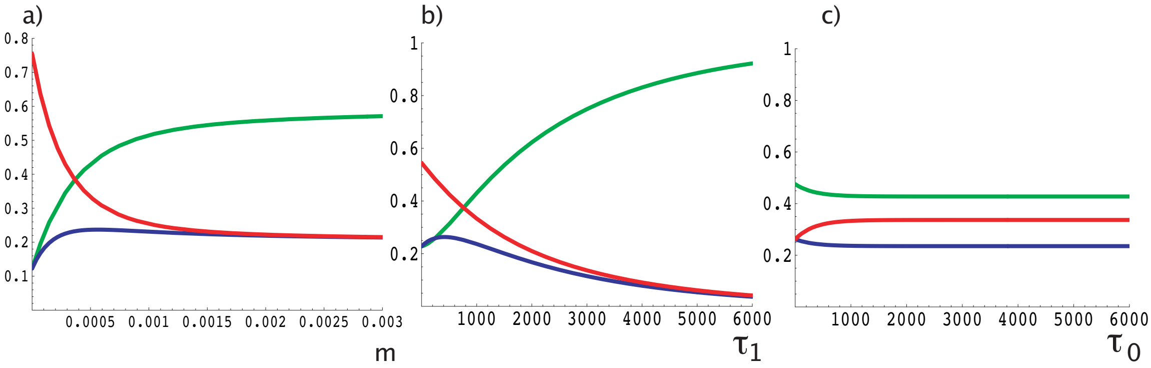

Figure 6.4: Analytical solutions (thick lines) for the probabilities of the three genealogical topologies (red $= (a(bc))$, green$= (c(ab))$, blue$=(b(ac))$) for the case of symmetric migration between A and B (Fig. 6.1c) plotted against the three model parameters: m (a), τ_1 (b) and τ_0 (c). In each case the other two model parameters respectively are held constant at $\tau_1 = 1000$, $\tau_1 = 1000$ and $m = 0.0005$. These parameter ranges correspond to those shown in figure 6.2 on the coalescence time scale.

chance that lineages from A and B coalesce during τ_1, it can be easily understood considering the events possible during τ_1. In the asymmetric case, coalescence during τ_1 must be preceded by the only possible migration event (backwards in time the lineage sampled in B jumps into A), so the chance of topology $(c(ab)))$ increases with the rate of migration until, with very high migration rates, this jump from B to A occurs almost instantaneously. In contrast, with symmetric migration, lineages will jump back and forth between the two demes and in the limit of high migration, the effective population size during τ_1 is effectively doubled.

6.2 Discussion

The main motivation for this analysis was to understand the influence of gene flow on topological probabilities. Compared to the special types of population structure required to produce asymmetries (Slatkin & Pollack, 2008), gene flow after initial divergence seems a biologically more relevant mechanism at least for populations of the same or closely related species. While, as in the case of population structure (Slatkin & Pollack, 2008), the analytical results are cumbersome, important insights can be gained simply by plotting topological probabilities against the three model parameters (Figs. 6.2, 6.4) and investigating the relevant limits. Firstly, asymmetries in topological probabilities can arise as a result of a small amount of gene flow, but only if the older species/population is involved and only if migration occurs after the more recent divergence. This makes intuitive sense given that migration before the more recent divergence event affects both b and c lineages equally. Gene flow therefore is an unlikely explanation for the topological asymmetries that have been reported from species triplets such as the human-chimp-gorilla case (Patterson *et al.*, 2006) or the *Drosophila* triplet considered by Pollard *et al.* (2006). Secondly, a simple expression can be derived for the difference between the probability of the two incongruent topologies in the case of asymmetric migration (eq. 6.5). This result could, at least in principle, be used to estimate migration rates from topological frequencies of intraspecific triplets.

However, the usefulness of such theory obviously depends on how common asymmetries actually are in intraspecific triplets. To explore this, I tested for asymmetries in genomic data from three *D. melanogaster* populations (Africa, Europe and North America) (Obbard *et al.*, 2009). *D. melanogaster* is a commensal species which broadly shares our own Out of Africa history. However, the colonisation of North America is thought to have occurred only a few hundred years ago (Stephan & Li, 2006). Two plausible migration scenarios would result in asymmetries in site counts in opposite directions. Secondary gene flow between Africa and Europe would lead to an excess of polymorphic sites shared between Africa and Europe, and likewise, migration between Africa and North America would increase the frequency of polymorphic sites shared by those populations. Both scenarios are possible given Human trading routes. The raw data (kindly provided by D. Obbard) consisted of polymorphism information (both coding and non-coding sequence) for 287 control genes. Extractions from 8 individuals per population were pooled and SOLEXA sequenced to 100x coverage each (for detailed methods see Obbard *et al.* (2009)). To produce triplet site counts, a single nucleotide per population and polymorphic site was sampled at random using the observed frequency of the segregating sites in the sequence pool.

Table 6.1 shows a slight, excess of mutations shared by African and European populations compared to sites shared between Africa and N. America (Tab. 6.1). Given the large distance between genes and the fact that linkage disequilibrium does not extend over a few hundred bases in *Drosophila*, one may for simplicity assume that each polymorphic site has a unique genealogy. However, even when ignoring linkage of nearby polymorphic sites, the observed difference is not significant ($\chi^2 = 0.731, p = 0.392$). In other words, although the data are consistent with low levels of migration between Africa and Europe following the divergence of the N. American population, they also fit a null model of divergence without subsequent migration.

The above example illustrates the limited power of detecting migration from topological asymmetries alone even with a relatively large number of loci sampled across the genome. In general, genome wide site counts in triplet alignments, are a rather inefficient way to extract information about gene flow for two reasons. Firstly, the presented analysis only deals with topologies which are of course not observed directly but rather inferred from the sequence information. Even if each site in the genome had its own

Table 6.1: Counts of mutations shared between population pairs of *D. melanogaster*.

Population pair	site counts
Europe/N. America	3215
Africa/Europe	1889
Africa/N. America	1836

genealogy, the chance of a mutation occurring would still depend both on the topology and the branch lengths of that genealogy. Therefore the expected coalescence times for pairs of sequences are more immediately useful for the analysis of genome wide site counts. Although these can been derived for the present model using existing results for the isolation with migration model (Wakeley, 1996; Wilkinson-Herbots, 2008), the results (not shown) do not lead to a simple expression for M as in the case of migration across a barrier within the ancestral populations (Slatkin & Pollack, 2008, eq. 9). It is easy to see that much of the signal about the relative magnitude of incomplete lineage sorting vs. migration is contained in the joint distribution of branch lengths and topologies. For example, as shown in figure 6.1b, in the case of asymmetric gene flow, genealogies affected by migration have on average a longer internal branch than those involving incomplete lineage sorting with a most recent common ancestor $> \tau_0 + \tau_1$. This increases the number of derived mutations shared by a and b. Secondly, in most cases we do not have independent information about the topology of the population tree which is usually inferred from the sequence data as well. This means that if migration rates are high ($M > 0.5$ in figure 6.2) so that the 'incongruent topology' becomes the most likely history, the population topology would automatically be mis-inferred, no matter whether one uses site counts, likelihood methods which assume no migration (Yang, 2002) or more realistic approximate methods to fit models of migration and divergence (Hey, 2010b).

However, despite these difficulties, the usefulness of full analytical results for divergence with migration models is illustrated by a recent genomic study on the history of our own species. Green *et al.* (2010) use a measure of asymmetry very similar to the D considered above to compare the recently sequenced Neanderthal genome to human genomes sampled from different populations. They find a significant excess of 4% of derived sites shared by Neanderthals and Eurasian *Homo sapiens* compared to sites shared by Neanderthals and African *H. sapiens* which they interpret as evidence for hybridisation between Neanderthals and ancient *H. sapiens* outside Africa (Green *et al.*, 2010, SOM 15). However, the authors do admit that this signal could equally be explained by ancestral population structure prior to the expansion of modern Humans out of Africa. Without full analytical results for both models it remains impossible to evaluate to what extent these two models can be distinguished and estimate the rate of hybridisation required to explain the data. Presumably, the comparatively brief period of Human-

Neanderthal coexistence implies that any hybridisation scenario would have to invoke rather high levels of gene flow compared to the very weak barriers in ancestral populations that can lead to asymmetries (Pollard *et al.*, 2006) provided such structure in our African ancestor persisted over a long timescale.

Chapter 7

Discussion

Each of the preceding chapters contains its own, extensive discussion. Below, I first give a brief summary of the main findings of the individual chapters and then discuss two general issues that emerge from this work: The effects of sampling and recombination on historical inference. Given that coalescent theory has always been driven by the availability of genetic data, it seems appropriate to view these in the light of the current revolution in sequencing technology.

7.1 Conclusions

7.1.1 Chapter 2

Chapter 2 shows that the degree of starshape of a genealogy is readily detectable using summary statistics and can be taken as a surrogate for the effect of past demography and other non-neutral forces. Although summary statistics such as Tajima's D (Tajima, 1989) and related measures are commonly used for this they are far from ideal (Felsenstein, 1992). Two types of simple new statistics are derived, which are based on the number of mutations on the rootward branches as inferred from polarized alignments by a straightforward algorithm or the properties of a perfectly starshaped genealogy respectively. Power analyses on data simulated under a history of exponential growth show that these measures are equal or superior to standard neutrality tests. In particular, this comparison reveals that genealogical ratios outperform standard summary statistics in tests based on the mean and variance across multiple unlinked loci. By grouping genealogies according to their (random) topology, it becomes clear that statistics which depend on pairwise measures such as Tajima's D are most severely confounded with the topology which explains their comparatively low power and dependence on large sample sizes. In contrast, genealogical ratios efficiently extract information from small numbers of individuals. Provided reliable outgroup information is available these statistics may constitute a useful alternative to full likelihood estimation and standard tests of neutrality and could form the basis for approximate methods of demographic inference.

7.1.2 Chapter 3

Chapter 3 investigates the phylogeographic history of a radiation of high alpine ground beetles (genus *Trechus*) on a single mountain range, the Orobian Alps in Northern Italy using sequence data from two loci. Bayesian stochastic search variable selection (BSSVS) (Lemey *et al.*, 2009; Ceiridwen *et al.*, 2010) is used to infer the most parsimonious set of directional location state changes together with standard mutational parameters and genealogies. While this inference is entirely based on the genealogy and as such blind to the underlying population level processes, a minimal set of location state changes which connects all populations has a straightforward and testable interpretation under a model of successive founder events originating from a refugium. Given the minimal set of location state changes which determines the order of population founding, the paraphyly constraints implicit in this model can be tested. Only three of the 12 sampled *Trechus* populations are incompatible with this scenario. This is remarkable given that the BSSVS approach is highly sensitive to location state changes which occur multiple times in the genealogy as expected from incomplete lineage sorting or migration but not under the founder event model. It also contrasts with previous phylogeographic studies on alpine insects, which have found extensive incomplete lineage sorting (Knowles, 2001; Carstens & Knowles, 2007a). Furthermore both mitochondrial and nuclear genealogies support separate refugial origins for populations on the western and eastern ends of the Orobian Alps, and mitochondrial node ages suggest persistence on the northern ridge for at least part of the last ice age. The deep phylogeographic structure within Orobian *Trechus* is in stark contrast to previous larger-scale phylogeographic studies particularly on high alpine plants (Schönswetter *et al.*, 2005) and suggests that dispersal-limited, high alpine arthropods may have quite different histories than the more dispersive alpine taxa previously studied. While BSSVS offers a quantitative way to extract directional information, the analysis also demonstrates the limited power of phylogeographic sampling schemes of small numbers of loci sampled for many individuals. In particular, it is not possible to distinguish between incomplete lineage sorting and migration.

7.1.3 Chapter 4

Chapter 4 describes how exon-primed, intron-crossing (EPIC) loci can be developed relatively straightforwardly for highly conserved genes using publicly available genomic data and expressed sequence tags

(ESTs). Amplification success of degenerate primers developed for 40 loci was scored on a diverse panel of Hymenoptera associated with oak galls and figs. Although amplification success declines with taxonomic distance from the species used for primer design (*Nasonia*), considerable numbers of loci amplify even in the gall and fig wasp hosts which are very distantly related to *Nasonia*. Estimates of divergence and diversity within Europe obtained for two Pteromalid parasitoids *C. fungosa* and *M. amaenus* suggest that these loci contain information about their phylogeographic history. Focusing on highly conserved genes for which degenerate primers can be built circumvents the need for species specific primer design or PCR optimisation (Papanicolaou *et al.*, 2005) required by alternative markers, in particular anonymous loci (Jennings & Edwards, 2005). Furthermore, these loci should make it possible in the future to investigate the history of entire natural communities in a quantitative framework.

7.1.4 Chapter 5

In chapter 5 sequence data from 20 of the newly developed nuclear loci are used to infer the historical relationships of three refugial populations (Middle East, the Balkans and Iberia) of the oak gall parasitoid *C. fungosa*. Previous studies on gall wasps (Rokas *et al.*, 2003; Challis *et al.*, 2007), their oak hosts (Dumolin-Lapegue *et al.*, 1997) and their chalcid parasitoids (Hayward & Stone, 2005; Nicholls *et al.*, 2010) as well as other temperate taxa (Michaux *et al.*, 2004; Culling *et al.*, 2006; Koch *et al.*, 2006) have found patterns of genetic diversity consistent with an eastern origin of refugial populations in southern Europe. This westwards expansion has been estimated to have begun in the early Pleistocene or before. Comparing the support for all possible population tree topologies using likelihood and Bayesian methods also suggests an ‘Out of the East’ history for *C. fungosa*. However, the estimated divergence times between refugial populations are surprisingly recent, coinciding with the last (Eemian) interglacial. The difference between population divergence times derived from model-based analyses and naïve interpretations of mitochondrial node ages can be entirely explained by the large ancestral population sizes inferred for *C. fungosa*. Given that most previous phylogeographic studies investigating the longitudinal history of temperate taxa have ignored this ancestral variation, the refugial populations of temperate taxa in Europe are likely to be younger than previously assumed in general.

The comparison of the two sampling schemes shows that there is significant information about population divergence in minimal samples. This is encouraging in two ways. In theory, full likelihood methods (Yang, 2002, 2010; Wang & Hey, 2010) are only tractable for minimal samples. In practice community wide studies are limited by their ability to include rare species, so the fact that single specimens are sufficient, provided a large number of loci is sampled, means that these methods can be used to test alternative models of parasitoid assemblage evolution in the future.

A problem with the assumed model of divergence is that it ignores migration between populations, an obvious possibility for refugial populations. However, the recency of the divergence time estimated for *C. fungosa* and the limited power to estimate ancestral N_e, suggest that much larger numbers of loci would be needed to fit more parameter-rich isolation with migration models (Nielsen & Wakeley, 2001; Hey & Nielsen, 2004). An interesting alternative history involving migration is a model of repeated episodes of gene flow occurring during interglacials (Jesus *et al.*, 2006). Although this model is of immediate interest given the Pleistocene climate cycles, it may be difficult to distinguish from a simpler history of recent divergence without migration. This is because a single episode of strong migration rapidly erases any signature of previous historical events. However, the two scenarios can potentially be distinguished on a community-wide scale by comparing species with different dispersal abilities. If geneflow during interglacials is important, one would expect to see more recent coalescence times in species with good dispersal abilities compared to poor dispersers.

7.1.5 Chapter 6

In chapter 6 the three population model used in the previous chapter is extended analytically by allowing for migration between the older population and one of the more recently diverged populations. The probabilities of genealogical topologies are derived for minimal triplet samples using a discrete time Markov-Chain. Plotting topological frequencies against model parameters gives a clear understanding of the effects of migration. As would be expected intuitively, migration disproportionally increases the probability and expected frequency of one of the two incongruent topologies ($(c(ab))$ in Fig. 6.1). This asymmetry in topological probabilities arises solely from migration after the more recent divergence event. The analysis illustrates the difficulty of obtaining even simple results for realistic, non-equilibrium models. The analytical difficulty arises directly from the lack of symmetry in the migration model which makes it necessary to consider all possible combinations of migration and coalescence events. Thus increasing the realism of these models (for example by relaxing the simplifying assumption of equal population size for all populations) introduces additional asymmetries which will further complicate analysis. However, since the main motivation for such theoretical work is the development of computational methods for the analysis of sequence data, complexity may not matter. For example, if expressions for the probability of full data patterns could be generated automatically (either by using a Matrix approach or by finding recursions for the moment generating function), they would be immediately useful even if they are complex.

7.2 Outlook

A common theme throughout this thesis has been the effect of sampling on statistical power. In chapter 2, the gain in power to infer past demography was shown to diminish rapidly with sample size. Similarly, the comparison of the two sampling schemes in chapter 4 and basic coalescent theory (Takahata *et al.*, 1995) suggests that the most efficient sampling scheme is one of a single individual sampled at a great number of loci. Although the importance of replicating across loci has been pointed out by many (e.g. Felsenstein, 1992, 2006; Wakeley, 2004b; Wang & Hey, 2010), there are surprisingly few thorough investigations of the effect of sampling. Felsenstein (2006) showed that the accuracy in estimating the scaled mutation rate in the neutral Wright-Fisher model only increases logarithmically with sample size, but is proportional to the number of loci. However, he points out that the optimal sampling schemes may differ between models and recommends that for histories involving migration "one would want to have larger sample sizes in each population to detect recent migration" (Felsenstein, 2006). In contrast, Wang & Hey (2010) conclude that Felsenstein's reasoning for a single population essentially extends to isolation with migration (IM) models. Knowing the optimal sampling scheme for parameter estimation under a particular model matters in two ways. Firstly, sequencing studies now face a genuine choice between sequencing a few moderately-sized genomes using next generation sequencing or obtaining sequences for a large number of individuals at a handful of selected loci using Sanger technology. If historical signal can be most efficiently extracted from a very large number of loci sequenced for a few individuals, even the most fragmented genome assemblies for two or three individuals contain vastly more information than traditional phylogeographic samples. Secondly, many of the theoretical complications that limit current inference methods disappear for small samples. In particular, likelihood methods which integrate over all possible histories and thus break down for moderate sample sizes are tractable for pairs and triplets (Wang & Hey, 2010; Yang, 2010). With the increasing availability of genomic data, such exact methods of historical inference will undoubtedly become more important in the future, not least because they are computationally more efficient than schemes based on simulations.

However, the analysis of genomic data comes with new challenges. In particular, recombination presents a conundrum for historical inference. On the one hand, it generates crucial replication by uncoupling the genealogical histories of nearby genomic regions. On the other hand, in practice linkage patterns can only be inferred incompletely and indirectly from polymorphism information, making it difficult to define blocks of shared ancestry. Almost all methods of historical inference assume that the history of a given locus can be described by a single bifurcating genealogy (Nielsen & Wakeley, 2001; Yang, 2002; Rannala & Yang, 2003; Hey & Nielsen, 2004). In other words, it is assumed that there is no recombination within, but free recombination between loci. Given that the rate of recombination is of the same order as the mutation rate in some organisms, this is obviously a gross oversimplification which is clearly violated when dealing with large continuous blocks of sequence. The ancestry of a sample of recombining sequences can be described as a graph (Griffiths, 1991; Wakeley, 2008). Unfortunately, results for divergence and/or migration models based on the ancestral recombination graph are not available. Thus in practice, the complications of recombination are avoided by trimming data into supposedly non-recombining segments (chapters 3,5) based on the four-gamete test (Hudson & Kaplan, 1985). However, what effects this has on inference is poorly understood. A recent simulation study (Strasburg & Riesenber, 2009) found no bias in parameter estimates under the IM model even for substantial amounts of recombination as long as loci were trimmed. However, this obviously throws information away in two ways. Firstly, shortening sequences reduces the mutational information available to infer genealogies. Secondly, the pattern of recombination itself contains information about the underlying history. In particular, the rate of recombination along a genealogy is proportional to its length and so the scale of correlation along the genome gives a clock that is independent of the mutation rate. Surprisingly perhaps, no current IM model uses linkage information. However, a powerful Hidden-Markov framework, which approximates the coalescent with recombination and uses information from linked sites to infer changes in topology along the genealogy has been developed for divergence models (Hobolth *et al.*, 2007). Similarly, Hellenthal *et al.* (2008) have developed a scheme to use recombination to fit a model of population founding in humans. In the Neanderthal case (Green *et al.*, 2010), recent hybridisation should be distinguishable from ancient population structure from the length of sequence blocks shared by Neanderthals and Humans.

To conclude, coalescent theory has become indispensable for the analysis of sequence data. It provides a sound quantitative description of the histories of samples and the population genetics processes shaping them which has made the historical divide between tree-based phylogeography and frequency based population genetics obsolete. Thinking in term of genealogies does indeed provide a deeper understanding of the historical signal in genetic data and ideally leads to new and more powerful ways to extract this information. However, we have only just begun to realise the full potential of the coalescent for historical inference. Massively parallel sequencing technologies are rapidly closing the practical gap between the study of model and non-model organisms. While these genomic datasets promise ever greater power for historical inferences, the limiting factor is now the availability of appropriate theory and efficient computational methods. This means that improving existing and developing new analytical methods based on the coalescent will remain a central task of population genetics for decades to come.

This work will undoubtedly have to tackle many difficult challenges, some new but many old. However, the potential rewards are immense. Improving our ability to see into the past opens up exciting possibilities for the study of community assembly and will shed new light on our own evolutionary journey.

Appendix

Bibliography

Aebi, A., Schönrogge, K., Melika, G., Alma, A., Bosio, G., Quacchia, A., Picciau, L., Abe, Y., Moriya, S., Yara, K., Seljak, G. & Stone, G.N. (2006). Parasitoid recruitment to the globally invasive chestnut gall wasp *Dryocosmus kuriphilus*. In *Galling Arthropods and Their Associates*, pages 103–121. Springer, Japan.

Allan, G.J., Francisco-Ortega, J., Santos-Guerra, A., Boerner, E. & Zimmer, E.A. (2004). Molecular phylogenetic evidence for the geographic origin and classification of Canary Island *Lotus* (Fabaceae: Loteae). *Molecular Phylogenetics and Evolution*, 32(1), 123–138.

Andolfatto, P. (2001). Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. *Molecular Biology and Evolution*, 18(3), 279–290.

Andolfatto, P. & Przeworski, M. (2000). A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. *Genetics*, 156(1), 257–268.

Askew, R.R. (1961). Some biological notes on the pteromalid (Hym. Chalcidoidea) genera *Caenacis* Förster, *Cecidostiba* Thomson and *Hobbya* Delucchi, with descriptions of two new species. *Entomophaga*, 6, 58–67.

Askew, R.R. (1980). The diversity of insect communities in leaf mines and plant galls. *The Journal of Animal Ecology*, 49, 145–152.

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

Bailey, R., Schönrogge, K., Cook, J.M., Melika, G., Csóka, G., Thúroczy, C. & Stone, G.N. (2009). Host niches and defensive extended phenotypes structure parasitoid wasp communities. *PLoS Biology*, 7(8), e1000179.

Barr, T.C.J. (1985). Pattern and process in speciation of trechine beetles in eastern North America (Coleoptera: Carabidae: Trechinae). In G.E. Ball, editor, *Taxonomy, Phytogeny and Zoogeography of Beetles and Ants*, pages 350–407. Dr W. Junk Publishers, Dordrecht, The Netherlands.

Barton, N.H., Depaulis, F. & Etheridge, A. (2002). Neutral evoluton in spatially continuous populations. *Theoretical Population Biology*, 61, 31–48.

Barton, N.H., Kelleher, J. & Etheridge, A.M. (2010). A new model for extinction and recolonisation in two dimensions: quantifying phylogeography. *Evolution*, *in press*.

Barton, N.H. & Wilson, I. (1995). Genealogies and geography. *Philosophical Transactions of the Royal Society of London Series B*, 349, 49–59.

Baudry, E. & Depaulis, F. (2003). Effect of misoriented sites on neutrality tests with outgroup. *Genetics*, 165(3), 1619–1622.

Beaumont, M.A., Zhang, W. & Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162, 2025–2026.

Becquet, C. & Przeworski, M. (2007). A new approach to estimate parameters of speciation models with application to apes. *Genome Research*, 17(10), 1505–1519.

Becquet, C. & Przeworski, M. (2009). Learning about modes of speciation from computational approaches. *Evolution*, 63(10), 2547–2562.

Bogdanowitcz, S.M., Wallner, W.E., Bell, J., O'Dell, T.M. & Harrison, R.G. (1993). Asian gypsy moth (Lepidoptera: Lymantriidae) in North America: evidence from molecular data. *Annals of the Entomological Society of America*, 86, 710–715.

Braverman, J.M., Hudson, R.R., Kaplan, N.L., Langley, C.H. & Stephan, W. (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, 140(2), 783–796.

Brower, A.V.Z. (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 Science of the United States of America*, 91, 6491–6495.

Burgess, R. & Yang, Z. (2008). Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Molecular Biology and Evolution*, 25(9), 1979–1994.

Campbell, B.C., Steffen-Campbell, J.D. & Werren, J.H. (1993). Phylogeny of the *Nasonia* (Hymenoptera: Pteromalidae) species complex inferred from an internal transcribed spacer (ITS2) and 28s rDNA sequences. *Insect Molecular Biology*, 2, 225–237.

Carstens, B.C. & Knowles, L.L. (2007a). Estimating phylogeny from gene tree probabilities in *Melanoplus* grasshoppers. *Systematic Biology*, 56, 400–411.

Carstens, B.C. & Knowles, L.L. (2007b). Shifting distributions and speciation: species divergence during rapid climate change. *Molecular Ecology*, 16(3), 619–627.

Carstens, B.C., Stoute, H.N. & Reid, N.M. (2009). An information-theoretical approach to phylogeography. *Molecular Ecology*, 18(20), 4270–4282.

Ceiridwen, J.E., Suchard, M., Lemey, P., Welch, J., Barns, I., Fulton, T.L., Barnett, R. O'Conell, T., Coxon, P., Monaghan, N., Valdioser, C.E., Baryshnikov, G.F., Rambaut, A., Thomas, M.G., Bradley, D.G. & Shapiro, B. (2010). Phylogenetic evidence for hybridisation between brown and polar bears during the late Pleistocene. *Nature*, *in review*.

Challis, R.J., Mutun, S., Nieves-Aldrey, J.L., Preuss, S., Rokas, A., Aebi, A., Sadeghi, E., Tavakoli, M. & Stone, G.N. (2007). Longitudinal range expansion and cryptic eastern species in the western palaearctic oak gallwasp *Andricus coriarius*. *Molecular Ecology*, 16(10), 2003–2014.

Charlesworth, B., Charlesworth, D. & Barton, N. (2003). The effects of genetic and geographic structure on neutral variation. *Annual Review of Ecology, Evolution, and Systematics*, 34, 99–125.

Chen, F.C. & Li, W.H. (2001). Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *American Journal of Human Genetics*, 68(2), 444–456.

Contreras-Diaz, H.G., Moya, O., Oromí, P. & Juan, C. (2007). Evolution and diversification of the forest and hypogean ground-beetle genus *Trechus* in the Canary Islands. *Molecular Phylogenetics and Evolution*, 42(3), 687–699.

Cook, J.M., Rokas, A., Pagel, M. & Stone, G.N. (2002). Evolutionary shift between host oak section and host-plant organs in *Andricus* gallwasps. *Evolution*, 56(9), 1821–1830.

Creer, S. (2007). Choosing and using introns in molecular phylogenetics. *Evolutionary Bioinformatics*, 3, 99–108.

Csóka, G., Stone, G.N. & Melika, G. (2005). The biology, ecology and evolution of gallwasps. In C. Raman, W. Schaefer & T.M. Withers, editors, *Biology, ecology and evolution of gall inducing insects*, pages 573–642. Science Publisher, Enfield, New Hampshire.

Culling, M.A., Janko, K., Boron, A., Vasil'ev, V. P.and Côté, I.M. & Hewitt, G.M. (2006). European colonization by the spined loach (*Cobitis taenia*) from Ponto-Caspian refugia based on mitochondrial DNA variation. *Molecular Ecology*, 15, 173–190.

Daniel, K. & Daniel, J. (1898). Beiträge zur Kenntnis der Gattung *Trechus* Clairville. *Coleopteren-Studien*, 2, 1–16.

Darwin, C. (1859). *The origin of species by means of natural selection*. John Murray, London.

Das, A., Mohanty, S. & Stephan, W. (2004). Inferring the population structure and demography of *Drosophila ananassae* from multilocus data. *Genetics*, 168(4), 1975–1985.

DeChaine, E.G. & Martin, A.P. (2006). Using coalescent simulation to test the impact of Quaternary climate cycles on divergence in an alpine plant-insect association. *Evolution*, 60(5), 1004–1013.

Degnan, J.H. & Rosenberg, N.A. (2009). Gene tree discordance, phylogenetic inference and the multi-species coalescent. *Trends in Ecology & Evolution*, 24(6), 332–340.

Degnan, J.H. & Salter, L.A. (1995). Gene tree distributions under the coalescent process. *Evolution*, 59(1), 24–37.

Depaulis, F., Mousset, S. & Veuille, M. (2003). Power of neutrality tests to detect bottlenecks and hitchhiking. *Journal of Molecular Evolution*, 57(0), S190–S200.

Di Rienzo, A., Donnelly, P., Toomajian, C., Sisk, B., Hill, A., Petzl-Erler, M.L., Haines, G.K. & Barch, D.H. (1998). Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. *Genetics*, 148(3), 1269–1284.

Din, W., Anand, R., Boursot, P., Darviche, D., Dod, B., Jouvin-Marche, E., Orth, A., Talwar, G., Cazenave, P.A. & Bonhomme, F. (1996). Origin and radiation of the house mouse: clues from nuclear genes. *Journal of Evolutionary Biology*, 9, 519–539.

Dowling, D.K., Friberg, U. & Lindell, J. (2008). Evolutionary implications of non-neutral mitochondrial genetic variation. *Trends in Ecology & Evolution*, 23(10), 546–554.

Drummond, A.J. & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7, 214.

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

Ebersberger, I., Galgoczy, P., Taudien, S., Taenzer, S., Platzer, M. & von Haeseler, A. (2007). Mapping human genetic ancestry. *Molecular Biology and Evolution*, 24(10), 2266–2276.

Edwards, S.V. (2009). Is a new and general theory of molecular systematics emerging? *Evolution*, 63(1), 1–19.

Emerson, B.C., Paradis, E. & Thebaud, C. (2001). Revealing the demographic histories of species using DNA sequences. *Trends in Ecology & Evolution*, 16(12), 707–716.

Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3, 87–112.

Faccini, S. & Sciaky, R. (2002). Note sulla variabilità morfologica dell'edeago in *Trechus modestus* Putzeys 1874 (Coleoptera, Carabidae, Trechinae). *Bollettino del Museo Regionale di Scienze Naturali di Torino*, 21, 103–113.

Fay, J.C. & Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, 155(3), 1405–1413.

Felsenstein, J. (1975). A pain in the torus: some difficulties with models of isoaltion by distance. *American Naturalist*, 109(976), 359–368.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17, 368–376.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics*, 22(1), 521–565.

Felsenstein, J. (1992). Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetical Research*, 59, 139–147.

Felsenstein, J. (2006). Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Molecular Biology and Evolution*, 23(3), 691–700.

Fisher, R.A. (1930). *The Genetical Theory of Natural Selection*. Clarenson, Oxford.

Focarile, A. (1949). 1o contributo alla conoscenza dei Trechini palearctici. *Bolletina della Soc. Entomol. Ital.*, 89, 71–77.

Focarile, A. (1950). 2o contributo alla concoscenza dei Trechini palearctici. *Bolletina della Soc. Entomol. Ital.*, 29, 52–67.

Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit 1 from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, 3, 294–299.

Forer, B. (1949). The fallacy of personal validation - a classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology*, 44, 118–123.

Fu, Y.X. (1996). New statistical tests of neutrality for DNA samples from a population. *Genetics*, 143(1), 557–570.

Fu, Y.X. & Li, W.H. (1993). Statistical tests of neutrality of mutations. *Genetics*, 133(3), 693–709.

Galtier, N., Depaulis, F. & Barton, N.H. (2000). Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics*, 155(2), 981–987.

Garrick, R.C., Rowell, D.M., Simmons, C.S., Hillis, D.M., Sunnucks, P. & Brown, J. (2009). Fine-scale phylogeographic congruence despite demographic incongruence in two low-mobility saproxylic springtails. *Evolution*, 62(5), 1103–1118.

Gifford, M.E. & Larson, A. (2008). In situ genetic differentiation in a hispaniolan lizard (*Ameiva chrysolaema*): a multilocus perspective. *Molecular Phylogenetics and Evolution*, 49(1), 277–291.

Glinka, S., Ometto, L., Mousset, S., Stephan, W. & De Lorenzo, D. (2003). Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: A multi-locus approach. *Genetics*, 165(3), 1269–1278.

Godfray, H.J.C. (1994). *Parasitoids. Behavioural and Evolutionary Ecology*. Princeton University Press, New Jersey.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Hober, B., Hoffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. & Paabo, S. (2010). A draft sequence of the Neanderthal genome. *Science*, 328(5979), 710–722.

Griffiths, R.C. (1981). The number of heterozygous loci bewteen two randomly chosen completely linked sequences of loci in two subdivided population models. *Journal of Mathematical Biology*, 12, 251–261.

Griffiths, R.C. (1991). The two-locus ancestral graph. In I.V. Basawa & R.I. Taylor, editors, *Selected Proceedings of the Symposium of Applied Probability*, pages 100–117. Institute of Mathematical Statistics, Haywards, CA, USA.

Griffiths, R.C. & Tavaré, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions: Biological Sciences*, 344(1310), 403–410.

Haddrill, P.R., Thornton, K.R., Charlesworth, B. & Andolfatto, P. (2005). Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Research*, 15(6), 790–799.

Halligan, D.L. & Keightley, P.D. (2006). Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Research*, 16(7), 875–884.

Hamblin, M.T., Mitchell, S.E., White, G.M., Gallego, J., Kukatla, R., Wing, R.A., Paterson, A.H. & Kresovich, S. (2004). Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. *Genetics*, 167(1), 471–483.

Harpending, H.C. (1994). Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Human Biology*, 66(4), 591–600.

Hayward, A. & Stone, G.N. (2005). Oak gall wasp communities: Evolution and ecology. *Basic and Applied Ecology*, 6(5), 435–443.

Hayward, A. & Stone, G.N. (2006). Comparative phylogeography across two trophic levels: the oak gall wasp *Andricus kollari* and its chalcid parasitoid *Megastigmus stigmatizans*. *Molecular Ecology*, 15(2), 479–489.

Hedrick, P.W. & Parker, J.D. (2003). Evolutionary genetics and genetic variation of haplodiploids and X-linked genes. *Annual Review of Ecology and Systematics*, 28(1), 55–83.

Heled, J. & Drummond, A.J. (2009). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3), 570–580.

Hellenthal, G., A., A. & Falush, D. (2008). Inferring human colonisation history using a copying model. *PLoS Genetics*, 4(5), e1000078.

Heuertz, M., De Paoli, E., Kallman, T., Larsson, H., Jurman, I., Morgante, M., Lascoux, M. & Gyllenstrand, N. (2006). Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of norway spruce [*Picea abies* (L.) Karst]. *Genetics*, 174(4), 2095–2105.

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

Hewitt, G.M. (1999). Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, 68(1-2), 87–112.

Hey, J. (2005). On the number of new world founders: a population genetics portrait of the peopling of the americas. *PLoS Biology*, 3(6), e193.

Hey, J. (2010a). The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. *Molecular Biology and Evolution*, 27, 921–933.

Hey, J. (2010b). Isolation with migration models for more than two populations. *Molecular Biology and Evolution*, 27, 905–920.

Hey, J. & Machado, C.A. (2003). The study of structured populations - new hope for a difficult and divided science. *Nature Reviews Genetics*, 4(7), 535–543.

Hey, J. & Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, 167(2), 747–760.

Hey, J. & Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America*, 104(8), 2785–2790.

Hickerson, M.J., Carstens, B.C., Cavender-Bares, J., Crandall, K.A., Graham, C.H., Johnson, J.B., Rissler, L., Victoriano, P.F. & Yoder, A.D. (2010). Phylogeography's past, present, and future: 10 years after Avise 2000. *Molecular Phylogenetics and Evolution*, 54(1), 291–301.

Hickerson, M.J., Stahl, E. & Takebayashi, N. (2007). msBayes: Pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation. *BMC Bioinformatics*, 8(1), 268.

Hickerson, M.J., Stahl, E.A., Lessios, H.A. & Crandall, K. (2006). Test for simulatenous divergence using approximate Bayesian computation. *Evolution*, 60(12), 2435–2453.

Higgins, D.J. & Sharp, P.M. (1988). Clustal: a package for performing multiple sequence alignments. *Gene*, 73, 273–244.

Hillis, D.M. & Bull, J.J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Zoology*, 42, 182–192.

Ho, S.Y.W., Phillips, M.J., Cooper, A. & Drummond, A.J. (2005). Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Molecular Biology and Evolution*, 22(7), 1561–1568.

Hobolth, A., Christensen, O.F., Mailund, T. & Schierup, M.H. (2007). Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genetics*, 3(2), e7.

Hodkinson, I.D. (2005). Terrestrial insects along elevation gradients: species and community response to altitude. *Biological Reviews*, 80, 489–513.

Holdhaus, K. (1954). Die Spuren der Eiszeit in der Tierwelt Europas. *Abhandlungen der Zoologisch-Botanischen Gesellschaft in Wien*, 18, 1–493.

Hudson, R.R. (1983). Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, 37, 203–217.

Hudson, R.R. (1993). *The How and Why of Generating Genealogies*. Japan Scientific Societies Press, Tokyo and Sinauer Associates.

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

Hudson, R.R., Boos, D.D. & Kaplan, N.L. (1992). A statsitiscal test for detecting geographic subdivison. *Molecular Biology and Evolution*, 9, 138–151.

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

Hudson, R.R. & Turelli, M. (2003). Stochasticity overrules the "three-times rule": genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. *Evolution*, 57(1), 182–190.

Hurst, G.D.D. & Jiggins, F.M. (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proceedings of the Royal Society B: Biological Sciences*, 272(1572), 1525–1534.

Innan, H., Zhang, K., Marjoram, P., Tavaré, S. & Rosenberg, N.A. (2005). Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. *Genetics*, 169(3), 1763–1777.

Irwin, D.E. (2002). Phylogeographic breaks without geographic barriers to gene flow. *Evolution*, 56(12), 2383–2394.

Jäckli, H. (1970). Die Schweiz zur letzten Eiszeit. *Eidgenössische Landestopographie*.

Janetschek, H. (1956). Das Problem der inneralpinen Eiszeitüberdauerung durch Tiere. *Zoologische Jahrbücher. Abteilung für Systematik, Geographie und Biologie der Tiere*, 70(177-226).

Jeannel, R. (1927). Monographie des Trechinae. morphologie comparee et distribution geographique d'une group des coleopters. *L'Abeille*, 32, 1–992.

Jennings, W.B. & Edwards, S.V. (2005). Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. *Evolution*, 59(9), 2033–2047.

Jesus, F.F., Wilkins, J.F., Solferini, V.N. & Wakeley, J. (2006). Expected coalescence times and segregating sites in a model of glacial cycles. *Genetics and Molecular Research*, 5(3), 466–474.

Jordan, S.D. (1905). The origin of species through isolation. *Science*, 22, 545–562.

Juste, J., Ibáñez, C., Muñoz, J., Trujillo, D., Benda, P., Karatas, A. & Ruedi, M. (2004). Mitochondrial phylogeography of the long-eared bats (*Plecotus*) in the mediterranean Palaearctic and Atlantic islands. *Molecular Phylogenetics and Evolution*, 31(3), 1114–1126.

Kalendar, R., Lee, D. & Schulman, A.H. (2009). FastPCR software for PCR primer and probe design and repeat search. *Genes, Genomes and Genomics*, 3(1).

Kass, R.E. & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.

Keightley, P.D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S. & Blaxter, M.L. (2009). Analysis of the genome sequences of three drosophila melanogaster spontaneous mutation accumulation lines. *Genome Research*, 19(7), 1195–1201.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to the steady flux of mutations. *Genetics*, 61, 893–903.

Kingman, J.F.C. (1982). The coalescent. *Stochastic Processes and their Applications*, 13, 235–248.

Kliman, R.M., Andolfatto, P., Coyne, J.A., Depaulis, F., Kreitman, M., Berry, A.J., McCarter, J., Wakeley, J. & Hey, J. (2000). The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics*, 156(4), 1913–1931.

Klinka, J. & Zink, R.M. (1997). The importance of recent ice ages in speciation: a failed paradigm. *Science*, 277(5332), 1666–1669.

Knowles, L.L. (2001). Did the Pleistocene glaciations promote divergence? tests of explicit refugial models in montane grasshopprers. *Molecular Ecology*, 10(3), 691–701.

Knowles, L.L. (2002). Statistical phylogeography. *Molecular Ecology*, 11, 2623–2635.

Knowles, L.L. (2004). The burgeoning field of statistical phylogeography. *Journal of Evolutionary Biology*, 17(1), 1–10.

Knowles, L.L. (2008). Why does a method that fails continue to be used? *Evolution*, 62, 2713–2717.

Knowles, L.L. (2009). Statistical phylogeography. *Annual Review of Ecology, Evolution, and Systematics*, 40(1), 593–612.

Koch, M.A., Kiefer, C. & Ehrlich, D. (2006). Three times out of Asia Minor: the phylogeography of *Arabis alpina* l. (Brassicaceae). *Molecular Ecology*, 15, 825–839.

Kubatko, L.S., Carstens, B.C. & Knowles, L.L. (2009). STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, 25(7), 971–973.

Kuhner, M.K., Yamato, J. & Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, 140(4), 1421–1430.

Kuo, C.H. & Avise, J. (2005). Phylogeographic breaks in low-dispersal species: the emergence of concordance across gene trees. *Genetica*, 124(2), 179–186.

Kuo, L. & Mallick, B. (1998). Variable selection for regression models. *Sankhya: The Indian Journal of Statistics, Series B*, 60(1), 65–81.

Lamm, K.S. & Redelings, Benjamin, D. (2009). Reconstructing ancestral ranges in historical biogeography: properties and prospects. *Journal of Systematics and Evolution*, 47(5), 369–382.

Lee, J.Y., Edwards, S.V. & Webster, M. (2009). Divergence across Australia's Carpentarian barrier: statistical phylogeography of the red-backed fairy wren (*Malurus melanocephalus*). *Evolution*, 62(12), 3117–3134.

Lemey, P., Rambaut, A., Drummond, A.J. & Suchard, M.A. (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology*, 5(9), e1000520.

Lennartsson, T. (2002). Extinction thresholds and disrupted plant-pollinator interactions in fragmented plant populations. *Ecology*, 83, 3060–3072.

Lessa, E.P. (1992). Rapid surveying of DNA sequence variation in natural populations. *Molecular Biology and Evolution*, 9(2), 323–330.

Leys, R., Cooper, S.J.B. & Schwarz, M.P. (2002). Molecular phylogeny and historical biogeography of the large carpenter bees, genus *Xylocopa* (Hymenoptera: Apidae). *Biological Journal of the Linnean Society*, 77(2), 249–266.

Li, Y.J., Satta, Y. & Takahata, N. (1999). Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method. *Genes & Genetic Systems*, 74(4), 117–127.

Liu, L. & Pearl, D.K. (2007). Species trees from gene trees: reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology*, 56(3), 504–514.

Lohse, K. (2009). Can mtDNA barcodes be used to delimit species? A response to Pons et al. (2006). *Systematic Biology*, 58(4), 439–442.

Lompe, A. (2004). Trechini. In H. Freude, K.W. Harde, G.A. Lohse & B. Klausnitzer, editors, *Die Käfer Mitteleuropas, Vol. 2*, pages 108–149. Spectrum Verlag, Heidelberg/Berlin.

Lopez-Vaamonde, C., Rasplus, Y.J., Weiblen, G. & Cook, J.M. (2001). Molecular phylogenies of fig wasps: partial co-cladogenesis of pollinators and parasites. *Molecular Phylogenetics and Evolution*, 21, 55–71.

Machado, C.A., Robbins, N., Gilbert, M.T.P. & Herre, E.A. (2005). Critical review of host specificity and its coevolutionary implications in the fig/fig-wasp mutualism. *Proceedings of the National Academy of Sciences of the United States of America*, 102(Suppl 1), 6558–6565.

Maddison, W.P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3), 523–536.

Maddison, W.P. & Knowles, L.L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, 55(1), 21–30.

Malécot, G. (1969). *The Mathematics of Heredity*. WF Freeman, San Francisco.

Margraf, N., Verdon, A., Rahier, M. & Naisbit, R.E. (2007). Glacial survival and local adaptation in an alpine leaf beetle. *Molecular Ecology*, 16(11), 2333–2343.

Matsen, F.A. & Wakeley, J. (2006). Convergence to the island-model coalescent in populations with restricted migration. *Genetics*, 172(1), 701–708.

Mena-Correa, J., Sivinski, J., Anzures-Dadda, A., Ramìrez-Romero, R., Gates, M. & Aluja, M. (2009). Consideration of *Eurytoma sivinskii* (Gates and Grissell), a eurytomid (Hymenoptera) with unusual foraging behaviors, as a biological control agent of tephritid (Diptera) fruit flies. *Biological Control*, 53(1), 9–17.

Michaux, J., Libois, R., E., P. & M.-G., F. (2004). Phylogeographic history of the yellow-necked field-mouse (*Apodemus flavicollis*) in Europe and in the Near and Middle East. *Molecular Phylogenetics and Evolution*, 32, 788–798.

Moya, O., Contreras-Diaz, H.G., Oromi, P. & Juan, C. (2004). Genetic structure, phylogeography and demography of two ground-beetle species endemic to the Tenerife laurel forest (Canary Islands). *Molecular Ecology*, 13(10), 3153–3167.

Muster, C., Maddison, W.P., Uhlman, S., Berendonk, T.U. & Vogler, A.P. (2009). Arctic-alpine distributions - metapopulations on a continental scale? *The American Naturalist*, 173(3), 313–326.

Nee, S., Holmes, E.C., Rambaut, A. & Harvey, P.H. (1995). Inferring population history from molecular phylogenies. *Philosophical Transactions of the Royal Society of London Series B*, 349(25-31).

Nei, M. & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3, 418–426.

Nei, M. & Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

Nepokroeff, M., Sytsma, K.J., Wagner, W.L. & Zimmer, E.A. (2003). Reconstructing ancestral patterns of colonization and dispersal in the Hawaiian understory tree genus *Psychotria* (Rubiaceae): a comparison of parsimony and likelihood approaches. *Systematic Biology*, 52(6), 820–838.

Nicholls, J.A., Preuss, S., Hayward, A., Melika, G., Csóka, G., Nieves-Aldrey, J.L., Askew, R.R., Tavakoli, M., Schönrogge, K. & Stone, G.N. (2010). Concordant phylogeography and cryptic speciation in two western Palaearctic oak gall parasitoid species complexes. *Molecular Ecology*, 19, 592–609.

Nichols, R. (2001). Gene trees and species trees are not the same. *Trends in Ecology & Evolution*, 16(7), 358–364.

Nielsen, R. & Beaumont, M.A. (2009). Statistical inferences in phylogeography. *Molecular Ecology*, 18(6), 1034–1047.

Nielsen, R. & Wakeley, J. (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, 158, 885–896.

Nordborg, M. (1998). On the probability of Neanderthal ancestry. *American Journal of Human Genetics*, 63(4), 1237–40.

Obbard, D.J., Welch, J.J., Kim, K.W. & Jiggins, F.M. (2009). Quantifying adaptive evolution in the *Drosophila* immune system. *PLoS Genet*, 5, e1000698.

Ochman, H. & Wilson, A.C. (1987). Evolution in bacteria - evidence for a universal substition rate in cellular genomes. *Journal of Molecular Evolution*, 26, 74–86.

Oliveira, D.C.S.G., Raychoudhury, R., Lavrov, D.V. & Werren, J.H. (2008). Rapidly evolving mitochondrial genome and directional selection in mitochondrial genes in the parasitic wasp *Nasonia* (Hymenoptera: Pteromalidae). *Molecular Biology and Evolution*, 25(10), 2167–2180.

Ometto, L., Glinka, S., De Lorenzo, D. & Stephan, W. (2005). Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Molecular Biology and Evolution*, 22(10), 2119–2130.

Pagel, M., Meade, A. & Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology*, 53(5), 673–684.

Palumbi, S.R. & S., B.C. (1994). Contasting population structure from nuclear intron sequence and mtDNA of humpback whales. *Molecular Biology and Evolution*, 11(3), 426–435.

Pamilo, P. & Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5(5), 568–583.

Panchal, M. & Beaumont, M.A. (2007). The automation and evaluation of nested clade analysis. *Evolution*, 61, 1466–1480.

Papanicolaou, A., Joron, M., Mcmillan, W.O., Blaxter, M.L. & Jiggins, C.D. (2005). Genomic tools and cDNA derived markers for butterflies. *Molecular Ecology*, 14(9), 2883–2897.

Parker, S.R. (1997). Sequence navigator. multiple sequence alignment software. *Methods in Molecular Biology*, 70, 145–54.

Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S. & Reich, D. (2006). Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441(7097), 1103–1108.

Pauls, S.U., Lumbsch, H.T. & Haase, P. (2006). Phylogeography of the montane caddisfly *Drusus discolor*: evidence for multiple refugia and periglacial survival. *Molecular Ecology*, 15(8), 2153–2169.

Pauw, A. (2007). Collapse of a pollination web in small conservation areas. *Ecology*, 88, 1759–1769.

Peters, J.L., Zhuravlev, Y.N., Fefelov, I., Humphries, E.M. & Omland, K.E. (2008). Multilocus phylogeography of a holarctic duck: colonization of North America from Eurasia by gadwall (*Anas strepera*). *Evolution*, 62(6), 1469–1483.

Pluzhnikov, A., Di Rienzo, A. & Hudson, R.R. (2002). Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics*, 161(3), 1209–1218.

Pollard, D.A., Iyer, V.N., Moses, A.M. & Eisen, M.B. (2006). Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genetics*, 2(10), e173.

Pulquério, M. & Nicholls, R.A. (2007). Dates from the molecular clock: how wrong can we be? *Trends in Ecology & Evolution*, 22(4).

Pybus, O.G., Rambaut, A., Holmes, E.C. & Harvey, P.H. (2002). New inferences from tree shape: numbers of missing taxa and population growth rates. *Systematic Biology*, 51(6), 881–888.

Rambaut, A. & Drummond, A.J. (2007). Tracer v1.4.

Ramirez-Soriano, A., Ramos-Onsins, S.E., Rozas, J., Calafell, F. & Navarro, A. (2008). Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics*, 179(1), 555–567.

Ramos-Onsins, S.E., Mousset, T., Mitchell-Olds, T. & Stephan, W. (2007). Population genetic inference using a fixed number of segregating sites: a reassessment. *Genetical Reserach*, 89, 231–244.

Ramos-Onsins, S.E. & Rozas, J. (2002). Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution*, 19(12), 2092–2100.

Rannala, B. & Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4), 1645–1656.

Raychoudhury, R., Baldo, L., Oliveira, D.C.S.G., Werren, J.H. & Wayne, M. (2009). Modes of aquisition of *Wolbachia*: horizontal transfer, hybrid introgression and codivergence in the *Nasonia* complex. *Evolution*, 63(1), 165–183.

Reich, D., Feldman, M. & Goldstein, D. (1999). Statistical properties of two tests that use multilocus data sets to detect population expansions. *Molecular Biology and Evolution*, 16(4), 453–466.

Reitter (1908). *Fauna Germanica - Die Käfer des deutschen Reiches*, volume 1. K. G. Lutz Verlag, Stuttgart.

Rokas, A., Atkinson, R., Brown, G., West, S.A. & Stone, G.N. (2001). Understanding patterns of genetic diversity in the oak gallwasp *Biorhiza pallida*: demographic history or a *Wolbachia* selective sweep? *Heredity*, 87, 294–305.

Rokas, A., Nylander, J.A., Ronquist, F. & Stone, G.N. (2002). A maximum-likelihood analysis of eight phylogenetic markers in gallwasps (Hymenoptera: Cynipidae): implications for insect phylogenetic studies. *Molecular Phylogenetics and Evolution*, 22, 1055–7903.

Rokas, A., Atkinson, R.J., Webster, L., Csóka, G. & Stone, G.N. (2003). Out of Anatolia: longitudinal gradients in genetic diversity support an eastern origin for a circum-mediterranean oak gallwasp *Andricus quercustozae*. *Molecular Ecology*, 12(8), 2153–2174.

Ronquist, F. (2004). Bayesian inference of character evolution. *Trends in Ecology & Evolution*, 19(9), 475–481.

Rosenberg, N. (2002). The probability of topological concordance of gene trees and species trees. *Theoretical Population Biology*, 61, 225–247.

Rozas, J. & Rozas, R. (1995). DNAsp, DNA sequence polymorphism: an interactive program for estimating population genetics parameteres from DNA sequence data. *Computer Applications in the Biosciences*, 11, 621–625.

Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X. & Rozas, R. (2003). DNAsp, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, 19(18), 2496–2497.

Rozen, S. & Skaletsky, H.J. (2000). Primer3 on the WWW for general users and biologist programmers. In S. Krawetz & S. Misener, editors, *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, pages 365–386. Humana Press, NJ.

Saitou, N. & Nei, M. (1986). The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. *Journal of Molecular Evolution*, 24(1), 189–204.

Satta, Y., Klein, J. & Takahata, N. (2000). DNA archives and our nearest relative: the trichotomy problem revisited. *Molecular Phylogenetics and Evolution*, 14(2), 259–275.

Saunders, I.W., Tavaré, S. & Watterson, G.A. (1984). On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability*, 16, 471–491.

Schaeffer, S.W. (2002). Molecular population genetics of sequence length diversity in the Adh region of *Drosophila melanogaster*. *Genetical Reserach*, 80, 163–175.

Schierup, M.H. & Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 156(2), 879–891.

Schmitt, T. (2009). Biogeographical and evolutionary importance of the European high mountain systems. *Frontiers in Zoology*, 6(1), 9.

Schmitt, T. & Hewitt, G. (2006). Disjunct distributions during glacial and intergalcial periods in mountain butterflies: *Erebia epiphron* as an example. *Journal of Evolutionary Biology*, 19(1), 108–113.

Schmölzer, K. (1962). Die Kleintierwelt der Nunatakker als Zeugen einer Eiszeitüberdauerung. *Mitteilungen des Zoologischen Museums in Berlin*, 38(2), 172–400.

Schneider, S., Roessli, D. & Excoffier, L. (2000). *Arlequin: a software for population genetics data analysis. Ver 2.000*. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva, Geneva.

Schneider, S. & Excoffier, L. (1999). Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics*, 152(3), 1079–1089.

Schönmann, R. (1937). Die Artsystematik und Verbreitung der hochalpinen Trechini der Ostalpen. *Zoologische Jahrbücher. Abteilung für Systematik, Geographie und Biologie der Tiere*, 70, 178–226.

Schönrogge, K., Stone, G.N. & Crawley, M.J. (1996a). Abundance patterns and species richness of the parasitoids and inquilines of the alien gall-former *Andricus quercuscalicis* (Hymenoptera: Cynipidae). *Oikos*, 77(3), 507–518.

Schönrogge, K., Stone, G.N. & 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(1), 51–60.

Schönrogge, K., Stone, G.N. & Crawley, M.J. (1996b). Alien herbivores and native parasitoids: rapid development of guild structure in an invading gall wasp, *Andricus quercuscalicis* (Hymenoptera: Cynipidae). *Ecological Entomology*, 21, 71–80.

Schönrogge, K., Walker, P. & Crawley, M.J. (1998). Invaders on the move: parasitism in the galls of four alien gall wasps in Britain (Hymenoptera, Cynipidae). *Proceedings of the Royal Society B: Biological Sciences*, 256, 1643–1650.

Schönswetter, P., Stehlik, I., Holderegger, R. & Tribsch, A. (2005). Molecular evidence for glacial refugia of mountain plants in the European Alps. *Molecular Ecology*, 14(11), 3547–3555.

Schönswetter, P., Tribsch, A., Barfuss, M. & Niklfeld, H. (2002). Several Pleistocene refugia detected in the high alpine plant *Phyteuma globulariifolium* Sternb. & Hoppe (Campanulaceae) in the European Alps. *Molecular Ecology*, 11(12), 2637–2647.

Schweiger, H. (1969). Gebirgssysteme als Zentren der Artenbildung. *Deutsche Entomologische Zeitschrift*, 16, 159–174.

Sha, Z.L., Zhu, C.D., Murphy, R.W. & Huang, D.W. (2007). *Diglyphus isaea* (Hymenoptera: Eulophidae): a probable complex of cryptic species that forms an important biological control agent of agromyzid leaf miners. *Journal of Zoological Systematics and Evolutionary Research*, 45(2), 128–135.

Sharanowski, B.J., Robbertse, B., Walker, J., Voss, S.R., Yoder, S.R., Spatafora, J. & Sharkey, M.J. (2010). Expressed sequence tags reveal Proctotrupomorpha (minus Chalcidoidea) as sister to Aculeata (Hymenoptera: Insecta). *Molecular Phylogenetics and Evolution*, *accepted*.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H. & Flook, P. (1994). Evolution, weighting and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Annals of the Entomological Society of America*, 87, 651–701.

Simonsen, K.L., Churchill, G.A. & Aquadro, C.F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, 141(1), 413–429.

Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical Research*, 58, 167–75.

Slatkin, M. & Hudson, R.R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, 129(2), 555–562.

Slatkin, M. & Maddison, W.P. (1998). A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics*, 123, 603–613.

Slatkin, M. & Pollack, J.L. (2006). The concordance of gene trees and species trees at two linked loci. *Genetics*, 172(3), 1979–1984.

Slatkin, M. & Pollack, J.L. (2008). Subdivision in an ancestral species creates asymmetry in gene trees. *Molecular Biology and Evolution*, 25(10), 2241–2246.

Sota, T. & Vogler, A.P. (2001). Incongruence of mitochondrial and nuclear gene trees in the carabid beetles *Ohomopterus*. *Systematic Biology*, 50(1), 39–59.

Stehlik, I., Blattner, F.R., Holderegger, R. & Bachmann, K. (2002). Nunatak survival of the high alpine plant *Eritrichium nanum* (l.) Gaudin in the central Alps during the ice ages. *Molecular Ecology*, 11(10), 2027–2036.

Stephan, W. & Li, H. (2006). The recent demographic and adaptive history of *Drosophila melanogaster*. *Heredity*, 98(2), 65–68.

Stone, G.N., Atkinson, R., Rokas, A., Csóka, G. & Nieves-Aldrey, J.L. (2001). Differential success in northwards range expansion between ecotypes of the marble gallwasp *Andricus kollari*: a tale of two lifecycles. *Molecular Ecology*, 10, 761–778.

Stone, G.N., Challis, R.J., Atkinson, R.J., Csóka, G., Hayward, A., Melika, G., Mutun, S., Preuss, S., Rokas, A., Sadeghi, E. & Schönrogge, K. (2007). The phylogeographical clade trade: tracing the impact of human-mediated dispersal on the colonization of northern Europe by the oak gallwasp *Andricus kollari*. *Molecular Ecology*, 16, 2768–2781.

Stone, G.N., Hernandez-Lopez, A., Nicholls, J.A., di Pierro, E., Pujade-Villar, J., Melika, G., Cook, J.M. & Abbot, P. (2009). Extreme host plant conservatism during at least 20 million years of host plant pursuit by oak gallwasps. *Evolution*, 63(4), 854–869.

Stone, G.N. & Sunnucks, P. (1993). Genetic consequences of an invasion through a patchy environment - the cynipid gallwasp *Andricus quercuscalicis* (Hymenoptera: Cynipidae). *Molecular Ecology*, 2(4), 251–268.

Stone, G.N., van der Ham, R.W.J.M. & Brewer, J.G. (2008). Fossil oak galls preserve ancient multitrophic interactions. *Proceedings of the Royal Society B: Biological Sciences*, 275(1648), 2213–2219.

Strasburg, J.L. & Riesenber, L.H. (2009). How robust are isolation with migration analyses to violations of the IM model? a simulation study. *Molecular Biology and Evolution*, 27(2), 297–310.

Suchard, M., Weis, Robert, E. & Sinsheimer, J.S. (2001). Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular Biology and Evolution*, 18, 1001–1013.

Sunnucks, P., Blacket, M., Taylor, J., C.J., S., Ciavaglia, S., Garrick, R., Tait, N. & Pavlova, A. (2006). A tale of two flatties: different responses to past environmental fluctuations at Tallaganda in montane southeastern Australia. *Molecular Ecology*, 15, 4513–4531.

Swofford, D.L. (2001). Paup*. phylogenetic analysis using parsimony (*and other methods). version 4.1. *Sinauer Associates, Sunderland, Massachusetts.*

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

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), 585–595.

Tajima, F. (1983). Evolutionary relationships of DNA sequences in finite populations. *Genetics*, 105(2), 437–460.

Tajima, F. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics*, 135, 599–607.

Takahata, N. & Nei, M. (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*, 110(2), 325–344.

Takahata, N., Satta, Y. & Klein, J. (1995). Divergence time and population size in the lineage leading to modern humans. *Theoretical Population Biology*, 48, 198–221.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, 24, 1596–1599.

Tamura, K., Subramanian, S. & Kumar, S. (2004). Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Molecular Biology and Evolution*, 21(1), 36–44.

Tanabe, K., Mita, T., Jombart, T., Eriksson, A., Horibe, S., Palacpac, N., Ranford-Cartwright, L., Sawai, H., Sakihama, N., Ohmae, H., Nakamura, M., Ferreira, M.U., Escalante, A.A., Prugnolle, F., Björkman, A., Färnert, A., Kaneko, A., Horii, T., Manica, A., Kishino, H. & Balloux, F. (2010). *Plasmodium falciparum* accompanied the human expansion out of Africa. *Current Biology*, *in press*.

Tavaré, S. (1984). Lines-of-descent and genealogical processes and their application in population genetic models. *Theoretical Population Biology*, 26, 119–164.

Templeton, A.R. (2010). Coalescent-based, maximum likelihood inference in phylogeography. *Molecular Ecology*, 19(3), 431–446.

Templeton, A.R., Routman, E. & Phillips, C.A. (1995). Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics*, 140(2), 767–782.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994). ClustalW - inproving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap-penalities and weight matrix choice. *Nucleic Acids Research*, 22, 4673–4680.

Thornton, K. & Andolfatto, P. (2006). Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics*, 172(3), 1607–1619.

Untergasser, A., Nijveen, H., Xiangyu, R., Bisseling, T., Geurts, R. & Leunissen, J.A.M. (2007). Primer3Plus, an enhanced web interface to primer3. *Nucleic Acids Research*, 35, W71–W74.

Uyenoyama, M.K. (1997). Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. *Genetics*, 147(3), 1389–1400.

van der Ham, R.W.J.M., Kuijper, W.J., Kortselius, M.J.H., van der Burgh, J., Stone, G.N. & Brewer, J.G. (2008). Plant remains from the Kreftenheye formation (Eemian) at Raalte, the Netherlands. *Vegetation History and Archaeobotany*, 17, 127–144.

Velichko, A.A., Novenko, E.Y., Pisareva, V.V., Zelikson, E.M., Boettger, T. & Junge, F.W. (2005). Vegetation and climate changes during the Eemian interglacial in central and eastern Europe: comparative analysis of pollen data. *Boreas*, 34(2), 207–219.

Villablanca, F.X., Roderick, G.K. & Palumbi, S.R. (1998). Invasion genetics of the mediterranean fruit fly: variation in multiple nuclear introns. *Molecular Ecology*, 7(5), 547–560.

Wakeley, J. (1996). Pairwise differences under a general model of subdivision. *Journal of Genetics*, 75(1), 81–89.

Wakeley, J. (1998). Segregating sites in Wright's island model. *Theoretical Population Biology*, 53(2), 166–174.

Wakeley, J. (1999). Nonequilibrium migration in human history. *Genetics*, 153, 1863–1871.

Wakeley, J. (2001). The coalescent in an island model of population subdivision with variation among demes. *Theoretical Population Biology*, 59(2), 133–144.

Wakeley, J. (2004a). Metapopulation models for historical inference. *Molecular Ecology*, 13(4), 865–875.

Wakeley, J. (2004b). Recent trends in population genetics: more data! more math! simple models? *Heredity*, 95(5), 397–405.

Wakeley, J. (2008). Complex speciation of humans and chimp. *Nature*, 452, E3–E4.

Wakeley, J. (2009). *Coalescent theory*. Roberts & Company Publishers, Greenwood Village, Colorado.

Wakeley, J. & Aliacar, N. (2001). Gene genealogies in a metapopulation. *Genetics*, 159, 893–905.

Wall, J.D. (2003). Estimating ancestral population sizes and divergence times. *Genetics*, 163(1), 395–404.

Wallace, A.R. (1876). *The geographic distributions of animals, with a study of the relations of living and extinct faunas as elucidating the past changes of the Earth's surface*. Harper and Brothers, New York.

Wang, Y. & Hey, J. (2010). Estimating divergence parameters with small samples from a large number of loci. *Genetics*, 184, 363–373.

Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7, 239–276.

Weiblen, G.D. (2002). How to be a fig wasp. *Annual Review of Entomology*, 47(1), 299–330.

Weinert, L.A., Werren, J.H., Aebi, A., Stone, G.N. & Jiggins, F.M. (2009). Evolution and diversity of *Rickettsia* bacteria. *BMC Biology*, 7(1), 6.

Weiss, G.H. & Kimura, M. (1965). A mathematical analysis of the stepping stone model of genetic correlation. *Journal of Applied Probability*, 2, 129–49.

Whitlock, M.C. & Barton, N. (1997). The effective size of a subdivided population. *Genetics*, 146, 427–441.

Whitlock, M.C. & McCauley, D.E. (1999). Indirect measures of gene flow and migration: FST not equal to 1/(4Nm+1). *Heredity*, 82(2), 117–125.

Wild, A.L. & Maddison, D.R. (2008). Evaluating nuclear protein-coding genes for phylogenetic utility in beetles. *Molecular Phylogenetics and Evolution*, 48(3), 877–891.

Wilder, J.A. & Hollocher, H. (2003). Recent radion of endemic Caribbean *Drosophila* of the *dunni* subgroup inferred from mutlilocus DNA sequence variation. *Evolution*, 57(11), 2566–2579.

Wilkins, J.F. (2004). A separation-of-timescales approach to the coalescent in a continuous population. *Genetics*, 168(4), 2227–2244.

Wilkinson-Herbots, H.M. (2008). The distribution of the coalescence time and the number of pairwise nucleotide differences in the "isolation with migration" model. *Theoretical Population Biology*, 73(2), 277–288.

Won, Y.J., Sivasundar, Y., Wang, Y. & Hey, J. (2005). On the origin of lake Malawi chilcid species: a population genetic analysis of divergence. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 6581–6586.

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

Wright, S. (1943). Isolation by distance. *Genetics*, 28(2), 114–138.

Wright, S. (1951). The genetic structure of populations. *Annals of Eugenics*, 15, 323–354.

Yang, Z. (2002). Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 162(4), 1811–1823.

Yang, Z. (2010). A likelihood ratio test of speciation with gene flow using genomic data. *Genome Biology and Evolution*, 2, 200–211.

Zhang, D.X. & Hewitt, G.M. (2003). Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology*, 12(3), 563–584.

Zhou, R., Zeng, K., Wu, W., Chen, X., Yang, Z., Shi, S. & Wu, C.I. (2007). Population genetics of speciation in nonmodel organisms: I. ancestral polymorphism in mangroves. *Molecular Biology and Evolution*, 24(12), 2746–2754.

Measuring the degree of starshape in genealogies – summary statistics and demographic inference

KONRAD LOHSE* AND JEROME KELLEHER
Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK

(*Received 27 August 2008 and in revised form 4 February 2009*)

Summary

The degree of starshape of a genealogy is readily detectable using summary statistics and can be taken as a surrogate for the effect of past demography and other non-neutral forces. Summary statistics such as Tajima's D and related measures are commonly used for this. However, it is well known that because of their neglect of the genealogy underlying a sample such neutrality tests are far from ideal. Here, we investigate the properties of two types of summary statistics that are derived by considering the genealogy: (i) genealogical ratios based on the number of mutations on the rootward branches, which can be inferred from sequence data using a simple algorithm and (ii) summary statistics that use properties of a perfectly star-shaped genealogy. The power of these measures to detect a history of exponential growth is compared with that of standard summary statistics and a likelihood method for the single and multi-locus case. Statistics that depend on pairwise measures such as Tajima's D have comparatively low power, being sensitive to the random topology of the underlying genealogy. When analysing multi-locus data, we find that the genealogical measures are most powerful. Provided reliable outgroup information is available they may constitute a useful alternative to full likelihood estimation and standard tests of neutrality.

1. Introduction

The motivation for studying the impact of past demography on sequence data is two-fold. Firstly, changes in population size are interesting in their own right, being intimately linked to processes such as speciation or geographic range shifts. Secondly, the standard neutral model (SNM) of a randomly mating Wright–Fisher population of constant size and discrete generations, hardly ever describes the patterns of diversity found in natural populations. Thus, studies aiming to detect loci under selection are faced with the considerable challenge of fitting realistic demographic models against which selection can be tested e.g. Glinka *et al.* (2003), Hamblin *et al.* (2004), Haddrill *et al.* (2005), Ometto *et al.* (2005) and Thornton & Andolfatto (2006). Since the rate of coalescence is inversely proportional to the effective population size, it is clear that demographic changes must leave a detectable signature in genealogies (Felsenstein, 1992). In general, positive population growth distorts genealogies towards a starshape with shorter internal branches, resulting in more low frequency variants and a unimodal rather than multi-peaked mismatch distribution (Slatkin & Hudson, 1991; Harpending, 1994; Schneider & Excoffier, 1999). In contrast to selective processes that act on single genetic variants, demography affects the whole genome, so one expects to find a concordant signature across loci (Tajima, 1989; Galtier *et al.*, 2000).

Approaches to demographic inference fall into three broad categories; for a review see Emerson *et al.* (2001). Firstly, likelihood methods, which are available for bottleneck and exponential growth models, make use of all the information in a sample by integrating over a large set of likely genealogies (Griffiths & Tavaré, 1994; Kuhner *et al.*, 1995). Although optimal in terms of statistical power and accuracy, likelihood estimation is computationally intensive and requires a fully specified alternative model. Therefore realistic growth histories often remain analytically intractable. Secondly, there are tree-based methods, which take the branch length information of a reconstructed tree as their starting point. Assuming that

* Corresponding author. Tel: +44 (0)131 650 5508. e-mail: K.R.Lohse@sms.ed.ac.uk

sequence evolution is clock-like, the number of lineages can be plotted against time and the shape of this trajectory compared with its neutral expectation (Nee *et al.*, 1995; Pybus *et al.*, 2002). Despite their conceptual appeal, these methods neglect any uncertainty in tree topology and are thus only as good as the reconstructed tree they are based on. Furthermore they cannot deal with recombination by definition. Finally, there are classical neutrality tests, most of which do not explicitly consider the genealogy but instead use more immediate aspects of the data such as the frequency spectrum of mutations, e.g. Tajima's D (Tajima, 1989) and Fu & Li's D (hereafter referred to as D_2) (Fu & Li, 1993), the haplotype distribution, e.g. Fu's F_S (Fu, 1996; Innan *et al.*, 2005), or the mismatch distribution, e.g. the raggedness statistic (Slatkin & Hudson, 1991). Compared with likelihood estimation, summary statistics are straightforward to calculate and their distribution can be simulated under almost any growth model.

Considering the zoo of statistics available and their wide use, there are surprisingly few studies that systematically compare their power, and those that do mainly consider bottlenecks and single locus data (Simonsen *et al.*, 1995; Fu, 1996; Ramos-Onsins & Rozas, 2002; Depaulis *et al.*, 2003; Ramirez-Soriano *et al.*, 2008). However, joint analysis of multiple loci is not only necessary to distinguish between selective and demographic events (Galtier *et al.*, 2000) but also potentially far more powerful than inferences based on a single locus. An added advantage of multi-locus analysis is that both means and variances of summary statistics can be used for testing. Variance based tests were first developed for microsatellite data (Di Rienzo *et al.*, 1998; Reich *et al.*, 1999) but are now routinely used to analyse sequence data from multiple loci (Pluzhnikov *et al.*, 2002; Haddrill *et al.*, 2005; Heuertz *et al.*, 2006) or even species (Hickerson *et al.*, 2006).

A general conclusion that has emerged from simulation studies is that tests based on the number and distribution of haplotypes have more power to detect bottlenecks than statistics based on π, in particular Tajima's D (Ramos-Onsins & Rozas, 2002; Innan *et al.*, 2005; Ramirez-Soriano *et al.*, 2008). Earlier, Felsenstein made a theoretical argument for the inferiority of pairwise measures (Felsenstein, 1992). Their large variance under neutrality arises both from their sensitivity to the last coalescence event and the random genealogical topology (Tajima, 1983). Under the SNM more symmetric genealogies are on average associated with higher π and more ragged mismatch distributions than asymmetric genealogies. It is important to realize that this topological variance is independent of the already large variance in coalescence times inherent in the genealogical process. In other words 'despite their aura of robustness' (Felsenstein, 1992), statistics based on π suffer from an unnecessarily large variance under neutrality, and hence have comparatively low power. Despite these results, D and mismatch distributions continue to be the methods of choice for demographic inferences in population genetics and phylogeography, respectively.

Following Felsenstein's recommendation that 'there is much to gain from explicitly taking the genealogical relationship of a sample into account' (Felsenstein, 1992), the aim of this study is to consider how genealogical information can be used for demographic inference in a summary statistics framework. Our premise here is that the mutation rate is sufficiently high relative to the per site recombination rate such that non-recombining blocks of sequences can be easily identified and treated as independent loci.

Given that there is usually not enough information in within-species sequences data to infer the full topology unambiguously it seems important to ask which part of the topology yields most information. The first part of the paper introduces some simple measures of starshape, which are based on the properties of a rooted genealogy. Using simulations, their power to detect a history of exponential growth is compared with standard neutrality tests for both the single and multi-locus cases. We focus on the exponential growth model for two reasons. Firstly, although it is a frequently used demographic model, the power of summary statistics to detect exponential growth has been little investigated. Secondly, likelihood methods are available, which can be taken as an absolute 'upper bound' of power for comparison. Such a direct comparison between summary statistics and the optimal likelihood methods is lacking so far.

2. Summary statistics

Several neutrality tests compare two different estimators of the scaled mutation rate (Fu & Li, 1993; Tajima, 1989; Fay & Wu, 2000) $\theta = 4N_e\mu$, where μ is the mutation rate and N_e the effective population size, which capture different aspects of the data. Most prominently, Tajima's D is defined as the difference between θ estimated as π, and $\theta_w = S/a_n$ (Watterson's θ, where $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$, n is the sample size and S the total number of polymorphic sites in the sample), normalized by the standard deviation of this difference. Genealogies from growing populations typically have relatively more low frequency variants and hence tend to have a negative D.

While neutrality tests are commonly based on the frequency spectrum and π, it is instructive to consider departures from the SNM in terms of their effect on the genealogy. Such tree-thinking necessarily underlies summaries that make use of outgroup information, e.g. D_2 has a straightforward genealogical interpretation. Below two different ways of employing

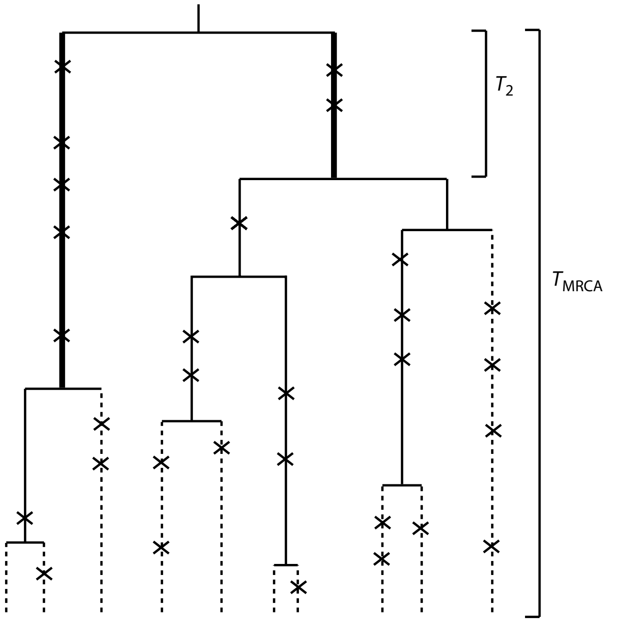

Fig. 1. Random genealogy of a sample of 20 sequences. The root partitions the sample into two subclades of size 3 and 7. Rootward branches are shown as bold, terminal branches as dotted lines, mutations are represented as crosses. The time interval until the last coalescence event, T_2, is shorter than average under the SNM. In this example $S=30$, $\eta_R=7$, $\eta_{R\text{min}}=2$ and $\eta_e=14$.

genealogical information in the construction of summary statistics are considered.

(i) *Genealogical ratios*

The rationale behind D_2 is to distinguish between two classes of mutations: those found on terminal branches, η_e and those on internal branches, η_i (Fig. 1) (Fu & Li, 1993). Suppose that some limited topological information can be inferred from the data. In particular, we will for now assume that the placement of the root is known. It is then possible to distinguish mutations found on the two rootward branches, which we shall denote η_R. Under the infinite sites assumption, these are all derived mutations that are shared by all individuals in either of the two sub-clades defined by the root. The advantage of considering the proximity of mutations to the root rather than the tips is twofold: firstly, rootward branches cover a greater proportion of the time to the most recent common ancestor of the sample (T_{MRCA}) and should, in general, be more informative about past changes in population size. Under the SNM, on average half of the T_{MRCA} is taken up by the coalescence of the last two lineages (T_2) (Fig. 1), whereas in a growing population, the smaller population size in the past forces the last two lineages to coalescence much more rapidly. Secondly, the average length of a branch connected to the root is less dependent on the sample size than the average length of a terminal branch.

Ideally, one wants to know the total number of mutations that have occurred during T_2, rather than the number of mutations on both rootward branches, η_R which is larger and depends on the topology, i.e. the order of the first node on the longer of the two branches (Uyenoyama, 1997, Appendix).

One possibility is to only consider the shorter of the two rootward branches that has exactly length T_2. Thus the number of mutations found on this branch, $\eta_{R\text{min}}$, over θw constitutes a very simple measure of starshape.

$$X=\frac{\eta_{R\text{min}}}{\theta_w}. \tag{1}$$

Such genealogical ratios have first been employed to study the effect of balancing selection on plant incompatibility loci (Uyenoyama, 1997). Being based on a single random event, X clearly neglects much of the information contained in the genealogy. Its power is limited by the probability of observing $\eta_{R\text{min}}=0$ under neutrality. In other words, X is unlikely to be of much use in the case of a single locus.

Alternatively, one can ignore the uncertainty in node order and take the number of mutations found on both rootward branches relative to θ_w:

$$X_1=\frac{\eta_R}{\theta_w}. \tag{2}$$

It is possible of course to construct various composite measures from the number of mutations found on different parts of the genealogy. Here, we only consider one additional statistic, the relative difference between rootward and terminal mutations:

$$X_2=\frac{\eta_{\text{R}}-\eta_{\text{e}}}{\theta_{\text{w}}}. \tag{3}$$

The X statistics assume some knowledge of the tree topology that is usually unknown. Of course one could use some standard method of tree reconstruction and infer η_R and $\eta_{R\text{min}}$ from the most likely topology. However, not only is it inefficient to reconstruct the full topology when all that is required is the placement of the root, conditioning on a single tree also ignores any topological uncertainty. We have therefore developed a simple scheme of inferring the root in a sample of polarized sequences that circumvents these problems.

Under the infinite sites assumption, a necessary criterion for the root-node is that no mutations are shared between the two subsets on either side. One can show that if both branches connected to the root carry mutations, i.e. $\eta_{R\text{min}}>0$ there exists exactly one bipartition of the sample with no mutational overlap. If however one or both of the rootward branches of the genealogy carry no mutations there may be multiple bipartitions that meet this criterion. In this case $\eta_{R\min}=0$ and the tree reconstructed from such a sample would have an unresolved polytomy at its base. To incorporate the topological uncertainty about the placement of the root, we compute the average value of η_R over all partitions that are compatible with the criterion of no mutational overlap. Note that in contrast to most tree reconstruction algorithms that join similar sequences (i.e. start from the tips down the tree), our scheme is divisive (i.e. it starts from the root). To avoid having to consider all possible bipartitions of the sample $(2^{n-1}-1)$, we make use of the fact that any sequences that share mutations have to be on the same side of the root. By first binning sequences that share at least one mutation, we can directly calculate η_R and the number of possible partitions.

(ii) *Starting from the limiting case*

A different approach is to construct summaries that measure departures from the limiting case of a perfectly star-shaped genealogy. Star-shaped genealogies have some convenient properties that can be used for this. Assuming that outgroup information is available, one can record the number of terminal mutations in each sequence *i* (because lineages are exchangeable, the labelling is arbitrary), V_i. In a perfectly star-shaped genealogy, all mutations must fall onto terminal branches by definition. Thus one expects the number of derived mutations in a sequence to be half the average pairwise diversity, i.e. $E[V_i]=\pi/2$. The statistic R_{2E} proposed by Ramos-Onsins and Rozas measures the average departure from this expectation:

$$R_{2E}=\frac{\left(\sum_{i=1}^{n}\left(V_i-\frac{\pi}{2}\right)^2/n\right)^{1/2}}{S} \tag{4}$$

(Ramos-Onsins & Rozas, 2002, eqn (2)). R_{2E} has proven superior to a wide range of summary statistics in detecting histories of bottlenecks (Ramos-Onsins & Rozas, 2002). However, because of its dependence on π, one may suspect it to suffer from a large variance under neutrality. We therefore consider a similar statistic that uses the observed S rather than π to assess the degree of starshape. Consider the total number of derived mutations in each sequence, D_i. Note that $\sum_{i=1}^{n}D_i=\sum_{i=1}^{n-1}i\xi_i$, in terms of the unfolded frequency spectrum, where ξ_i denotes derived mutations that occur *i* times in the sample. Using the fact that $E[D_i]=S/n$ in a star-shaped genealogy we can define a new statistic:

$$R_S=\frac{\left(\sum_{i=1}^{n}\left(D_i-\frac{S}{n}\right)^2/n\right)^{1/2}}{S}. \tag{5}$$

Since under neutrality a large proportion of mutations will be found on inner branches, i.e. be shared by many sequences, $E[D_i]=S/n$. In other words, R_S is such that smaller values are expected under a history of growth.

3. Methods

(i) *Summary statistics and demographic model*

We carried out coalescent simulations in ms (Hudson, 2002) to compare the power of a range of summary statistics to distinguish between the SNM and a history of exponential growth. In addition to D, D_2, R_{2E} and the new statistics defined above, F_S, (Fu, 1996) and H (Fay & Wu, 2000) were considered. F_S is based on the number of haplotypes in the sample and has previously been found to be more powerful than statistics based on the frequency distribution (Fu, 1996; Ramos-Onsins & Rozas, 2002). H was conceived as a test for the effect of selection on linked neutral sites (Fay & Wu, 2000) and is not expected to have power to detect continuous growth. However, other demographic scenarios such as moderate bottlenecks may perturb genealogies in ways similar to genetic hitchhiking resulting in significant values of H. We assume that the population size has grown exponentially with rate α to its current size N_0:

$$N(t)=N_0\,\mathrm{e}^{-\alpha t}. \tag{6}$$

Following standard practice, this exponential growth is incorporated through a re-scaling of time (Slatkin & Hudson, 1991). We define a rescaled time T_coal relative to N_0 and α:

$$T_\mathrm{coal}=\int_0^t\frac{\mathrm{e}^{\alpha t}}{2N_0}\mathrm{d}t=\frac{(\mathrm{e}^{\alpha t}-1)}{2N_0\alpha}. \tag{7}$$

This represents the total amount of genetic drift that has occurred. It is convenient to define a growth rate relative to N_0 as $A=2N_0\alpha$, which gives:

$$T_\mathrm{coal}=\frac{\mathrm{e}^{A^t/2N_0}-1}{A}. \tag{8}$$

(ii) *Power test*

Critical values of 5% confidence for each statistic were determined from 10 000 replicate genealogies simulated under the SNM for each of a wide range of S values (1–250) (Hudson, 1993; Braverman *et al*., 1995; Ramos-Onsins *et al*., 2007). Genealogies from growing populations were simulated conditional on θ. For each replicate the alternative hypothesis of positive growth was tested by comparing the observed value of a statistic to the critical value given the observed S. Power was estimated as the proportion of

10 000 replicate genealogies for which a statistic was below its critical value in a one-tailed test. Power to reject the SNM was recorded for a large range of parameter combinations. We compared the performance of statistics for different growth rates, ($0<A<50$), sample sizes ($n=10$, 50) and values of θ (5–50). When varying θ, we chose a fixed value of $A=8$. This seems compatible with growth rates estimated from empirical data. For example, variation at silent sites in the *Adhr* region and X-linked genes in *Drosophila pseudoobscura* is consistent with $A=7$ (Schaeffer, 2002). While θ can be arbitrarily high for mitochondrial data, $\theta=20$ may be unrealistic for nuclear loci in out-crossing species. Therefore, power was evaluated for a range of θ values (5–50) again keeping the growth rate fixed at $A=8$.

When using means and variances of summary statistics across loci, power was determined analogously to the single locus case. Critical values of 5% confidence of means and variances of statistics were determined from 10 000 replicate sets of loci with the exact same combination of S values. Although computationally expensive, this avoids making any assumptions about the distribution of mutation rates between loci. However, given that mutation rates vary along the genome assuming the same θ for all loci to simulate the alternative history of growth seems unrealistic and may lead to overestimation of power. We checked for the influence of heterogeneity in mutation rates on power by repeating the multilocus power tests with θ values drawn from a gamma distribution with $\alpha=2$ (Pluzhnikov *et al.*, 2002) and a scale parameter equivalent to a mean of $\theta=20$. This combination of growth and mutation rates is roughly comparable to mutation rate estimates for nuclear loci in *Drosophila melanogaster* (Galtier *et al.*, 2000). As before we assumed no recombination within loci as well as absence of linkage between loci, i.e. replicate genealogies were simply treated as multiple loci.

(iii) *Likelihood method*

In practice, both θ and A are unknown, and their likelihood should, in principle, be estimated jointly. However, because of the non-independence of these two parameters, this is not a practical option. Following standard practice we alternated between maximum likelihood estimation of A and θ (Griffiths & Tavaré, 1994). First a maximum likelihood estimate (MLE) for θ under the SNM was estimated using the program GENETREE (http://www.stats.ox.ac.uk/griff/software.htm). In a second step, this MLE for θ was fixed to run a likelihood surface for A. Finally, the MLE value for A was used to re-evaluate θ. This scheme yields two MLEs for θ for each replicate, one under the assumption of no growth and one given the most likely growth rate, which were compared in a likelihood ratio test (LRT). We did not find that the MLE estimates for A and θ improved upon repeated re-evaluation suggesting that a single round of estimation is sufficient for this moderate growth scenario. 100,000 runs were performed for each likelihood surface evaluation. Again, the proportion of replicate genealogies for which the null hypothesis could be rejected was taken as a measure of statistical power. Due to the long computing time, 100 replicates per parameter combination were used.

4. Results

(i) *Single locus*

In general, both the likelihood method and summary statistics have low power to detect a history of moderate ($A<8$) exponential growth for $n=10$ (Fig. 2). As expected, the likelihood method is most powerful overall, although its superiority is surprisingly small. For example, based on the LRT the SNM is rejected for 30% of genealogies simulated under exponential growth of $A=4$. In comparison, R_S and R_{2E} detect this history of growth in 23% of cases (Fig. 2).

Consistent with previous results, F_S, R_{2E}, and the new measure R_S, are considerably more powerful than both D and D_2 (Ramos-Onsins & Rozas, 2002; Ramirez-Soriano *et al.*, 2008). For $\theta=20$, F_S is the most powerful statistic. The new measure R_S has consistently higher power than R_{2E}. As expected, H and X have no power to distinguish between the SNM and the growth case (not shown). However, the other two genealogical ratios perform surprisingly well. X_1 has higher power than D_2 and the power of X_2 is between that of R_{2E} and R_S (Fig. 2). The complete lack of power of D for $n=10$ is somewhat surprising. Comparison with the result for $n=50$ (Fig. 3) reveals that its performance is strongly dependent on sample size. We ran additional simulations (not shown) and found that for $n<15$ extremely negative values of D are more likely under neutrality than under growth resulting in a rejection rate of the SNM of less than 5%. In other words, when n is small, the variance of D under neutrality is too large to detect exponential growth.

In general, all statistics have considerably higher power for $n=50$ (Fig. 3). Interestingly, it never reaches 100% even when growth is extreme ($A=50$). However, the relative effect of the sample size on power differs between statistics. For instance, X_1 improves relatively little in comparison to other measures. This is to be expected given that even small samples are likely to include the deepest split in the genealogy of the whole population (Saunders *et al.*, 1984). For $n=10$, the power of all statistics decreases

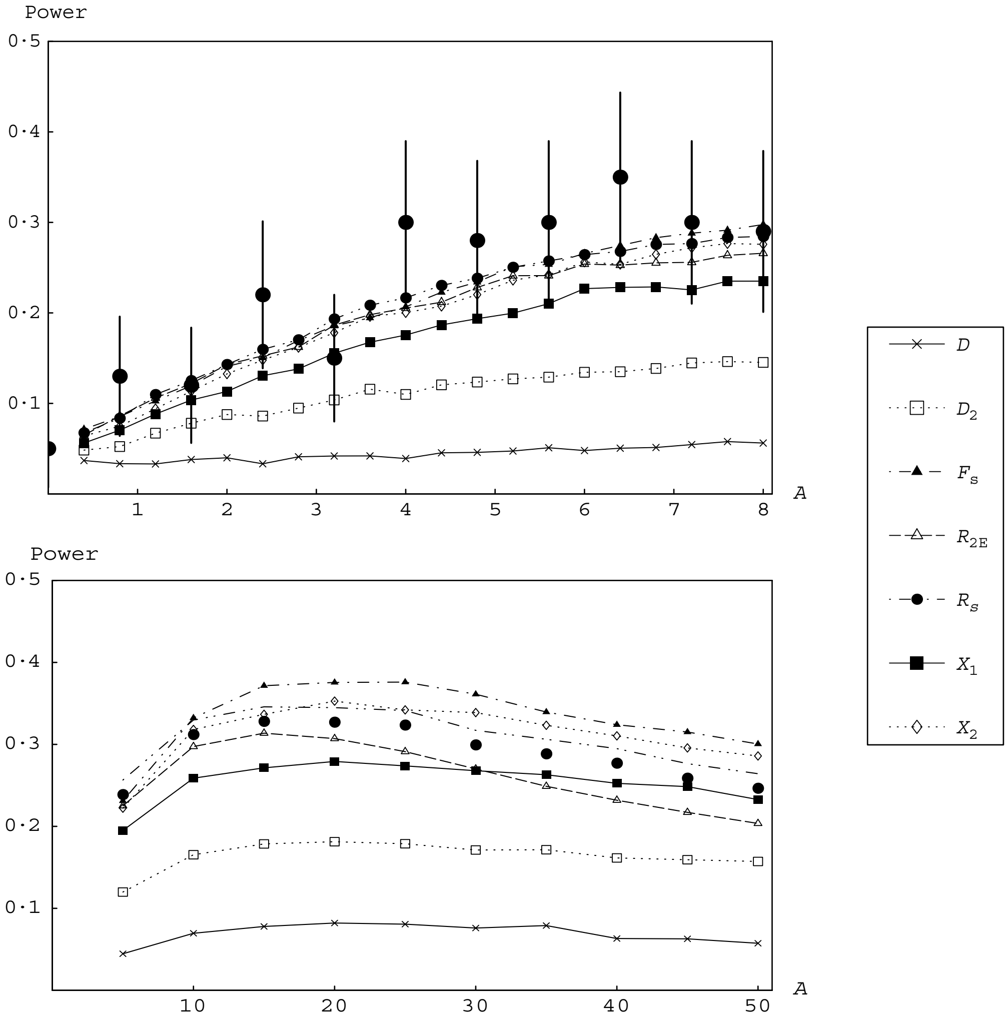

Fig. 2. Power of summary statistics and likelihood method against exponential growth rate $A=0$–50. $n=10$, $\theta=20$. Each point is based on 10 000 replicate simulations. The power of the likelihood method was estimated from 100 replicates (see large filled circles and error bars).

for histories of extreme growth ($A>25$) (Fig. 2). This is due to the overall shortening of genealogies under rapid growth.

The mutation rate has a relatively small influence on power. In general, the power of all measures increases with θ (Fig. 4). However, the trajectories X_1 and F_S level off while the power of the other statistics continues to improve with increasing values of θ. The power of F_S is limited by the number of haplotypes (which cannot exceed n).

To check how statistics are affected by the topological variance, genealogies simulated under the alternative history of growth were sorted according to the bipartition by the root and the proportion of significant values determined for each topology class. Figure 5 clearly shows that the two statistics based on π, D and R_{2E} as well as D_2 are sensitive to asymmetric topologies. The chance of observing a significant value increases markedly with topological asymmetry. This effect is most pronounced for D, which has no 'power' to reject the SNM unless genealogies are very asymmetric and growth is weak. In contrast, the dependency of X_1 on the rootward partition is relatively slight and in the opposite direction, i.e. the chance of rejecting the SNM is smaller for asymmetric genealogies (Fig. 5).

(ii) *Multiple loci*

Compared with the relatively subtle effect both θ and n have on statistical power, increasing the number of loci improves power dramatically. In the mean-based test, all statistics apart from D have a power of close to 100% to detect a history of moderate exponential growth ($A=8$) for 10 loci. However, the relative performance of statistics changes slightly compared with the single locus case. Notably, X_2 has higher power than all other summary statistics (Fig. 6). The power of X is slightly lower than that of X_1 (not shown). Analogously to the results for a single locus, power

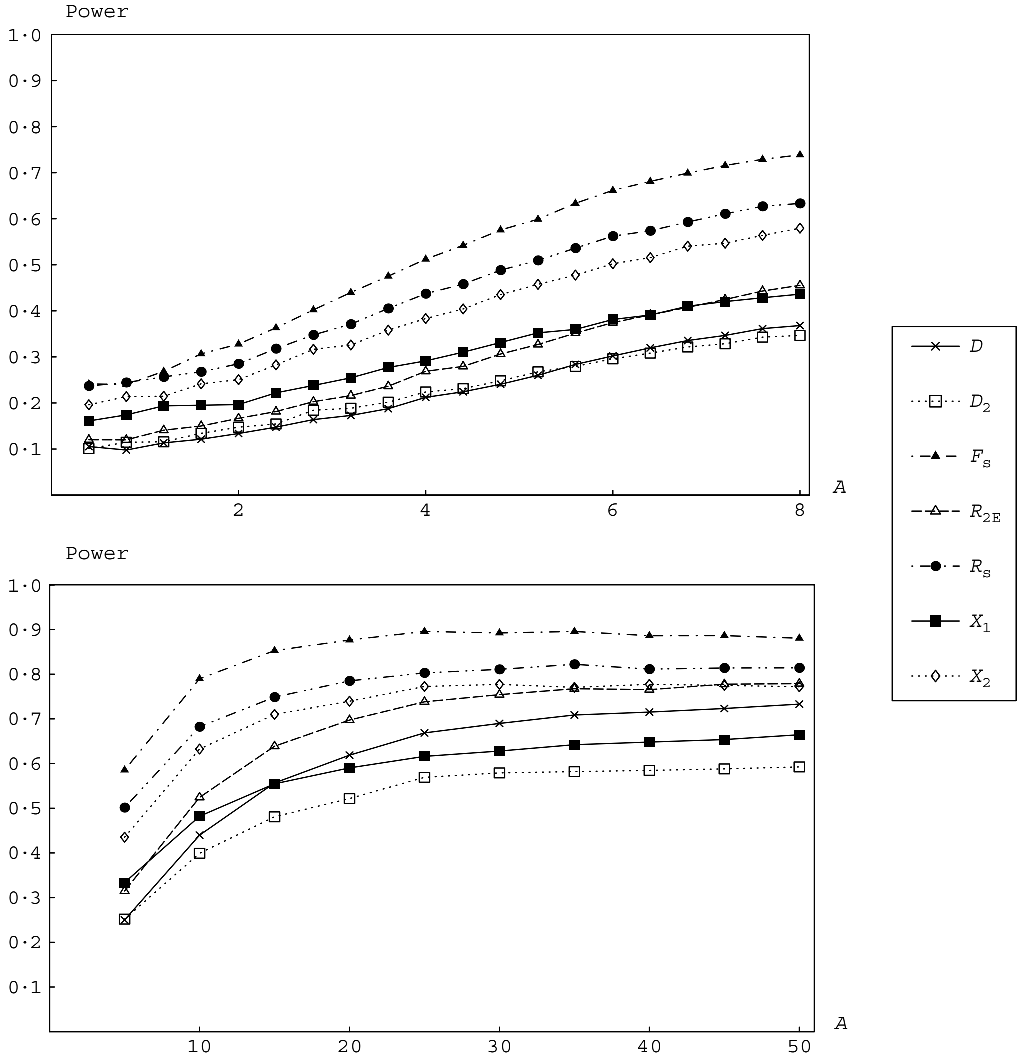

Fig. 3. Power of summary statistics against exponential growth rate $A=0$–50. $n=50$, $\theta=20$. Note the different range (0–1) on the y-axis compared with Fig. 2.

increases both with more extreme growth scenarios and larger n (not shown).

As one may suspect, the increase in power with the number of loci is slower for the variance test. More importantly, the relative performance of statistics is very different. By far the most powerful statistic in the variance test is X_1 followed by D and X (Fig. 7). This indicates a general trade-off. Statistics with a high variance under the SNM have comparatively low power in the single-locus case and the mean test, but high power in the variance test and vice versa.

Allowing for heterogeneity in mutation rates between loci affects both the relative performance of summary statistics and their overall power. As one may expect, heterogeneity in θ generally results in a decrease in power. In the mean-based test, the three X statistics are most affected. However, in the variance test the performance of X_1 is little affected. This statistic even has slightly higher power when mutation rates vary between loci. This appears to be due to the non-normal distribution of X_1 under growth. Genealogies with more than one possible root-partition generally have a very low value of X_1, since we take an average over all possible partitions most of which will be associated with $X_1=0$.

5. Discussion

It is important to distinguish between the general limitations that genealogical and mutational stochasticity impose on demographic inference from genetic data and problems associated with particular methods. Two main conclusions emerge from comparing the performance of the new 'genealogical statistics' to classical neutrality tests and the LRT.

(i) *General limits to demographic inference*

The signatures that changes in population size leave in genealogies are typically subtle compared with the

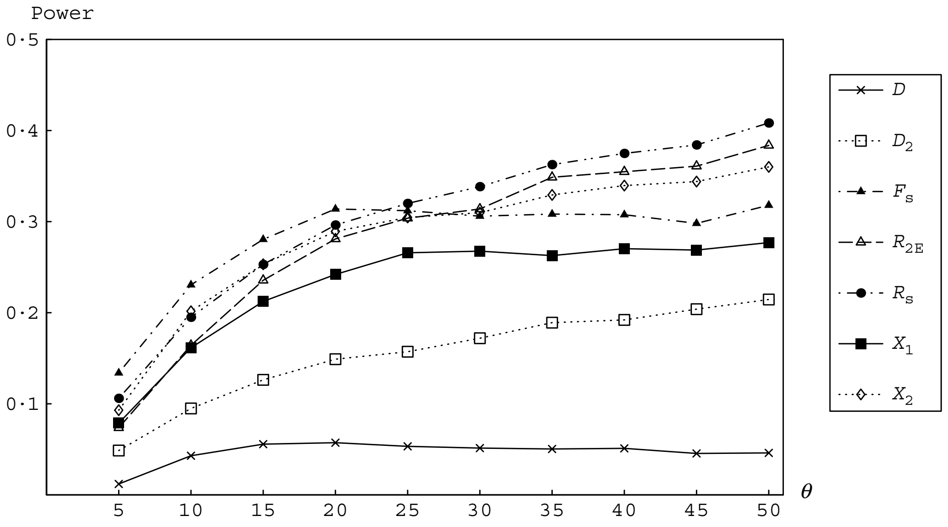

Fig. 4. Power of summary statistics to detect a history of exponential growth ($A=8$) against θ. $n=10$.

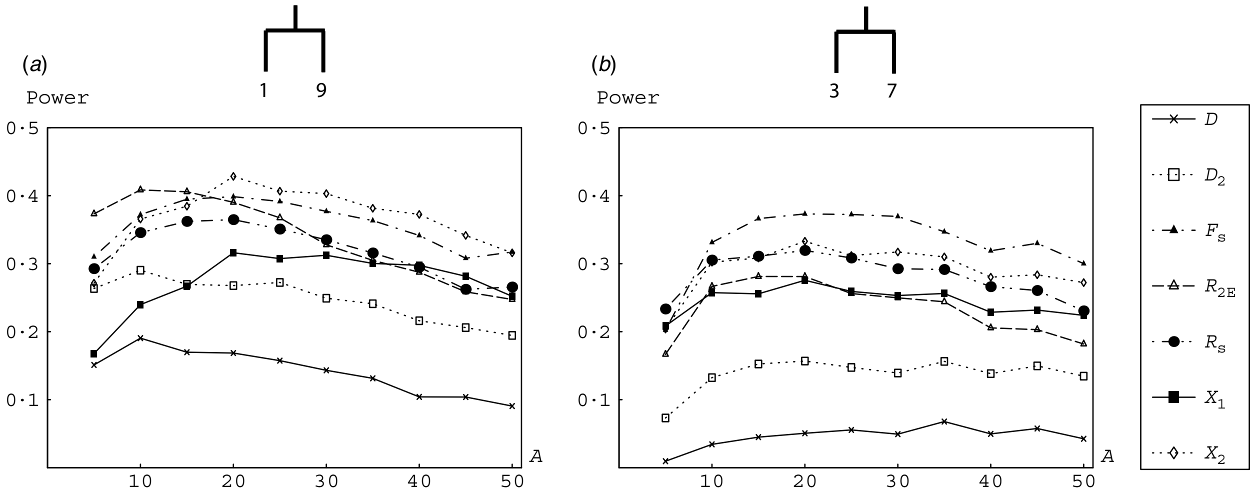

Fig. 5. The effect of topological asymmetry on statistical power (simulation parameters as in 2). Genealogies of Fig. 2 were sorted according to the partition by the root (shown above plot). Only the most asymmetrical partition (9, 1) (a) and one other case (7, 3) (b) are shown. Results for the other three partitions were very similar to (b). Note that since lineages are exchangeable all asymmetrical partitions have the same probability $P_a=2/(n-1)$ (Tajima, 1983, eqn (2)).

randomness of the ancestral process. Thus all methods have low power to distinguish between the SNM and histories of moderate growth in the single locus case. A surprising finding of this study was that the full likelihood method only works marginally better than the most powerful summary statistics. Changes in N_e disproportionally affect the length of the basal branches of a genealogy. However, because these rootward branches also contribute most to the variance in total tree length, inferences based on a single locus will be weak at best. It is telling that the *X* statistics which only considers the last coalescence events in the history, outperform standard neutrality tests in the variance test when multiple realizations of this event, i.e. loci, are available. As has been argued before, most statistical power can be gained by increasing the number loci, which represent independent realizations of the ancestral process, rather than the sample size or the length of sequence (Felsenstein, 1992; Kliman *et al.*, 2000; Wakeley, 2004).

(ii) *Pairwise measures*

Independent of the general limits to demographic inference, pairwise measures such as *D* have particularly low power to infer demography. This has been found in previous simulation studies, which consider other demographic scenarios such as strong bottlenecks and rapid logistic growth (Fu, 1996; Ramos-Onsins & Rozas, 2002; Ramirez-Soriano *et al.*, 2008). The fundamental flaw of pairwise measures can be best understood in terms of the underlying genealogy.

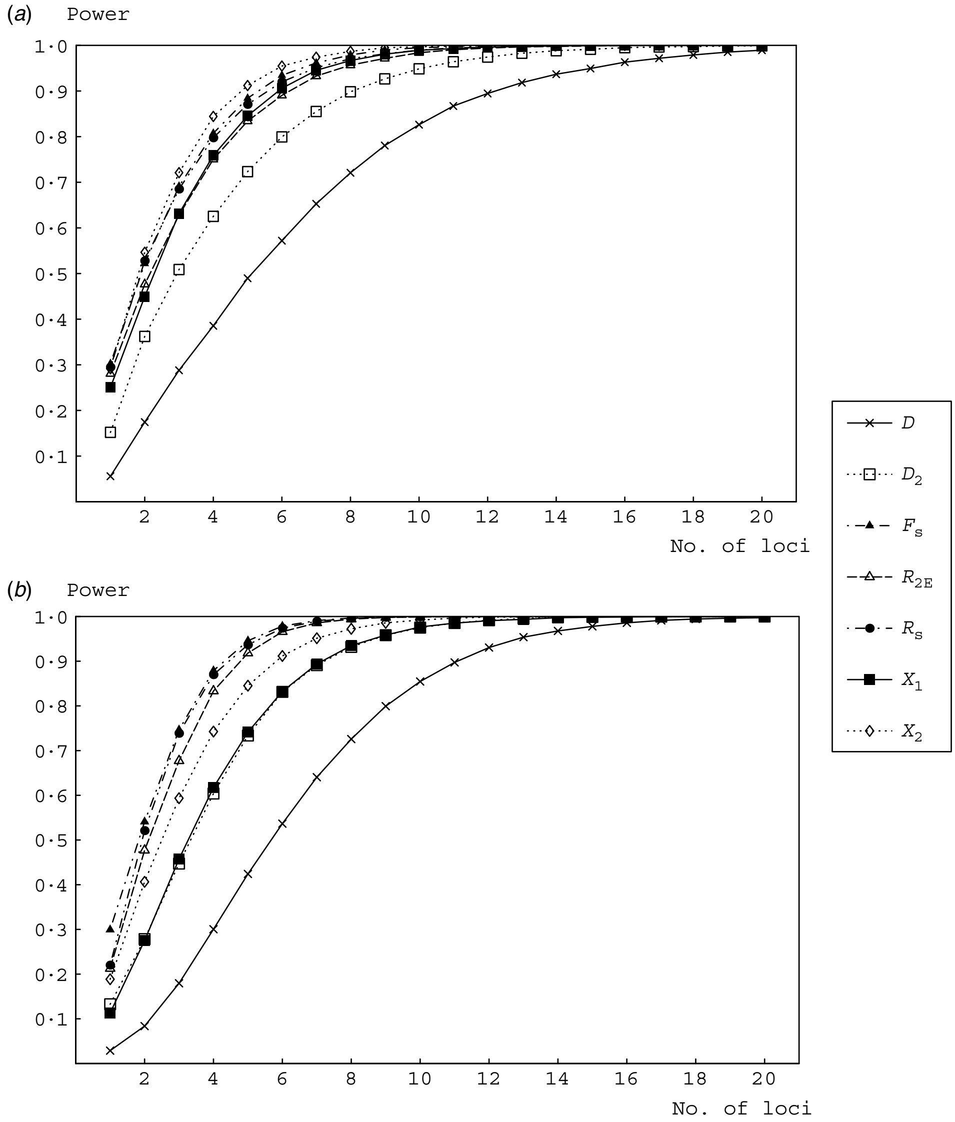

Fig. 6. Power of summary statistics to detect a history of growth $A=8$ using the mean across multiple loci against the number of loci, $n=10$ (A) and $\theta=20$ (B). Assuming mutational rate heterogeneity (θ gamma distributed with $\alpha=2$ and $E[\theta]=20$).

In contrast to selection and population structure, changes in N_e on their own only alter the distribution of branch lengths without affecting the topology, which can be regarded as a random nuisance parameter. While the full topology can rarely be reconstructed, there is potentially a lot of topological information in sequence data. Thus, the challenge that any efficient inference method has to meet is to separate this topological information from the relevant branch length information while taking topological uncertainty into account. Tree-based methods such as lineage-through time plots clearly fall short of the latter because they rely on a fully resolved topology. Pairwise measures on the other hand simply ignore the confounding effect of the topology (Felsenstein, 1992). It is thus easy to see why D has power only when sample sizes are large. While increasing sample size adds increasingly shorter external branches and therefore little additional information, it does reduce the chance of extremely asymmetric bipartitions by the root which are responsible for much of the variance in π and hence D.

Perhaps worryingly, this sensitivity to the topology not only translates into a loss of statistical power but also means that negative D values may in fact be more informative about the topological asymmetry of the genealogy (which may be caused by other non-neutral forces, e.g. selection) underlying the sample than about past growth. In order to distinguish between the effects of selection and demography, topology needs to be separated from branch length information. One approach is to explicitly account for the topology information if possible. For instance, one could determine confidence intervals of statistics conditional

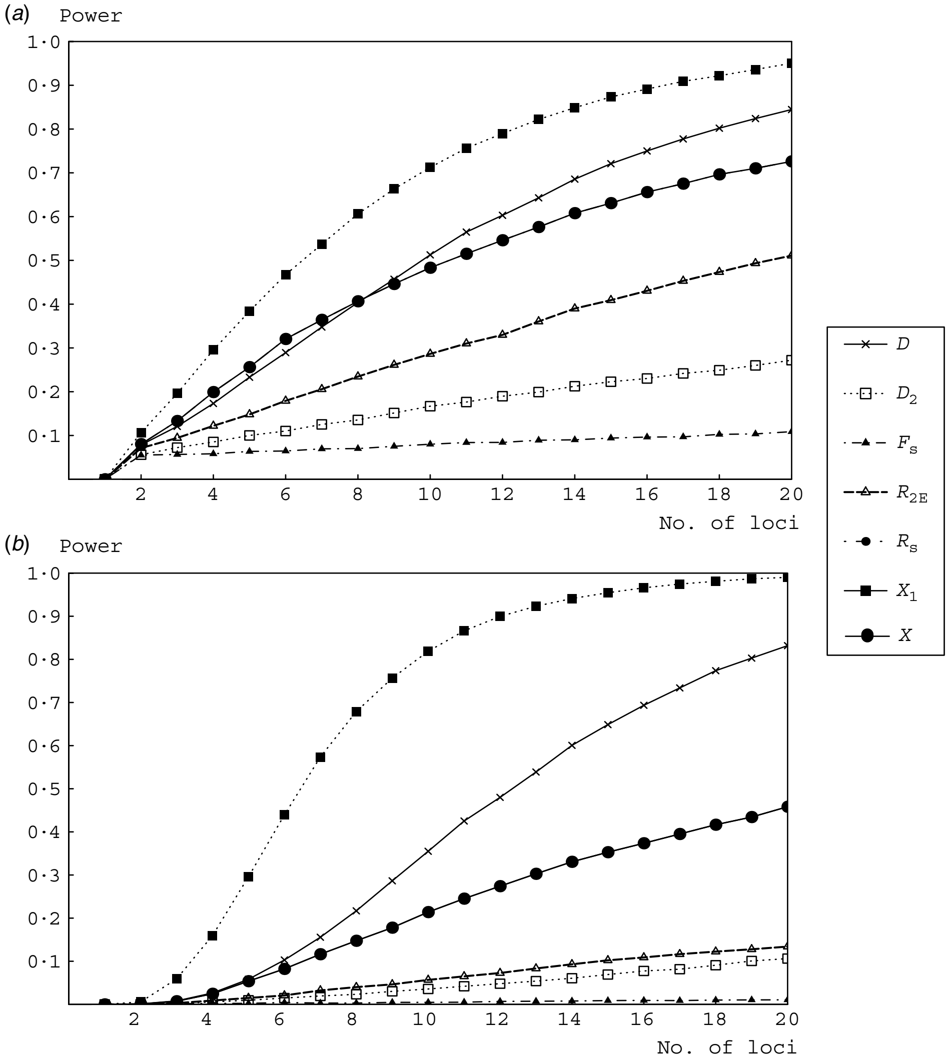

Fig. 7. Power of summary statistics in the variance-based tests across multiple loci for three different growth rates (from left to right $A = 2$, 4, 8). (A) $\theta = 20$. (B) Assuming mutational rate heterogeneity (θ gamma distributed with $\alpha = 2$ and $E[\theta] = 20$).

on the bipartition by the root if this is known. Not surprisingly, this improves the power of D, but has little effect on statistics that are not based on π (not shown). The alternative is to use measures that are less sensitive to the topology. F_S and other haplotype statistics have previously been shown to be more powerful than frequency spectrum statistics for this very reason (Depaulis *et al.*, 2003; Innan *et al.*, 2005). However, it has also been noted that F_S sometimes behaves erratically (Fu & Li, 1993; Ramos-Onsins & Rozas, 2002). As mentioned earlier, its power levels off with increasing θ (Fig. 4), because the sample size sets an upper bound to the number of haplotypes.

(iii) *Recombination and topological uncertainty*

The X statistics presented here fall somewhere in between tree-based methods and classical summary statistics. They exploit the fact that changes in population size disproportionally affect the relative length of the deepest branches in the genealogy and make use of topological information, without sacrificing the simplicity of the summary statistics framework. Given their high power in the multilocus case, how useful are such genealogical ratios in practice?

Recombination presents a fundamental problem to tree-based methods like the X statistics, which are defined only for non-recombining sequences. Similarly, likelihood methods that can deal with recombination are currently not available. To wrongly reconstruct trees from recombining data can potentially be severely misleading especially in the context of demographic inference. In fact, genealogical ratios similar to the ones presented here have been used to show that recombination can mimic the effect population growth has on the shape of inferred genealogies. Internal branches will appear relatively shorter and the tree overall more star-shaped (Schierup & Hein, 2000; Ramirez-Soriano *et al.*, 2008). Ideally one would like to model recombination explicitly when making demographic inferences. However estimates of recombination rates are usually associated with a large uncertainty. Furthermore, it is notoriously difficult to distinguish between recombination and back-mutations.

One approach to circumvent these problems is to test for recombination beforehand (e.g. using the four gamete test) and exclude recombinant regions from the analysis if necessary. One can then both condition on there being no within-locus recombination and afford to use more powerful statistics such as the ones presented here. This strategy of identifying non-recombining stretches of sequence is increasingly used to analyse multilocus data, e.g. Galtier *et al.* (2000) or Jennings & Edwards (2005). Fortunately, many organisms appear to have lower recombination rates than model species such as Drosophila. For instance in a recent study on Australian birds only 6 out of 30 loci of intergenic sequence showed evidence for recombination (Jennings & Edwards, 2005). How profitable this scheme is ultimately depends on the relative magnitude and distribution of recombination and mutation rates. Before the genealogical ratios can be used on multiple loci, which have been pruned to exclude recombinant stretches, both the potential bias of such pruning and the effect of undetected recombination events on the genealogical ratios need to be properly evaluated. Interestingly, our method of inferring the root does in itself constitute a test for recombination and may help to focus on those recombination events that matter to the statistical test.

A related problem concerns the infinite sites assumption. Although the algorithm we have developed to compute the X statistics takes topological uncertainty into account, ignoring the possibility of back-mutations may underestimate the length of basal branches (Baudry & Depaulis, 2003). Although this source of error has been ignored here it should in principle be possible to account for back-mutations considering that they are independent of the assumptions of the genealogical process. In fact, any mutational model can be used to define statistics analogous to the genealogical ratios presented here. The problem with more complicated mutation models is in estimating the basal topology needed to calculate these measures.

(iv) *Conclusions*

In summary, the results confirm that only the most extreme demographic events leave a sufficient signature to be detectable in single locus data. Still, instead of the excessive and often non-quantitative employment of mismatch distributions, phylogeographic studies could benefit from using more powerful statistics such as R_S and R_{2E} to test demographic hypotheses. Conversely, population genetics studies of sequence data from multiple, unlinked loci could benefit from using summary statistics that incorporate genealogical information explicitly. When outgroup information is available and the assumptions of no within-locus recombination and infinite sites mutations can be justified, simple genealogical ratios are potentially more powerful than standard statistics. In taking the relative number of mutations found on specific parts of the genealogy as a measure of the degree of starshape, the demographic signal can be separated from irrelevant and confounding topological information. Extensions of this approach are feasible. For instance, one could consider the covariance between the number of basal and terminal mutations. Such simple statistics may be profitable for approximate likelihood or Bayesian approaches (Thornton & Andolfatto, 2006). There remains a need to understand the effect of pruning and undetected recombination events on tree reconstruction in general and tree-based measures such as the X statistics presented here in particular.

Many thanks to N. Barton, P. Haddrill, J. Polechova, D. Charlesworth, Kai Zeng and D. Obbard for helpful advice and discussion. Thanks also to Kelly Dyer for help with Genetree. Detailed comments and valuable suggestions from two anonymous reviewers on an earlier version of this manuscript greatly improved this work. K.L. is funded by a studentship from the Biotechnology and Biological Sciences Research Council. J.K. is funded by EPSRC.

References

Baudry, E. & Depaulis, F. (2003). Effect of misoriented sites on neutrality tests with outgroup. *Genetics* **165**, 1619–1622.

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. & Stephan, W. (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**, 783–796.

Depaulis, F., Mousset, S. & Veuille, M. (2003). Power of neutrality tests to detect bottlenecks and hitchhiking. *Journal of Molecular Evolution* **57**, S190–S200.

Di Rienzo, A., Donnelly, P., Toomajian, C., Sisk, B., Hill, A., Petzl-Erler, M. L., Haines, G. K. & Barch, D. H. (1998). Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. *Genetics* **148**, 1269–1284.

Emerson, B. C., Paradis, E. & Thebaud, C. (2001). Revealing the demographic histories of species using DNA sequences. *Trends in Ecology and Evolution* **16**, 707–716.

Fay, J. C. & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413.

Felsenstein, J. (1992). Estimating effective popuation size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetical Research* **59**, 139–147.

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

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

Galtier, N., Depaulis, F. & Barton, N. H. (2000). Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics* **155**, 981–987.

Glinka, S., Ometto, L., Mousset, S., Stephan, W. & De Lorenzo, D. (2003). Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics* **165**, 1269–1278.

Griffiths, R. C. & Tavaré, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions: Biological Sciences* **344**, 403–410.

Haddrill, P. R., Thornton, K. R., Charlesworth, B. & Andolfatto, P. (2005). Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Research* **15**, 790–799.

Hamblin, M. T., Mitchell, S. E., White, G. M., Gallego, J., Kukatla, R., Wing, R. A., Paterson, A. H. & Kresovich, S. (2004). Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. *Genetics* **167**, 471–483.

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

Heuertz, M., De Paoli, E., Kallman, T., Larsson, H., Jurman, I., Morgante, M., Lascoux, M. & Gyllenstrand, N. (2006). Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of norway spruce [*Picea abies* (L.) Karst]. *Genetics* **174**, 2095–2105.

Hickerson, M. J., Dolman, G. & Moritz, C. (2006). Comparative phylogeographic summary statistics for testing simultaneous vicariance. *Molecular Ecology* **15**, 209–223.

Hudson, R. R. (1993). The how and why of generating gene genealogies. In *Mechanisms of Molecular Evolution* (Eds. N. Takahata & A. G. Clark), pp. 23–36. Sinauer, Sunderland, Mass.

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

Innan, H., Zhang, K., Marjoram, P., Tavaré, S. & Rosenberg, N. A. (2005). Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. *Genetics* **169**, 1763–1777.

Jennings, W. B. & Edwards, S. V. (2005). Speciation history of Australian grass finches (*Poephila*) inferred from thirty gene trees. *Evolution* **59**, 2033–2047.

Kliman, R. M., Andolfatto, P., Coyne, J. A., Depaulis, F., Kreitman, M., Berry, A. J., McCarter, J., Wakeley, J. & Hey, J. (2000). The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* **156**, 1913–1931.

Kuhner, M. K., Yamato, J. & Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using metropolis-hastings sampling. *Genetics* **140**, 1421–1430.

Nee, S., Holmes, E. C., Rambaut, A. & Harvey, P. H. (1995). Inferring population history from molecular phylogenies. *Philosophical Transactions of the Royal Society of London Series B* **349**, 25–31.

Ometto, L., Glinka, S., De Lorenzo, D. & Stephan, W. (2005). Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Molecular Biology and Evolution* **22**, 2119–2130.

Pluzhnikov, A., Di Rienzo, A. & Hudson, R. R. (2002). Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics* **161**, 1209–1218.

Pybus, O. G., Rambaut, A., Holmes, E. C. & Harvey, P. H. (2002). New inferences from tree shape: numbers of missing taxa and population growth rates. *Systematic Biology* **51**, 881–888.

Ramirez-Soriano, A., Ramos-Onsins, S. E., Rozas, J., Calafell, F. & Navarro, A. (2008). Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics* **179**, 555–567.

Ramos-Onsins, S. E., Mousset, T., Mitchell-Olds, T. & Stephan, W. (2007). Population genetic inference using a fixed number of segregating sites: a reassessment. *Genetical Research* **89**, 231–244.

Ramos-Onsins, S. E. & Rozas, J. (2002). Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution* **19**, 2092–2100.

Reich, D., Feldman, M. & Goldstein, D. (1999). Statistical properties of two tests that use multilocus data sets to detect population expansions. *Molecular Biology and Evolution* **16**, 453–466.

Saunders, I. W., Tavaré, S. & Watterson, G. A. (1984). On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability* **16**, 471–491.

Schaeffer, S. W. (2002). Molecular population genetics of sequence length diversity in the adh region of *Drosophila melanogaster*. *Genetical Research* **80**, 163–175.

Schierup, M. H. & Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879–891.

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

Simonsen, K. L., Churchill, G. A. & Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429.

Slatkin, M. & Hudson, R. R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555–562.

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

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

Thornton, K. & Andolfatto, P. (2006). Approximate bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics* **172**, 1607–1619.

Uyenoyama, M. K. (1997). Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. *Genetics* **147**, 1389–1400.

Wakeley, J. (2004). Recent trends in population genetics: more data! more math! simple models? *Journal of Heredity* **95**, 397–405.

doi:10.1111/j.1558-5646.2010.01024.x

QUANTIFYING THE PLEISTOCENE HISTORY OF THE OAK GALL PARASITOID *CECIDOSTIBA FUNGOSA* USING TWENTY INTRON LOCI

Konrad Lohse,[1,2] Barbara Sharanowski,[3] and Graham N. Stone[1]

[1]Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, West Mains Road, EH9 3JT, United Kingdom

[2]E-mail: konrad.lohse@gmail.com

[3]Department of Entomology, North Carolina State University, 2317 Gardner Hall, Campus Box 7613, Raleigh, North Carolina 27695

Received September 17, 2009
Accepted April 2, 2010

The longitudinal spread of temperate organisms into refugial populations in Southern Europe is generally assumed to predate the last interglacial. However, few studies have attempted to quantify this process in nonmodel organisms using explicit models and multilocus data. We used sequence data for 20 intron-spanning loci (12 kb per individual) to resolve the history of refugial populations of a widespread western Palaearctic oak gall parasitoid *Cecidostiba fungosa* (Pteromalidae). Using maximum likelihood and Bayesian methods we assess alternative population tree topologies and estimate divergence times and ancestral population sizes under a model of divergence between three refugia (Middle East, Balkans and Iberia). Both methods support an "Out of the East" history for *C. fungosa*, matching the pattern previously inferred for their gallwasp hosts. However, coalescent-based estimates of the ages of population divides are much more recent (coinciding with the Eemian interglacial) than nodal ages of single gene trees for *C. fungosa* and other species. We also find that increasing the sample size from one haploid sequence per refugial population to three only marginally improves parameter estimates. Our results suggest that there is significant information in the minimal samples currently analyzable with maximum likelihood methods, and that similar methods could be applied to multiple species to test alternative models of assemblage evolution.

KEY WORDS: Ancestral population size, coalescent theory, parasitoid assemblages, population divergence times, statistical phylogeography.

Many western palaearctic taxa have their current centers of genetic diversity to the east of Europe, suggesting that refugial populations around the Mediterranean basin are ultimately derived from a more eastern source (Din et al. 1996; Rokas et al. 2003; Juste et al. 2004; Michaux et al. 2004; Culling et al. 2006; Koch et al. 2006; Challis et al. 2007; Stone et al. 2007). Westwards dispersal of such taxa into southern European refugia is often thought to have occurred in the early Pleistocene, if not before (Taberlet et al. 1998; Rokas et al. 2003; Juste et al. 2004; Culling et al. 2006; Challis et al. 2007) and of necessity must predate the well-documented latitudinal range shifts associated with the last ice age (Taberlet et al. 1998; Hewitt 1999) by at least one glacial cycle. However, the few studies that have attempted to estimate the age of this older longitudinal dispersal are largely qualitative, being based on a small set of (primarily mitochondrial) gene trees (e.g., Taberlet et al. 1998; Hewitt 1999; Nichols 2001; Rokas et al. 2003; Juste et al. 2004; Culling et al. 2006; Challis et al. 2007). It has been noted that species differ considerably in their mitochondrial divergence between refugia and this has been attributed to species-specific responses to Pleistocene climate cycles (Taberlet et al. 1998). However, an obvious alternative explanation for the observed lack of interspecific temporal congruence is that mitochondrial gene trees are dominated by incomplete lineage sorting, the extent of which may be large in general and/or different between species (Nichols 2001).

Because polymorphism within ancestral populations must originate before daughter populations diverge, branches of gene trees are necessarily longer than those of population trees and a naïve interpretation of node ages may severely overestimate population divergence (Pamilo and Nei 1988; Maddison 1997). Similarly, gene tree topologies may be incongruent with the order of population divergence (Tajima 1983; Pamilo and Nei 1988; Rosenberg 2002). Because the magnitude of both these effects depends on the size and stability of the ancestral populations (Tajima 1983; Maddison 1997; Nichols 2001), they are likely to be exaggerated when resolving the origins of—and relationships among—refugial populations, which are stable by their very nature (Hewitt 1999). Thus, assessing the generality of an "Out of the East" pattern ideally requires replication both at the level of species and loci.

Assemblages of parasitoids associated with oak cynipid galls offer unmatched replication at the species level. In the Western Palaearctic, an estimated 120 species of chalcidoid wasps are obligate natural enemies of the inhabitants of oak cynipid galls (Csóka et al. 2005; Hayward and Stone 2005). Phylogeographic studies on Western Palaearctic oak gallwasps show their populations to be divided into three major refugial areas: the Iberian Peninsula in the west, Central Europe and the Balkans in the center, and Asia Minor and Iran in the east (Rokas et al. 2001, 2003; Stone et al. 2001, 2008; Challis et al. 2007), broadly paralleling patterns seen in oak phylogeography (Dumolin-Lapegue et al. 1997). In the gallwasps, allele frequency data for multiple nuclear markers support the conclusion that there has been very little subsequent gene flow between these regions (Rokas et al. 2001, 2003; Stone et al. 2001, 2008; Challis et al. 2007). Oak gallwasps are thought to have diversified in regions to the east of Europe prior to the Pleistocene (Stone et al. 2009), and pre-Pleistocene or early Pleistocene westwards range expansion across Europe has been suggested by patterns of genetic variation in several widespread species (Rokas et al. 2001, 2003; Challis et al. 2007). An obvious question is whether gall-associated parasitoids have pursued their hosts from the east. At least two of them, the torymids *Megastigmus stigmatizans* and *M. dorsalis,* appear to have done so (Rokas et al. 2003; Hayward and Stone 2006; Nicholls et al. 2010). The challenge now is to reconstruct longitudinal colonization processes in the Western Palaearctic for a broader taxonomic spread of oak gall-associated parasitoids, to assess the generality of an "Out of the East" pattern, and to determine whether parasitoids dispersed over a similar timescale to their hosts, or after a delay—so allowing their hosts a measure of "enemy-free space" (Hayward and Stone 2006). One reason for caring which of these scenarios is true is that close phylogeographic concordance increases the potential for coevolution among community members, and such communities are inherently sensitive to disturbance by species gain (Stone and Sunnucks 1993; Schönrogge et al. 1996b, 1998) or loss (Lennartsson 2002; Pauw 2007).

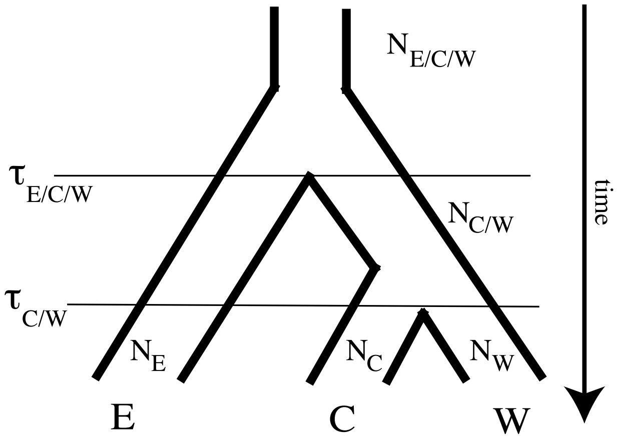

Figure 1. Model of successive population divergence between major Palearctic refugia from East to West: Asia Minor and Iran (E) Balkans and Central Europe (C), Iberia (W). With minimal sampling of one individual per population, topological probabilities of gene trees are determined by only two model parameters, the time between population divergences ($\tau_{E/C/W} - \tau_{C/W}$) and the effective sizes of the ancestral population during this time ($N_{C/W}$).

Here, we use sequence data from 20 intronic loci to study the history of refugial populations in the pteromalid parasitoid *Cecidostiba fungosa*, a widespread species in oak gall communities (Askew 1961; Schönrogge et al. 1996a; Bailey et al. 2009). The three-refuge phylogeographic pattern of oak gallwasp communities allows us to compare two analytical methods—a maximum likelihood (ML) approach (Yang 2002), and an analogous, Bayesian approach (Rannala and Yang 2003). Both estimate ancestral population parameters (population sizes and divergence times) directly from patterns of polymorphism in sequence data (rather than from gene trees inferred for each locus) and assume a model of divergence between three populations (Fig. 1). The order of population divergence or the topology of the population tree can be viewed as an additional model parameter and the likelihoods in both methods can be used to compare statistical support for different topologies. We address the following, specific questions: (1) Do data for *C. fungosa* support an "Out of the East" population history, such that refugial populations in the center and west of Europe are derived from a shared ancestral population in the center which in turn is derived from a common ancestral population further east (Fig. 1)? (2) When did refugial populations split from each other, and how large were their ancestral populations? (3) How different are multilocus estimates of population divergence times from gene divergence times (both nuclear and mitochondrial)?

A strategy of sampling many loci from a single individual per taxon has been used extensively to study divergence between closely related species, in particular the Great Apes (Yang 2002; Jennings and Edwards 2005; Patterson et al. 2006). There are two reasons why such minimal sampling is of interest. First, going backwards in time, only lineages that persist into the ancestral species/population contribute to estimates of ancestral population parameters. Coalescent theory shows that samples taken from the same species or population quickly coalesce down to a small number of lineages (Griffiths 1981; Tavaré 1984; Norborg 1998) (Fig. 2). This means that even if divergence is relatively recent, that is, less than N_e generations ago, the power gained by increasing within-population sampling levels off relatively rapidly. In contrast, each additional sampled locus provides an independent replicate of the coalescent process in the ancestral population irrespective of the divergence time (Wakeley 2004). So if the total cost of sampling is number of loci × number of sampled individuals, the optimal sampling scheme is one of few individuals sequenced at a large number of loci. Second, minimal sampling is currently the only sampling scheme for which a statistically optimal likelihood method allowing parameter estimation directly from site patterns exists (Yang 2002). In contrast, Bayesian approaches (Rannala and Yang 2003) or gene tree–species tree methods (Degnan and Salter 1995; Maddison and Knowles 2006; Liu and Pearl 2007; Degnan and Rosenberg 2009; Kubatko et al. 2009) have the advantage that they can deal with arbitrary sample sizes and numbers of populations. However, this comes at the potential cost of prior assumptions and/or difficulty in integration over topological uncertainty in the gene trees.

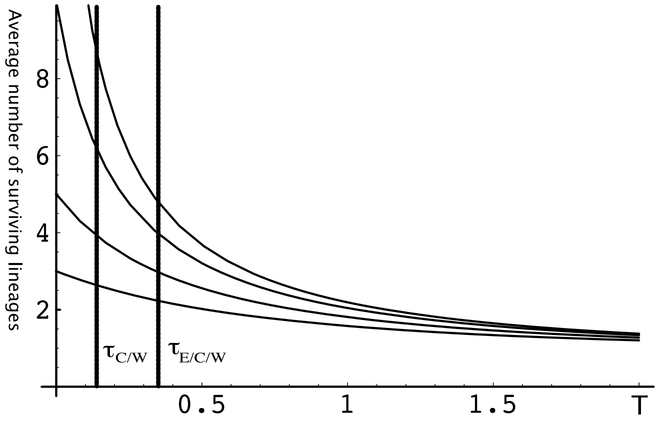

Figure 2. The expected mean number of lineages surviving coalescence into an ancestral population (Tavaré 1984, equation 5.5) plotted against divergence time (T) in coalescence units (2N_e generations) for four different sampling sizes (from top to bottom, $n = 20, 10, 5, 3$). Because only surviving lineages contribute to the estimation of ancestral parameters and their number decreases rapidly, the expected gain in power from increasing sample size is limited even if divergence is relatively recent ($T < 0.5$). The solid lines show the divergence time estimates (scaled by twice the mean of population sizes N_E, N_C, and N_W) obtained for *C. fungosa* in this study (priors *a*).

These issues are relevant in selecting an appropriate study design in systems in which there is a trade off between sampling multiple individuals and generating data for multiple loci or species. Ability to obtain informative population parameters from small numbers of individuals is likely to be particularly important in comparative studies of communities, such as the oak gall system, in which some taxa are rare enough that increasing sample size is not an option. It is therefore useful to ask how much information about ancestral population parameters over phylogeographic timescales can be obtained with minimal sampling. To investigate the influence of sample size, we compared minimal sampling of a single individual per population with an extended sample of three individuals per population. We then use theoretical expectations for the number of surviving lineages given the estimated divergence history (Fig. 2) to consider the likely gain in power for larger sample sizes in our Discussion.

Methods

CHOICE OF LOCI

We obtained sequences for 20 newly developed intronic loci for *C. fungosa* (Table 1) and the closely related species *Caenacis lauta*, which was used as an outgroup in some analyses. These loci included 12 ribosomal protein genes (*RpL10ab, RpL13a, RpL15, RpL27a, RpL37, RpL37a, RpL39, RpS15, RpS18, RpS23, RpS4, RpS8*) and eight regulatory genes (*AntSesB, bellwether, nAcRbeta-64B, Rack1, Ran, sansfille, SUI, Tctp*) (for primer sequences and CG indentifiers see Table S1), all of which are thought to be single copy genes with no known paralogs in insects. Primer development and testing will be described in detail elsewhere (K. Lohse, B. Sharanowski, M. Blaxter, and G. Stone, unpubl. ms.). In short, primers were designed using alignments of Hymenoptera EST data (Sharanowski et al. 2010) and insect sequences from public databases (NCBI). No or little polymorphism at a particular locus may arise either as a result of a low mutation rate (so limiting signal), or a recent coalescent event (and so important to demographic inference), or both. Excluding loci that are invariant in *C. fungosa* results in an upward bias in estimates of population divergence time. To avoid such bias, we used all nuclear loci available for *C. fungosa* (K. Lohse, B. Sharanowski, M. Blaxter, and G. Stone, unpubl. ms.) and tested whether accounting for differences in mutation rate between loci influenced our estimates.

Table 1. Summary statistics of nuclear loci used in the analysis. Loci for which a larger sample of three individuals per population was obtained are shown in bold. Diversity in the minimal single individual sample and divergence to *C. lauta* were calculated for introns (π_{Intron}, K_{Intron}) and synonymous exon sites (π_S, K_S) separately. Also shown are the number of introns (#In) and the total number of polymorphic sites (S) for the single individual samples and locus-specific mutation rate (μ). The normalized product of μ and the total locus length can be taken as a measure of information content (Info). The last column (rec) gives the number of bases that were excluded to trim each locus to the largest nonrecombining fragment according to the four-gamete tests.

Locus	primers	#In	Length (bp)			Diversity			Divergence/mutation rate				
			Total	Intron	Exon	π_s	π_{Intron}	S	K_s	K_{Intron}	μ	Info	rec (bp)
AntSesB	40fb, 40rb	2	606	171	435	0.000	0.008	2	0.076	0.148	0.984	0.981	0
bellwether	33fb, 33rb	1	549	214	335	0.000	0.003	2	n/a	n/a	n/a	n/a	0
nAcRbeta-64B	39f, 39r, 39fb, 39rb	2	728	113	615	0.004	0.000	1	0.371	0.227	1.703	2.039	0
Rack1	18fb, 18rb	2	560	304	256	0.000	0.007	3	0.087	0.052	0.627	0.578	0
Ran	32f, 32r	1	499	202	297	0.011	0.003	2	0.090	0.091	0.802	0.659	0
RpL10ab	19f, 19r	2	955	807	29	0.000	0.003	3	0.072	0.043	0.641	1.001	0
RpL13a	6f, 6r	2	849	718	131	0.000	0.019	21	0.000	0.097	1.414	1.975	0
RpL15	2fb, 2rb	2	618	412	206	0.000	0.002	2	0.233	0.056	1.047	1.065	16
RpL27a	28fb, 28r	2	501	332	169	0.017	0.030	16	0.155	0.101	1.309	1.078	0
RpL37	27f, 27r	1	866	785	81	0.033	0.020	24	0.017	0.123	1.882	2.681	0
RpL37a	36f, 36r	1	220	91	129	0.000	0.000	0	0.408	0.069	1.203	0.436	0
RpL39	16f, 16r	1	463	442	21	0.000	0.000	0	0.000	0.086	1.386	1.055	0
RpS15	20fb, 20rb	1	739	476	263	0.058	0.035	30	0.073	0.091	1.076	1.308	0
RpS18	22f, 22r	1	812	658	154	0.020	0.005	6	0.072	0.052	0.757	1.011	132
RpS23	21f, 21r	1	268	79	189	0.016	0.042	6	0.119	0.127	0.926	0.408	0
RpS4	11f, 11r	1	754	483	271	0.000	0.000	1	0.094	0.083	1.040	1.290	117
RpS8	5f, 5r	1	422	242	180	0.029	0.008	6	0.060	0.034	0.447	0.311	0
sans_fille	35f, 35r	1	446	84	362	0.017	0.000	2	0.140	0.037	0.501	0.367	0
SUI	24f, 24r	1	823	636	186	0.000	0.006	6	n/a	n/a	n/a	n/a	0
Tctp	25f, 25r	2	493	148	345	0.000	0.014	3	0.134	0.088	0.826	0.670	0
Total		**28**	**12171**	**7397**	**4774**			**136**					**265**
MEAN per locus			**608.5**	**369.9**	**238.6**	**0.0092**	**0.0105**	**6.8**	**0.1387**	**0.0727**			
Cox1	pF2/C2413d	n/a	698	n/a		0.090	n/a	24	0.353				

MOLECULAR METHODS

Whole genomic DNA was extracted from specimens stored in 98% ethanol in 50 μl of extraction buffer containing 5% Chelex^{TM}100 resin (Bio-Rad, Hercules, CA). To allow for direct sequencing of PCR products without the need to discriminate between haplotypes in heterozygotes, we used males, which are haploid in Hymenoptera, whenever possible. The exceptions were three female *C. fungosa*, for which haplotypes were distinguished by cloning of PCR products as necessary (see below).

Polymerase chain reactions (PCRs) were performed in 20 μl reactions using the following mix for all primer combinations: 2.0 mL 10× Bioline PCR buffer, 2.0 μl bovine serum albumin (10 mg/mL), 0.8 μl $MgCl_2$ (50 mM), 0.16 μl dNTPs (25 mM each), 0.1 μl Taq Polymerase (5 U/μl, Bioline), 0.2 μl of each primer (20 uM), and 1 μl DNA template.

A generic touchdown PCR protocol was used for all loci: 94°C for 3 min, followed by cycles of 94°C for 15 sec, an annealing step of 40 sec, 72°C for 3 min, and a final step at 72°C for 10 min. The annealing temperature was varied as follows: The first 10 cycles decreased in 1°C increments from 65°C to 55°C, followed by 30 cycles each with an annealing step at 55°C.

To allow comparison of information content in the nuclear loci with a frequently used mitochondrial locus, we also sequenced a 689 bp region of the cytochrome *c* subunit 1 gene (*Cox1*) using primers COI_pF2 and COI_2413d, a modified version of C1-J-2441 (Simon et al. 1994, Table S1). These primers were designed to amplify a fragment largely overlapping the LCO/HCO region of *Cox1* (Folmer et al. 1994), but excluding a poly-T repeat at its $5'$ end present in Chalcidoidea, which causes slippage during PCR resulting in uninterpretable sequence.

All PCR products showing single amplified bands were sequenced directly in both directions using ABI BigDye chemistry (Perkin Elmer Biosystems, Waltham, MA) on ABI 3700 and

3730 sequencers in the GenePool Edinburgh. Chromatograms were checked by eye and complimentary reads aligned using Sequencher version 4.8.

For five loci (*RpS4, RpL27a, RpL37, RpL15b, nAcRbeta*) sequences from female individuals of *C. fungosa* contained putative heterozygous sites or were not readable due to indels. These PCR products were cloned using a mini-Prep kit (Qiagen, Valencia, CA). Five clones were sequenced per locus and individual, one of which was chosen at random for subsequent coalescent analyses. In one case (sample C3, locus *RpS4*) none of the sequenced clones matched the expected product. This sample was excluded from the analysis.

MODEL OF POPULATION DIVERGENCE AND POPULATION SAMPLING STRATEGIES

We consider a simple model of divergence between three putative refugial populations of *C. fungosa*: Asia Minor and Iran (east, E), Balkans and Central Europe (center, C), and Iberia (west, W). This is analogous to a model of divergence between three species (Takahata et al. 1995; Yang 2002) that has been used to estimate divergence times and ancestral population sizes in Great Apes (Rannala and Yang 2003; Patterson et al. 2006), fruit flies (Villablanca et al. 1998; Li et al. 1999), birds (Jennings and Edwards 2005), and plants (Zhou et al. 2007). The model makes the standard population genetics assumptions of random mating within each population, fixed population sizes between divergence events, and no migration after divergence. The first and last assumptions at least are supported by multilocus allele frequency data for the gallwasp hosts in this system (Stone and Sunnucks 1993; Rokas et al. 2003; Stone et al. 2008).

Following recent studies on Hominids and model organisms (Chen and Li 2001; Takahata et al. 1995; Li et al. 1999; Rannala and Yang 2003; Patterson et al. 2006; but see Jennings and Edwards 2005), we initially adopted a sampling scheme that maximizes the number of loci available by using only a single haploid male from each of the three refugial populations listed above. To examine the impact of increased sampling within populations, we generated an extended dataset, comprising three haploid sequences per population for 13 loci and a single sequence per population for the remaining seven loci as before (Table 1 and Table S2). Impacts of further increases in sample size will be considered based on the theoretical expectation of the number of surviving lineages (Fig. 2).

We used ML (Yang 2002) and Bayesian approaches (Rannala and Yang 2003) (described below) (1) to test whether the most likely order of population divergence is compatible with an "Out of the East" scenario, and (2) to estimate divergence times and ancestral population sizes under this scenario using the single individual per population sampling. To investigate the impact of sample size on parameter estimation, Bayesian analyses were repeated using the extended dataset as defined above (Table S2).

ALIGNMENT AND MUTATION RATE

Cecidostiba fungosa and *C. lauta* sequences were aligned in ClustalW and checked by eye (GenBank accession numbers HM208872-HM209026). Exonic regions were assigned by comparison with *D. melanogaster* protein sequences and checked for an open reading frame. Indels in the alignment were treated as missing data.

In the ML and Bayesian analyses, all model parameters are scaled by the per site mutation rate, μ. Conversion of the scaled time between divergence events (γ) into real times (τ), and of the scaled mutation rate (θ) into effective population sizes (N_e), therefore requires an estimate of μ and its incorporation into the relationships $\gamma = \tau\mu$ and $\theta = 4N_e\mu g$, where g is the average generation time in years. Note that for haplodiploids $N_{e\_hd} = (9N_fN_m)/(2N_f + N_m)$, where N_f and N_m are the number of males and females, respectively, in a randomly mating population. Assuming equal sex ratio and variance in fitness between sexes, N_{e_hd} is 0.75 N_{e_d} (Hedrick and Parker 2003).

To calculate a mean estimate of μ for our loci, we first estimated a synonymous genome-wide mutation rate for the closely related pteromalid wasp genus *Nasonia,* using a divergence time of 0.4 million years ago (mya) between *N. giraulti* and *N. longicornis* (Campbell et al. 1993; Oliveira et al. 2008; Raychoudhury et al. 2009) and a nuclear genome-wide distance at synonymous sites (K_s) of 0.011 between these species (Oliveira et al. 2008). With $\mu = K_s/2t$, these values give $\mu = 1.375\ 10^{-8}$ b/yr. The *Nasonia* divergence time was derived by applying observed bacterial mutation rates to *Wolbachia* symbionts infecting the two *Nasonia* species (Raychoudhury et al. 2009). However, the resulting mutation rate estimate is also remarkably consistent with the few other molecular clock calibrations that exist for insects, such as the calibration of 1.11×10^{-8} b/yr for Hawaiian Drosophilids using island ages (Tamura et al. 2004).

To apply the *Nasonia* mutation rate to our intron-rich (and so partially noncoding) sequences, we scaled it by the ratio of the observed average divergence between *C. fungosa* and *C. lauta* at synonymous sites, K_s over the average divergence across all sites K_{Total}. This yields a factor of 0.478, so the total average mutation rate for our loci is $\mu = 1.375\ 10^{-8} \times 0.478 = 6.27\ 10^{-9}$ b/yr. Note that because this is an average across all sites, it is lower than the mutation rate for synonymous coding sites. This calculation incorporates any mutational constraints on introns and coding sites in *C. fungosa* without making a priori assumptions about intron evolution. We estimated a relative mutation rate for each locus as the observed K_{Total} at each locus over the average K_{Total} (Chen and Li 2001; Yang 2002; Jennings and Edwards 2005), shown in Table 1.

To calculate ancestral effective population sizes, we assumed an average generation time of $g = 0.5$ years for *Nasonia* and *C. fungosa*. This is reasonable for *C. fungosa,* which attacks both sexual spring galls and asexual autumn galls (Askew 1961; Schönrogge et al. 1995, 1996a) (as synonyms *C. adana* and *C. hilaris*), and for temperate populations of *Nasonia*. For comparison with mitochondrial node ages, we calculated a mutation rate for *Cox1* using the Jukes-Cantor-corrected distance between *N. giraulti* and *N. longicornis* at this locus and a divergence time of 0.4 mya as before. This gives 22.3% (Oliveira et al. 2008) divergence per site and million years. We compared this locally calibrated clock with estimates obtained in previous studies using the commonly assumed arthropod mitochondrial clock of 2.3% per site and million years (Brower 1994). Despite the obvious shortcomings of the "Brower clock," comparison of relative node ages in this way is valid as long as the same calibration is used across taxa, and a molecular clock assumption is tested and supported in each taxon, as here.

RECOMBINATION TESTS AND GENE TREE RECONSTRUCTION

Both phylogenetic reconstruction and the coalescent analyses described below make the crucial assumption of no recombination within loci. We determined the minimum number of recombination events using a four-gamete test in DNAsp (Rozas and Rozas 1995) on the largest alignment of each locus. Three loci (*RpS4, RpS18, RpL15*) showed evidence for recombination and were trimmed to the largest nonrecombining block (Galtier et al. 2000; Jennings and Edwards 2005) (shown in Table 1).

Although both the ML and Bayesian approaches described below use site patterns directly and do not rely on estimated gene trees, we reconstructed trees to visualize the data and to test the molecular clock hypothesis that is implicit in both approaches. ML trees were reconstructed for each locus in PAUP* (Swofford 2001). For single individual alignments (triplets), this was done using exact searches, whereas for the three individual per population alignments branch and bound searches were used. Loci varied considerably in relative intron length and hence in base composition. We therefore assumed a single substitution rate but unequal base frequencies (Felsenstein 1981). To test the support for internal nodes in each triplet gene tree, 1000 bootstrap replicates were performed taking a bootstrap value of >70% to indicate strong nodal support (Hillis and Bull 1993). We compared rooting with a strict molecular clock to rooting with *C. lauta* for the triplet gene trees (Tajima 1993; Jennings and Edwards 2005; Tamura et al. 2007). To further test the validity of the molecular clock assumption, we performed Tajima's 1 − degree of freedom test on each triplet (Tajima 1993; Jennings and Edwards 2005; Tamura et al. 2007). This nonparametric test is designed for triplet samples given a known species topology and is simpler and more powerful than similar model-based tests (Tajima 1993; Nei and Kumar 2000; Jennings and Edwards 2005).

MAXIMUM LIKELIHOOD ANALYSIS

For minimal sampling, only four parameters in the three-population divergence model matter: the two divergence times $\tau_{C/W}$ and $\tau_{E/C/W}$ and the sizes of the two ancestral populations $N_{C/W}$ and $N_{E/C/W}$ (Fig. 1) and an exact likelihood approach to inference is possible. The program Ne3sML numerically maximizes the likelihood for a given population/species topology (Yang 2002). By default the method assumes an infinite sites mutation model and a molecular clock. Given the level of polymorphism observed in *C. fungosa* (Table 1), this simple model of sequence evolution seems appropriate. For example, if diversity at silent sites (synonymous exon sites and introns) is 0.01 (Table 1), the chance of a back mutation is 10^{-4} per site. Because we are analyzing slightly fewer than 10^4 silent sites in total, we expect to see at most a single back-mutation in the entire dataset and can safely ignore more complicated mutation models.

The likelihood approach of Yang (2002) differs crucially from methods that estimate a species tree conditional on a set of reconstructed gene trees (Degnan and Salter 1995; Maddison and Knowles 2006; Carstens and Knowles 2007; Liu and Pearl 2007; Degnan and Rosenberg 2009; Kubatko et al. 2009) in that it uses the site information directly. The method integrates over all possible gene tree topologies and branch lengths at each locus and computes the joint log likelihood for a given population history (topology and parameter estimates) as the sum over the log likelihoods of individual loci (Yang 2002; Rannala and Yang 2003). The advantage of this is that in contrast to gene tree species tree approaches (Liu and Pearl 2007; Degnan and Rosenberg 2009; Kubatko et al. 2009), information from unresolved or poorly resolved loci is incorporated automatically. This is particularly important in recently diverged populations. For example, a monomorphic locus resulting from a recent coalescence event would be excluded from analyses conditional on gene tree reconstruction as uninformative, resulting in upwardly biased estimates of divergence time.

We first compared the likelihood of all three possible population tree topologies. Although assessing the statistical significance of nonnested models is difficult in a likelihood setting, models may be ranked by their likelihood (Carstens et al. 2009). Under the "Out of the East" scenario, central and western populations are derived from a shared ancestral population in the center, which in turn split from a common ancestral population in the east, that is, the population tree topology is (E, (C, W)) (Fig. 1). The two alternative topologies are (W, (C, E)), which corresponds to an "Out of the West" scenario, and (C, (E, W)) which is difficult to interpret in the geographic context of *C. fungosa* populations, because it is unclear where the two ancestral populations would be located.

ML analyses under the most likely population history were performed for two different mutational models. The simplest model assumes a single mutation rate across all loci. We reran this analysis using the relative rates calculated for each locus as described above (Table 1), thereby accounting for possible rate heterogeneity (Table 3).

BAYESIAN ESTIMATION OF DIVERGENCE TIMES AND ANCESTRAL POPULATION SIZES

MCMCcoal (Rannala and Yang 2003) is the Bayesian equivalent of the ML approach described above. The program uses Markov chain Monte Carlo (MCMC) sampling to estimate posterior probabilities for all model parameters conditional on prior distributions. If multiple individuals per population are sampled, the three population sizes between the present and the most recent divergence event (i.e., N_E, N_C, N_W) (Fig. 1) are modeled as additional parameters. Note that the parameterization in MCMCcoal differs slightly from Ne3sML, as the former uses divergence times rather than internode intervals.

In a Bayesian framework, support for alternative but nonnested models can be compared using Bayes factors (Kass and Raftery 1995). Natural logarithms (ln) of harmonic mean likelihoods (HML) were calculated for each population tree topology (using prior means in analysis *a* described below) to test support for the "Out of the East" scenario. Following Kass and Raftery (1995), values of twice the difference in lnHML (2ΔlnHML) of 2–6, 6–10, and >10 represent, respectively, positive, strong, and very strong support for the model with higher likelihood.

Because in the case of *C. fungosa* we have no prior knowledge of the model parameters, we used exponentially distributed priors (shape parameter $\alpha = 1$) for all parameters (Jennings and Edwards 2005). To check how sensitive posterior estimates are to prior settings, all analyses were performed twice using different prior means by adjusting β, the scale parameter of the gamma distribution (Table 4). In the first analysis (*a*), we set prior means to ∼0.150 mya and ∼0.050 mya for $\tau_{E/C/W}$ and $\tau_{C/W}$, respectively ($\beta = 380$) and ∼215,000 for both ancestral population sizes ($\beta = 1520$). In the second analysis (*b*), the prior means for all parameters were increased by an order of magnitude (i.e., changing β to 38 and 152) (Table 4). Although the individual parameter values are arbitrary, these two sets of priors should be different enough to assess the robustness of the Bayesian estimation (Jennings and Edwards 2005). Given that incorporating relative mutation rates did not improve estimation using the ML method (see Results), for simplicity all Bayesian analyses were performed assuming a single mutation rate across all loci. Runs were continued for 10^6 generations with a burn-in of 10^5 and repeated using different random number seeds to check for convergence.

Results

GENE TREES

When only a single individual was sampled from each refugial population, phylogenetic reconstructions for eight of the 18 polymorphic nuclear loci supported the "Out of the East" topology (E, (C, W)) (Fig. 3A), as did the mitochondrial locus *Cox1* (Fig. 2D). Of the remaining loci, two supported each of the two incongruent topologies (Fig. 2B, C) and six showed an unresolved topology (*RpL15, RACK1, ran, Tctp, sansfille, SUI*). Clock-rooted and outgroup-rooted topologies agreed for all resolved loci, but bootstrap support was generally weaker for outgroup rooting (Fig. 3). Although this is not a formal test, the majority of resolved gene trees thus support the "Out of the East" hypothesis (Fig. 1). Tajima's $1 - D$ test rejected a strict molecular clock for only two of 20 loci (*RpS15, RpL 37*). Thus the majority of loci meet the clock assumption implicit in the ML and Bayesian approaches used here.

Increasing sample size to three individuals from each refugial population resulted in increased variation in gene tree topology (Fig. 4). Despite the many unresolved nodes in some trees, Figure 4 reveals extensive incomplete lineage sorting between *C. fungosa* populations, resulting in a "forest" of largely incongruent gene trees.

MAXIMUM LIKELIHOOD ANALYSES

The population tree topology (E, (C, W)) had a higher likelihood than either of the two alternative topologies (C, (E, W)) and (W, (C, E)), consistent with the "Out of the East" hypothesis (Table 2). The maximum likelihood estimates (MLEs) of model parameters are broadly consistent between the variable rate (18 loci) and single rate mutational models (using the same 18 loci). However, because the variable rates model has a lower log likelihood, the simpler single rate model was used in all subsequent analyses including the Bayesian runs (Table 3). This also allowed the loci *SUI* and *bellwether,* for which no outgroup sequences could be obtained, to be included in the analyses, giving a total of 20 loci.

Under the "Out of the East" topology (E, (C, W)), the MLE for the older population splitting time between the Iranian population and the ancestor of Hungary and Spain, $\tau_{E/C/W}$, is estimated as 0.110 mya (Table 3). The MLE for $\theta_{E/C/W}$ corresponds to an ancestral population with an effective size of 614,000 before this first split. However, both the MLE for the time between the two population splits, $\tau_{E/C/W} - \tau_{C/W}$ and the population size during that time, $N_{C/W}$ are close to zero, suggesting that Iberian and Hungarian populations may have split almost

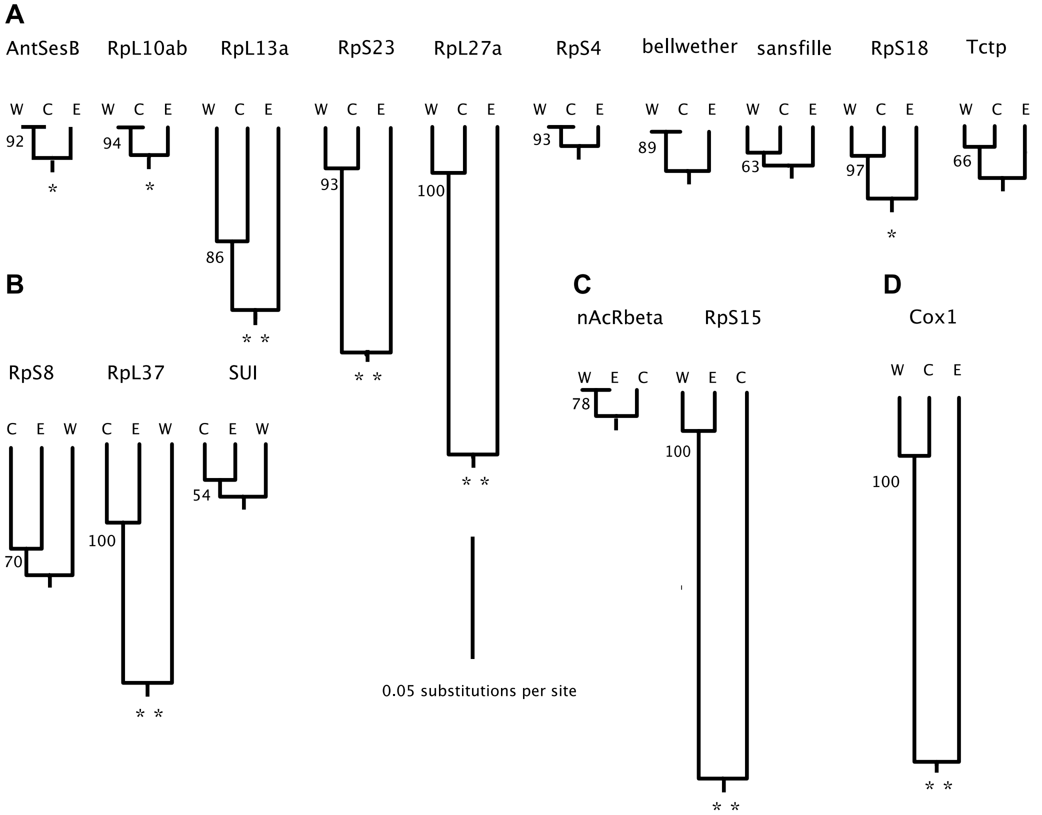

Figure 3. ML trees reconstructed for nuclear loci and *Cox1* assuming a strict molecular clock. Bootstrap proportions for the internal node are shown next to each tree. Loci with unresolved topologies (<50% bootstrap support) are not shown. Eight loci have a topology congruent with the "Out of the East" hypothesis (E, (C, W)) (A), two each have topology (W, (C, E)) (B) and (C, (E, W)) (C). The mitochondrial locus *Cox1* is also congruent with "Out of the East" (D). Bootstrap support using rooting with *C. lauta* is indicated with asterisks (* > 50%, ** > 70%) below each tree.

immediately after the initial divergence from the ancestral Eastern population (Table 3).

BAYESIAN ESTIMATION OF DIVERGENCE TIMES AND ANCESTRAL POPULATION SIZES

Minimal sampling

Bayes factor comparison of lnHML (Table 2) shows that the "Out of the East" model fits the data significantly better than either of the alternative population tree topologies. The contrasting sets of priors *a* and *b* had little impact on posterior estimates of three of the four model parameters (Table 4, Fig. 5A, B, D). Posterior mean ages for the split between eastern populations and the common ancestor of central and western populations $\tau_{E/C/W}$ were 0.118 mya and 0.134 mya in analyses *a* and *b* respectively, with values of 0.043 mya and 0.046 mya for the divide between central and western populations $\tau_{C/W}$ (Table 4). This comparatively long interval between the two divergence times ($\tau_{E/C/W} - \tau_{C/W}$) is in apparent contrast to the results of the ML analysis. However, the 95% confidence intervals for the two divergence times overlap in both prior settings *a* and *b*, such that the lower confidence interval for $\tau_{E/C/W} - \tau_{C/W}$ includes zero, compatible with divergence between western and central populations occurring immediately after the initial split from the ancestral eastern population. Likewise, the posterior estimate for the effective size of the population ancestral to all three refugial populations ($N_{E/C/W}$) was minimally influenced by the prior (Table 4, Fig. 5D) (551,000 for *a* and 585,000 for *b*).

In contrast, posterior distributions for the effective size of the population ancestral to central and western populations, $N_{C/W}$, differed considerably between prior settings *a* and *b* (197,000 and 698,000) (Table 4, Fig. 5C). $N_{C/W}$ was also the parameter with the largest variance, the 95% confidence interval spanning two orders of magnitude (priors *b*, Table 4). Notably, with both prior settings, posterior distributions of $N_{C/W}$ peak at the origin (Fig. 5C). This suggests that there is little information about $N_{C/W}$ in the data, with posterior distributions largely reconstructing the prior.

To investigate whether the uncertainty in $N_{C/W}$ can account for the apparent difference in ML and Bayesian estimates of the interval between population splits ($\tau_{E/C/W} - \tau_{C/W}$), we carried out a third MCMCcoal run (Table 4, priors *c*). When the prior mean for $N_{C/W}$ is set to a very low value (2100), the posterior distribution for $\tau_{C/W}$ shifts markedly toward the right (Fig. 5A) such that the

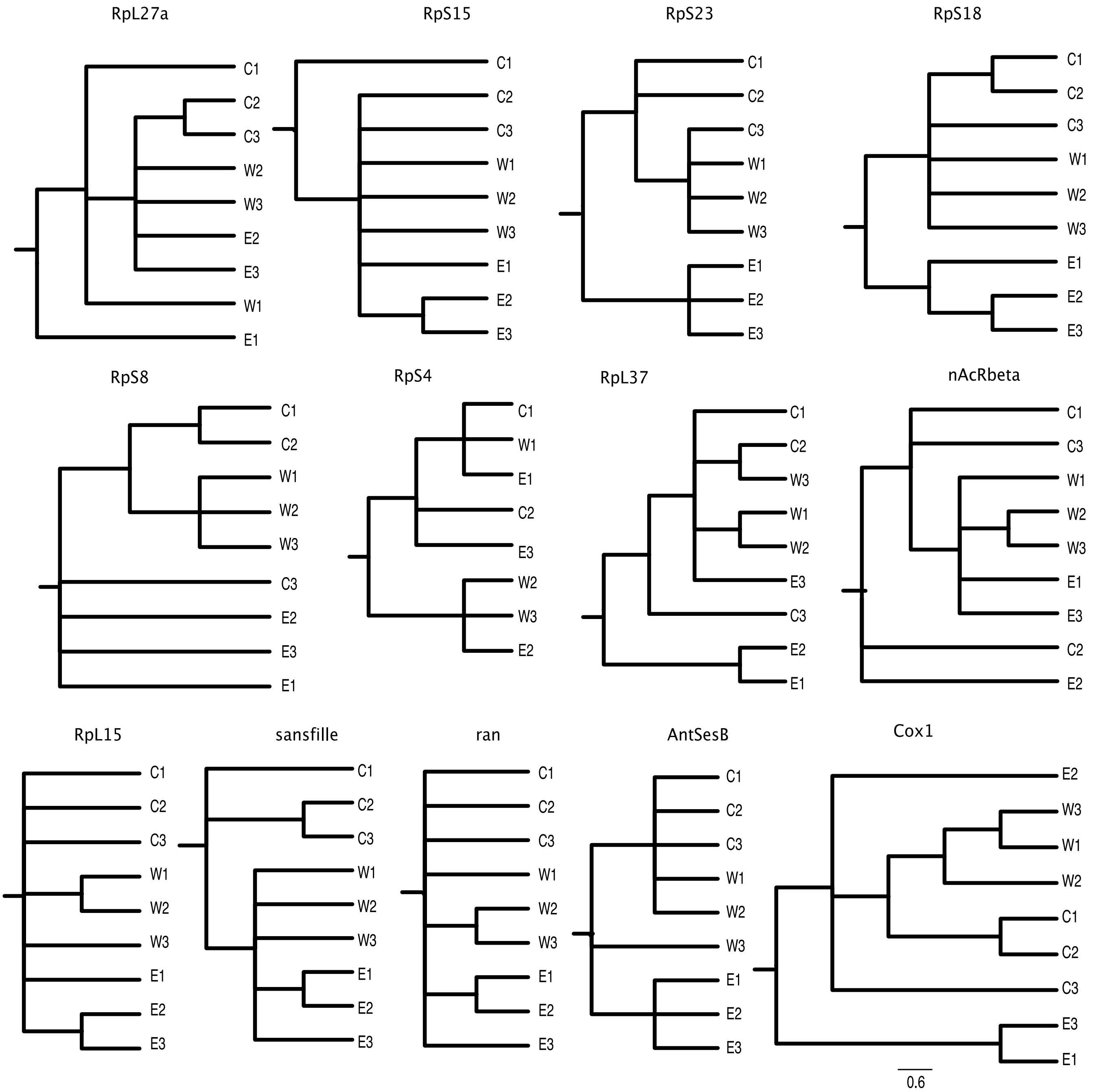

Figure 4. **ML trees for the extended sampling of three individuals (labeled 1–3) per population for 12 nuclear loci and *Cox1* rooted using *C. lauta*. *RpL37a* is monomorphic and not shown. Although on average samples from the same population are more closely related than those from different populations, there is extensive lineage sorting, resulting in a "forest" of partially incongruent gene trees.**

two divergence events are estimated to have happened in close succession (0.091 and 0.089 mya) in agreement with the ML results (Table 3).

Extended (three individual) sampling

MCMCcoal analyses of the extended (three individual per population) dataset again gave strongest support to the "Out of the East" scenario (Table 2). Although Bayes factor comparison strongly rejects the "Out of the West" topology (W, (C, E)), the second alternative topology (C, (E, W)) does not provide a significantly worse fit to the data (Table 2).

Parameter estimates agree well with those obtained when only a single individual per population was sampled (Table S3 and Fig. S1). However, increased sampling does have some influence on parameter estimation. First, estimates of $N_{C/W}$, are larger and less sensitive to prior settings when three individuals per

Table 2. Comparison of support for alternative population tree topologies, using the lnL of the maximum likelihood estimation (NeML3s) and the harmonic mean likelihood (lnHML) in the Bayesian analyses. In each case the "Out of the East" topology has the highest likelihood (in bold). Values in parentheses show the ln Bayes factor (2ΔlnHML) of the "Out of the East" hypothesis relative to alternatives. Topologies that fit significantly worse than the "Out of the East" hypothesis are indicated with asterisks, using a ln Bayes factor of 2–6 to indicate positive support (*), 6–10 to indicate strong support (**), and >10 to indicate very strong support (***), following Kass and Raftery (1995).

	Population tree topology		
	Out of the East (E, (C, W))	Out of the West (W, (C, E))	(C, (E, W))
NeML3s (single triplet) lnL	**−796.94**	−799.06	−799.05
MCMCcoal (*a*, single triplet) ln(har.mean)	**−19100.692**	−19103.820 (lnBF=6.25)**	−19103.060 (lnBF=4.73)*
MCMCcoal (*a*, extd. triplet) ln(har.mean)	**−19558.237**	−19563.899 (lnBF=11.324)***	−19558.997 (lnBF=0.76)

population are sampled for both prior sets *a* and *b* (Table S3). Second, the posterior distributions for $\tau_{C/W}$ are now unimodal, rather than L-shaped with a maximum at the origin (Fig. S1). However, this has little impact on the variance of the posterior. For example, the 95% confidence interval for $\tau_{C/W}$ is 0.005–0.136 mya (priors *a*) in the analysis of the extended samples of three individuals per population, compared with 0.002–0.121 mya when sampling a single individual (Table 4). Taken together this suggests that increasing sample size per population to three haploid individuals adds some, but not much, power to the estimation of model parameters.

Sampling multiple individuals per population we can also estimate the effective sizes of the three sampled populations between the present and the first divergence events, N_E, N_C, N_W. (Table S3). Although estimates of these parameters had fairly wide confidence intervals and were sensitive to prior settings, their relative magnitude was consistent across analyses. N_C was always the largest followed by N_E and N_W. It is also noteworthy that all three estimates were smaller than those obtained for ancestral populations, paralleling the findings of Jennings and Edwards (2005) and previous results in Great Ape studies (Chen and Li 2001; Yang 2002; Patterson et al. 2006).

GENE DIVERGENCE TIMES

Following Jennings and Edwards (2005), we calculated Jukes Cantor distances (D) to estimate coalescence times for each divergence event (D/2) and compared the average distance across loci with the estimated population divergence time and the mitochondrial (*Cox1*) node ages for both single and three individual samples. In both cases, nuclear genes sampled from central and western populations diverged on average almost 0.4 million years (or three glacial periods) prior to the estimated population divergence (Fig. 6). Coalescence times estimated for *Cox1* depend on the assumed mutation rate. Applying the calibration by Oliveira et al. (2008), both coalescence times for *Cox1* (0.013 MY and 0.145 MY respectively) are younger than the average coalescence at nuclear genes but are well within the 95% of the estimated population divergence (Table 4). Using Brower (1994), mitochondrial coalescence between the ancestor of central and western samples and the eastern sample (1.433 mya) predates the average coalescence times for nuclear genes (0.714 mya), whereas the mitochondrial coalescence time between central and western samples (0.125 mya) is still more recent than that for nuclear genes (0.467 mya) (Fig. 6).

Discussion

We analyzed a large multilocus dataset under the simplest possible model of divergence between three populations to make quantitative inferences about the longitudinal history of *C. fungosa*. Reconstructing the genealogical histories of individual loci leads to a "forest" of largely incongruent and often poorly resolved

Table 3. Maximum Likelihood estimates (MLEs) of ancestral population sizes and population divergence times for refugial populations of *C. fungosa* assuming a population tree topology (E, (C, W)). Corresponding Ne and τ values are shown in bold in brackets. The simplest mutational model assumes a single rate for all loci. In the variable rates analysis, a relative mutation rate was computed for each locus from divergence to *C. lauta*.

	MLE, single rate (20 loci)	MLE, single rate (18 loci)	MLE, variable rates (18 loci)
$\theta_{E/C/W}$ ($\boldsymbol{N_{E/C/W}}$)	0.0076979 (**614,000**)	0.007995 (**637,000**)	0.008933 (**712,000**)
$\theta_{C/W}$ ($\boldsymbol{N_{C/W}}$)	0.000008 (<**1000**)	0.000002 (<**1000**)	0.000003 (<**1000**)
$\gamma_{E/C/W}-\gamma_{C/W}$ ($\tau_{E/C/W}$ in my)	0.0000032 (<**0.001**)	0.000001 (<**0.001**)	0.000001 (<**0.001**)
$\gamma_{C/W}$ $\tau_{C/W}$ in my	0.0006924 (**0.110**)	0.000712 (**0.114**)	0.000756 (**0.121**)
lnL	−853.486	−794.948	−796.913

Table 4. Prior and posterior means and 95% confidence intervals for divergence times and ancestral population sizes in Bayesian analyses using minimal sampling of a single individual per population and assuming an "Out of the East" population tree topology (E, (C, W)). Corresponding Ne and τ values are shown in bold below. All analyses (*a–c*) assumed exponentially distributed priors (α=1), but differed in their prior means. The population size in between the two divergence events $N_{C/W}$ is the parameter most sensitive to prior choice and has the widest confidence interval.

Parameter	(α, β)	Prior mean (95% confidence interval)	Posterior mean (95% confidence interval)
		priors *a*	
$\theta_{E/C/W}$	(1, 380)	0.00271 (0.00011, 0.00968)	0.00691 (0.00239, 0.01830)
$N_{E/C/W}$		**216,000 (10,000, 772,000)**	**551,000 (190,000, 1,459,000)**
$\theta_{C/W}$	(1, 380)	0.00267 (0.00009, 0.00982)	0.002477 (0.00033, 0.00727)
$N_{C/W}$		**213,000 (8,000, 783,000)**	**197,000 (26,000, 580,000)**
$\gamma_{E/C/W}$	(1, 1519)	0.00095 (0.00012, 0.00276)	0.00074 (0.00019, 0.00139)
$\tau_{E/C/W}$		**0.151 my (0.019 my, 0.440 my)**	**0.118 my, (0.030 my, 0.221 my)**
$\gamma_{C/W}$	(1, 1519)	0.000329 (0.00001, 0.00119)	0.00027 (0.00001, 0.00076)
$\tau_{C/W}$		**0.052 my, (0.002 my, 0.189 my)**	**0.043 my, (0.002 my, 0.121 my)**
		priors *b*	
$\theta_{E/C/W}$	(1, 38)	0.02664 (0.00083, 0.09691)	0.00734 (0.00464, 0.01121)
$N_{E/C/W}$		**2,124,000, (66,000, 7,726,000)**	**585,000 (370,000, 894,000)**
$\theta_{C/W}$	(1, 38)	0.02639 (0.00064, 0.09669)	0.00875 (0.00050, 0.05260)
$N_{C/W}$		**2,104,000 (51,000, 7,709,000)**	**698,000 (40,000, 4,141,000)**
$\gamma_{E/C/W}$	(1, 152)	0.00980 (0.00113, 0.02918)	0.00084 (0.00023, 0.00156)
$\tau_{E/C/W}$		**1.563 my (0.180 my, 4.653 my)**	**0.134 my (0.037 my, 0.249 my)**
$\gamma_{C/W}$	(1, 152)	0.00326 (0.00008, 0.01198)	0.00029 (0.00001, 0.00084)
$\tau_{C/W}$		**0.520 my (0.131 my, 1.910 my)**	**0.046 my (0.002 my, 0.134 my)**
		priors *c*	
$\theta_{E/C/W}$	(1, 380)	0.00257 (0.00004, 0.00961)	0.00741 (0.00485, 0.01088)
$N_{E/C/W}$		**205,000 (3,000, 766,000)**	**591,000, (387,000, 868,000)**
$\theta_{C/W}$	(1, 38000)	0.00003 (<0.00001, 0.00009)	0.00005 (0.00001, 0.00015)
$N_{C/W}$		**2,100 (<1000, 7,000)**	**5,000, (<1,000, 13,000)**
$\gamma_{E/C/W}$	(1, 1519)	0.00096 (0.00011, 0.00277)	0.00057 (0.00011, 0.00111)
$\tau_{E/C/W}$		**0.153 my (0.017 my, 0.442 my)**	**0.091 my (0.018 my, 0.177 my)**
$\gamma_{C/W}$	(1, 1519)	0.00033 (0.00001, 0.00122)	0.00056 (0.00011, 0.00108)
$\tau_{C/W}$		**0.053 my (0.013 my, 0.195 my)**	**0.089 my (0.018 my, 0.172 my)**

gene trees (Fig. 4), which individually contain little information about the underlying population history. However, analyzing these data jointly in a coalescent framework, the relationship between major refugial populations of *C. fungosa* can be described as a quantified population tree, which includes relevant population genetic parameters (Fig. 7). This is a considerable improvement over previous phylogeographic studies in this system, which have largely been based on mitochondrial sequence data and allozymes (Rokas et al. 2001, 2003; Stone et al. 2001; Challis et al. 2007; Stone et al. 2009) and allows us to quantify important aspects of the phylogeographic history of *C. fungosa*.

First, both likelihood and Bayes factor comparisons of population tree topologies (Table 2) support the "Out of the East" scenario for *C. fungosa*.

Second, both ML and Bayesian estimates for the time of the first population split between the eastern population and the common ancestral population of central and western populations $\tau_{E/C/W}$ fall well within the late Pleistocene. Likewise, both methods suggest that the more recent divergence between central and western populations ($\tau_{C/W}$) occurred either during the last interglacial or glacial period. However, because the MLE for the time between population splits ($\tau_{E/C/W} - \tau_{C/W}$) is effectively zero and the 95% confidence intervals for the two divergence times overlap in all Bayesian analyses, we cannot exclude the possibility that the two population splits happened in close succession.

Finally, the present coalescent analyses provide information about the effective sizes of ancestral and present populations. Although our estimates of both ancestral population sizes, in particular $N_{C/W}$, have large confidence intervals and, in the case of $N_{C/W}$, are sensitive to prior settings (discussed below), they provide an important comparison with model organisms. For example the observed diversity in *C. fungosa* ($\pi_s = 0.92\%$, Table 1) is comparable with that in non-African populations of *D. melanogaster* ($\pi_s = 1.33\%$) (e.g., Andolfatto 2001, Table 3). Similarly, estimates

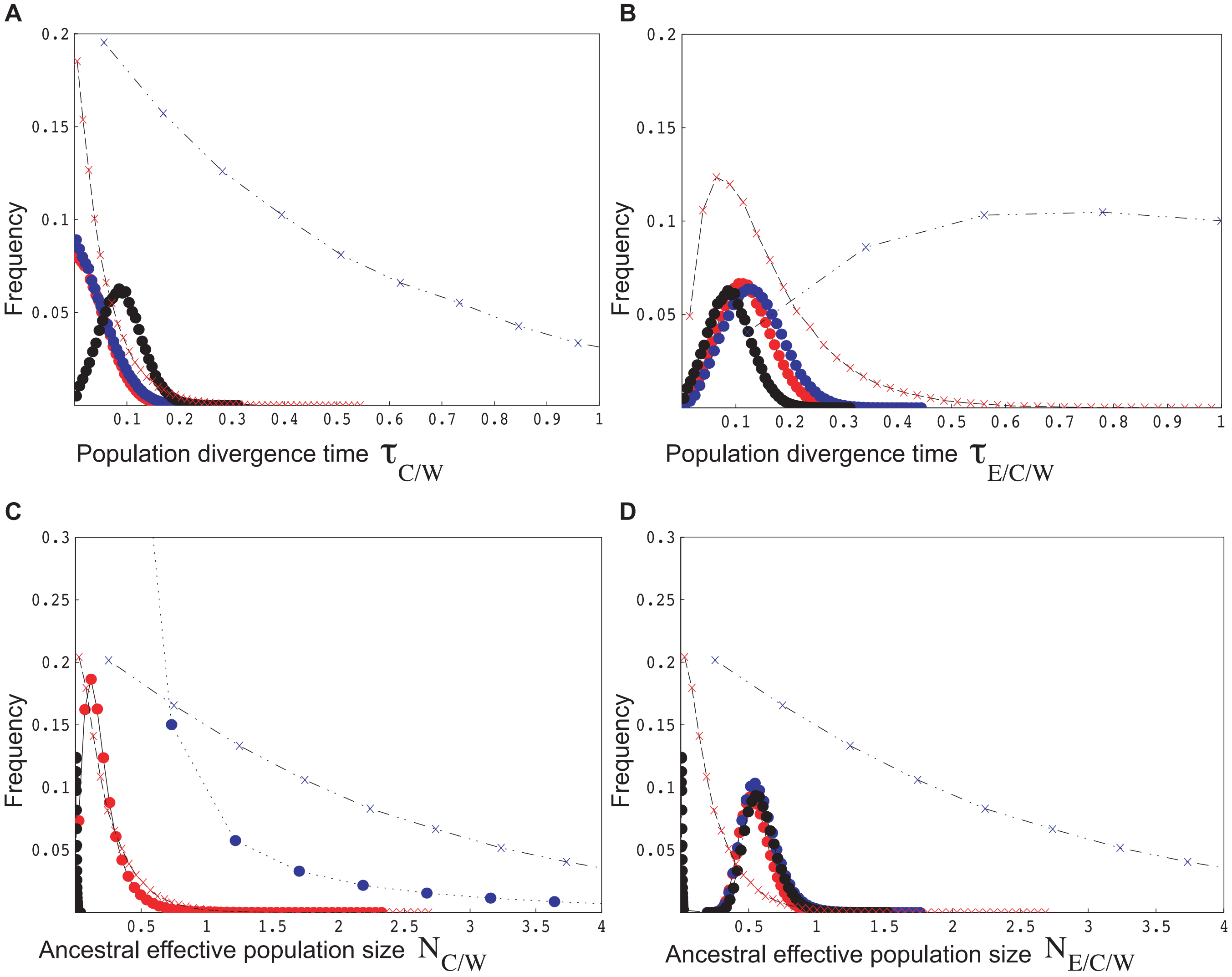

Figure 5. Prior and posterior distributions of parameters under the "Out of the East" model of population divergence using minimal sampling of a single individual per population. Prior distributions for the first two MCMCcoal analyses are shown as dashed lines (a, mixed long and short dashes between blue symbols; b, long dashes between red symbols), posterior distributions for the single triplet analysis are in color (*a*, red; *b*, blue; *c*, black). Whereas $\tau_{E/C/W}$ (B) and $N_{E/C/W}$ (D) are little influenced by the prior means, $N_{C/W}$ (C) is extremely sensitive. This parameter is also confounded with $\tau_{C/W}$. When setting a low prior mean for $N_{C/W}$ (analysis *c*) the posterior distribution for $\tau_{C/W}$ shifts markedly toward the right (see black line in A). Note that despite $\alpha = 1$ for all model parameters, the prior distribution for $\tau_{E/C/W}$ (B) is not exponential because of the constraint $\tau_{E/C/W} > \tau_{C/W}$.

for the effective population sizes of *D. melanogaster* of 10^6 (Andolfatto and Przeworski 2000) and for effective size of the ancestor of *D. melanogaster* and *D. simulans* of $Ng = 3.9 \times 10^5$ (Li et al. 1999) agree with our results for *C. fungosa* in order of magnitude. If effective population sizes of 10^6 are the rule in insect parasitoids, their longitudinal histories will inevitably involve extensive incomplete lineage sorting, strengthening the case for multilocus approaches for meaningful phylogeographic inferences.

How do these results compare with those obtained from single gene trees both in *C. fungosa* and in other codistributed oak gall parasitoids and their hosts? In *C. fungosa,* the topology of the inferred population tree (Fig. 7) is congruent with both the majority of resolved nuclear gene trees as well as the mitochondrial gene tree when a single individual per refugial population was sampled (Fig. 3). More generally, the eastern origin of *C. fungosa* is consistent with the mitochondrial gene tree for another oak gall parasitoid, *M. stigmatizans* (Hayward and Stone 2006), with mitochondrial and nuclear gene trees in the parasitoid *M. dorsalis* (Nicholls et al. 2010) and three species of host gall wasps (Rokas et al. 2003; Challis et al. 2007; Stone et al. 2007, 2009).

Although by definition gene divergence must predate the divergence of populations, our results suggest that the magnitude of this difference is considerable in *C. fungosa* and very relevant

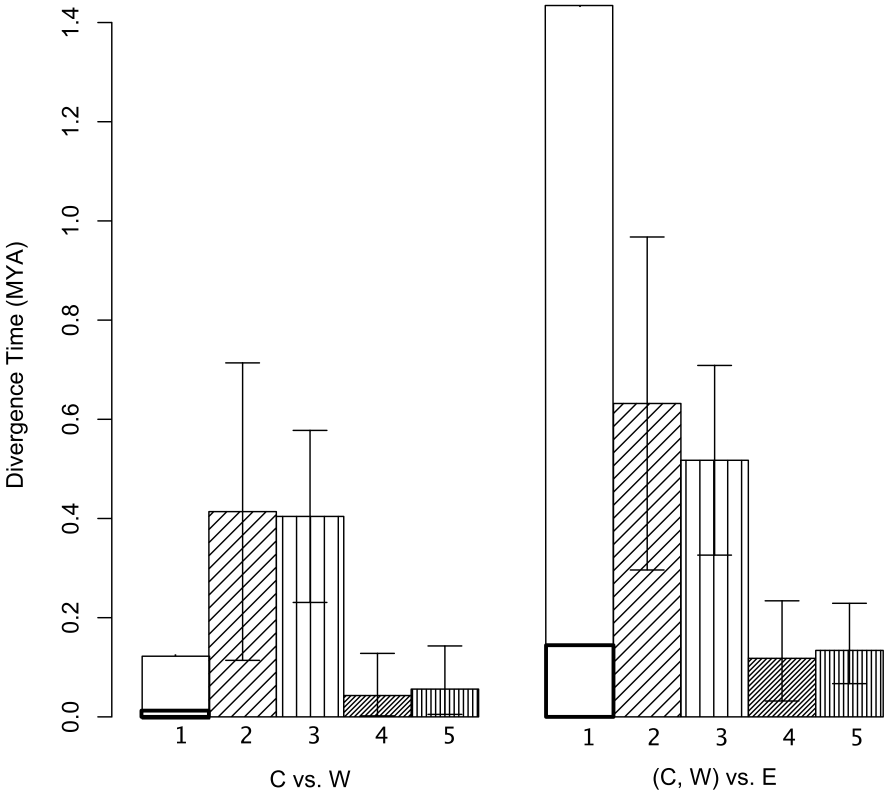

Figure 6. Divergence times for the two splits in the Out of the East model (C vs. W left and (C,W) vs. E right). The figure shows that Bayesian estimates (prior settings *a*) of population divergence times for both single and extended triplet samples (columns 4 and 5 in each figure, respectively) are more recent than the mean coalescence time across nuclear loci for both sampling schemes (columns 2 and 3 in each figure). Mitochondrial divergence (column 1) was calculated from node ages in the single triplet tree using both Oliveira et al.'s (2008) rate calibrated from *Nasonia* sister species (lower estimates, bold bars in column 1) and the widely applied rate estimate of Brower (1994) (higher estimates, column 1). Error bars show ±95% confidence limits.

for our interpretation of its Pleistocene history. It is noteworthy that the estimates for $\tau_{E/C/W}$ coincide with the last (Eemian) interglacial, 0.130–0.115 mya, which suggests that divergence between refugial populations is as recent as it possibly can be (given the definition of glacial refugia). We know from the fossil record that both oaks (Velichko et al. 2005) and associated gall wasp species (Stone et al. 2008; van der Ham et al. 2008) known to be attacked by *Cecidostiba* expanded their range in Central and

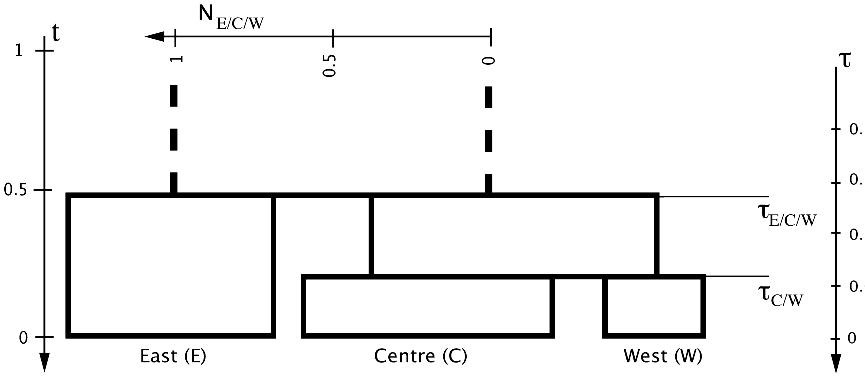

Figure 7. Population tree for Western Palearctic *C. fungosa* inferred from 20 genetrees. Means of posterior distributions of model parameters were obtained from the Bayesian analysis (priors *a*, extended sampling of three sequences per population, Table S3 and figure S4). The widths of blocks correspond to effective population sizes (scale at top). Divergence times are shown on two different scales: τ in MY (right-hand scale), and $t = t/(2N_{E/C/W})$ generations assuming two generations per year, that is, $g = 0.5$ (left-hand scale). Note that all blocks have a greater width than height such that pairs of lineages sampled from the same population are more likely to coalesce in their ancestral population.

Northern Europe during this period. It is thus plausible for population divergences associated with westward range expansions of *C. fungosa* to have occurred over a similar timescale.

Although the unknown error in the mitochondrial clock and the large discrepancy between different calibrations (Brower 1994; Oliveira et al. 2008) make a direct comparison with mitochondrial dates problematic, it is nevertheless reassuring that the mitochondrial ages obtained for *C. fungosa* fall within the 95% confidence interval of (Oliveira et al. 2008) or predate (Brower 1994) the estimated time of population divergence (Fig. 6), as they should. A mitochondrial divergence more recent than that inferred for the population would be inconsistent with the assumed model, and require gene flow between populations. However, it is noteworthy that regardless of the mitochondrial mutation rate used, the *Cox1* divergence times are very different from the average divergence times at nuclear genes (Fig. 6). This demonstrates the extremely large variance in coalescence times and highlights the danger of over-interpreting node ages of single gene trees. An additional problem with mitochondrial mutation rate calibrations is that they are likely to be confounded by the selective dynamics of bacterial endosymbionts (Oliveira et al. 2008), the prevalence of which is known to differ both between populations and closely related species of Pteromalids (Weinert et al. 2009, A. Aebi, unpubl. data). It is therefore not clear to what extent the *Nasonia* rate applies to *C. fungosa*. In contrast, the nuclear estimates for *Nasonia* are broadly consistent with those obtained for other insects.

The fact that divergence at a single locus can only provide an upper bound of the population divergence time may well explain why mitochondrial dates found in previous studies on other species of European gall parasitoids and their gall wasp hosts (Hayward and Stone 2006) are considerably older than the population divergence estimates for *C. fungosa* obtained here. For instance, mitochondrial divergence between Central European and Iberian clades of the parasitoid *M. stigmatizans* has been estimated at 0.264 mya (Hayward and Stone 2006). Mitochondrial divergence estimates between Central Europe and Iberia for gall wasp host species are still older; for example, 0.383 mya in *Andricus kollari* (Hayward and Stone 2006) and 1.6 mya in *Andricus coriarius* sensu stricto (Challis et al. 2007). Analyses of multilocus datasets are clearly required to provide better estimates of population divergence times in these species. As our results show, the fact that the variance in coalescence time is lower for mitochondrial loci given their smaller N_e may reduce but does not alleviate this problem. This underlines the possibility raised by Nichols (2001) that between-taxon variation in mtDNA-inferred dates of divergence between glacial refugia may well be attributable to coalescent variance rather than taxon-specific differences in postglacial dispersal. Rigorous testing of the hypothesis of taxon-specific variation in divergence times requires broader application of multilocus approaches.

ANCESTRAL N_e AND SAMPLING

The results of the Bayesian analyses show that estimates of $\tau_{C/W}$, or rather the time between the population splits ($\tau_{E/C/W} - \tau_{C/W}$) and the population size during that time, $N_{C/W}$, are confounded. Considering that it is the ratio of the two parameters which determines the chance of coalescence between population splits (Hudson 1983; Saitou and Nei 1986; Yang 2002), this makes intuitive sense and may explain the poor ability to estimate $N_{C/W}$ independently. A large variance in ancestral N_e has also been reported by most earlier multilocus analyses of divergence models (Chen and Li 2001; Yang 2002; Rannala and Yang 2003). In general, explanations for the low power to estimate this parameter fall into two categories: (1) violations of the model assumptions; and (2) limited signal in the data.

Ignoring within-locus recombination and mutational rate heterogeneity, for example, can in principle overestimate ancestral population sizes (Satta et al. 2000; Yang 2002; Wall 2003). However, the few studies that have incorporated these factors suggest that they have little influence on estimates of ancestral N_e (Satta et al. 2000; Yang 2002; Wall 2003). Similarly, the fact that our ML results for the variable mutation model are in agreement with those assuming a single rate despite large differences in relative mutation rates (Table 1) suggests that any impact of mutational heterogeneity between loci is greatly outweighed by coalescence and mutational variance and therefore an unlikely explanation for the low power to estimate $N_{C/W}$.

In general, there are two factors that determine statistical power to infer ancestral parameters; (1) the number of lineages that contribute to the estimate (Fig. 2) and (2) the mutational information available to infer their relationships. Both clearly depend on the timescale of divergence. Relating the estimated population divergence times (scaled by the mean of current population sizes) for *C. fungosa* to the theoretical expectation for the number of surviving lineages, we can ask how much power could potentially be gained by further increasing sample sizes. For example, Figure 2 shows that sampling three instead of a single individual per population roughly doubles the expected number of eastern lineages that survive into the common ancestral population, whereas 16 more individuals are required for a further twofold increase. For the more recent divergence at $\tau_{C/W}$, the increase in the number of surviving lineages from additional samples is of course more substantial (Fig. 2). However, if our analysis was limited by sample size, we would expect to see an improvement in parameter estimation proportional to the increase in the number of surviving lineages when sampling three individuals. The fact that this is not the case (i.e., the variance in the estimates of three of the four model parameters is little affected despite the doubling of surviving lineages) suggests that the power to infer ancestral parameters is largely limited by the mutational variation available rather than the sample size. However, our finding of a markedly higher posterior mean $N_{C/W}$ for the three individual sampling suggests that the estimation of this parameter may indeed be sensitive to the sample size. This makes intuitive sense if we extend the "number of surviving lineage" argument above and consider that only lineages that survive into $N_{C/W}$ and coalesce before they reach $N_{E/C/W}$ contribute to the estimate of $N_{C/W}$. One would therefore expect increased power to estimate this parameter with increasing sample sizes both in *C. fungosa* and in the bird divergence studied by Jennings and Edwards (2005). Thorough investigation of the effect of sampling on statistical power in divergence models both theoretically and using empirical data is required to inform sample designs of future population genetic and phylogeographic studies. In particular, disentangling the effects of mutational limitation and those of sample size (both the number of sampled loci and individuals) would be useful. If mutational information is not limiting, gene tree species tree methods (Degnan and Salter 1995; Maddison and Knowles 2006; Liu and Pearl 2007; Degnan and Rosenberg 2009; Kubatko et al. 2009) should converge to the same answer as the inference methods used here.

Another way to improve power may be to use outgroup information in the likelihood calculation. At present Ne3sML and MCMCcoal rely on clock rooting (Yang 2002), which, given the small number of polymorphic sites in some loci, results in large topological uncertainty. Being able to distinguish between parsimony informative sites and singleton mutations by reference to an outgroup should in principle enhance the power of both approaches.

ASSUMPTIONS AND EXTENSIONS OF THE MODEL

Considering the large confidence intervals in parameter estimates, it is clear that quantitative inference of population history is a data-hungry problem, particularly if divergence is recent. It is therefore questionable how much scope there is to probe more realistic models without increasing the amount of data drastically. In general, inferences of ancestral population parameters are likely to be much more sensitive to violations of the divergence model than they are to violations of the model of sequence evolution. Because there are key population processes omitted from the present analyses that render population history less tree-like, one could argue that the notion of a "population tree" as such is an unrealistic description of phylogeographic history.

First, the model assumes that there is no migration after divergence. Although at least in the host gallwasps, allele frequency data support this assumption (Rokas et al. 2001, 2003; Stone et al. 2001, 2008; Challis et al. 2007), we cannot exclude the possibility of migration after divergence for *C. fungosa*. It would therefore be interesting to relax this assumption and IMa, which uses the algorithm of MCMCcoal, has recently been extended to estimate divergence with migration for more than two populations (Hey 2010). However, modeling migration explicitly in a three-population model introduces six additional parameters. Considering the low divergence between *C. fungosa* populations for our loci, there would appear to be little power in the data to distinguish between a divergence model with a very recent split as inferred here and more complicated models involving both divergence and subsequent gene flow. Clearly, much larger amounts of data are needed to successfully explore such models. An additional problem with analyzing models of migration is that, in contrast to strict divergence models, they are sensitive to unsampled populations (Wilkinson-Herbots 2008; Lohse 2009). With the advent of nextgen sequencing technologies, the volumes of data required to explore divergence with gene flow on such recent timescales should soon be available.

Second, the model assumes constant population sizes between divergence events. Again, allowing for changes in population size opens up a myriad of possible historical scenarios and potentially increases the number of parameters dramatically.

Fortunately however, the *C. fungosa* data allow us to at least exclude drastic demographic events. For instance, under a model of colonization through extreme founder events (without subsequent migration), widespread incongruence between gene trees and population trees would not be expected. Thus the mere presence of all possible gene tree topologies in our data allows us to reject this scenario for *C. fungosa*.

And finally, the model assumes panmixia within populations, which may be unrealistic over short timescales and large geographic areas. Recent theoretical work (Slatkin and Pollack 2008) and simulations (Becquet and Przeworski 2009) have demonstrated that subdivision in ancestral populations can lead to misinference under simple divergence models.

In general, any model-based analysis faces the challenge of choosing models that contain sufficient realism to capture key features in the data while being simple enough to be useful. We have shown that in the case of *C. fungosa* a simple divergence model between three populations can explain the observed genetree incongruence and be used to estimate both the origin and divergence time of refugial populations despite the recency of this history. We hope that this study motivates similar analyses of more realistic models.

TOWARD A MULTILOCUS APPROACH TO COMMUNITY PHYLOGEOGRAPHY

The close ecological dependence of oak gall parasitoids on their hosts and the large number of species involved make this and similar host–parasitoid communities valuable systems in which to study the evolution of ecological interactions (Schönrogge et al. 1995; Hayward and Stone 2005). Unlike most organisms for which similar multilocus analyses have been conducted (Li et al. 1999; Rannala and Yang 2003; Jennings and Edwards 2005), the ecology of chalcidoid parasitoids involves intricate interactions with codistributed species at different trophic levels. Linking the extensive information on species composition and food web structure for these communities (Schönrogge et al. 1995, 1996a; Bailey et al. 2009) with population genetic and phylogeographic inferences at the species level opens up an exciting opportunity to address novel and general questions about coevolution and assembly of communities. For instance, do particular lineages or guilds within trophic levels show earlier longitudinal range expansion than others? And if so, what are the ecological properties of such species? For example, are they generalists rather than specialists, and so less likely to go locally extinct (Hayward and Stone 2006)? Further questions arise when considering multiple trophic levels. How correlated are phylogeographic histories between hosts and parasitoids? Is there a general lag between the arrival of gallwasp (or other herbivore) hosts and associated parasitoids such that herbivores experience periods of enemy-free space (Hayward and Stone 2006)? We are currently working on obtaining multilocus data for codistributed chalcidoid parasitoid species and their gallwasp hosts to address these questions in a quantitative framework. The rarity of many of the species involved (e.g., Schönrogge et al. 1995; Stone et al. 1995; Schönrogge et al., 1996a,b, 1998; Stone et al. 1995) means that we will have to make the most of small sample sizes.

ACKNOWLEDGMENTS

Many thanks to Z. Yang for help with Ne3sML, D. Obbard, M. Blaxter, P. Fuentes-Utrilla, and N. Barton for valuable advice and discussion and J. Nicholls for designing *Cox1* primers. We also thank L. L. Knowles and three anonymous reviewers for insightful comments on earlier versions of this manuscript. We are indebted to G. Melika, M. Tavakoli, and J. Pujade-Villar for providing specimens, F. Sinclair and D. Askew for help with identification and A. Raza for help with cloning. This work was supported by grants from the U.K. Natural Environment Research Council to GNS (NE/E014453/1 and NER/B/504406/1). KL is supported by a BBSRC studentship.

LITERATURE CITED

Andolfatto, P. 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. 18:279–290.

Andolfatto, P., and M. Przeworski. 2000. A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. Genetics 156:257–268.

Askew, R. R. 1961. Some biological notes on the pteromalid (Hym. Chalcidoidea) genera *Caenacis* Förster, *Cecidostiba* Thomson and *Hobbya* Delucchi, with descriptions of two new species. Entomophaga 6: 58–67.

Bailey, R., K. Schönrogge, J. M. Cook, G. Melika, G. Csóka, C. Thúroczy, and G. N. Stone. 2009. Host niches and defensive extended phenotypes structure parasitoid wasp communities. PLoS Biol.7:e1000179. doi:10.1371/journal.pbio.1000179.

Becquet, C., and M. Przeworski. 2009. Learning about modes of speciation from computational approaches. Evolution 63:2547–2562.

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. USA 91:6491–6495.

Campbell, B. C., J. D. Steffen-Campbell, and J. H. Werren. 1993. Phylogeny of the *Nasonia* (Hymenoptera: Pteromalidae) species complex inferred from an internal transcribed spacer (ITS2) and 28s rDNA sequences. Insect Mol. Biol. 2:225–237.

Carstens, B. C., and L. Knowles. 2007. Estimating phylogeny from gene tree probabilities in *Melanoplus* grasshoppers. Syst. Biol. 56:400–411.

Carstens, B. C., H. N. Stoute, and N. M. Reid. 2009. An information-theoretical approach to phylogeography. Mol. Ecol. 18:4270–4282.

Challis, R. J., S. Mutun, J.-L. Nieves-Aldrey, S. Preuss, A. Rokas, A. Aebi, E. Sadeghi, M. Tavakoli, and G. N. Stone. 2007. Longitudinal range expansion and cryptic eastern species in the western Palaearctic oak gallwasp *Andricus coriarius*. Mol. Ecol. 16:2103–2114.

Chen, F.-C., and W.-H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. 68:444–456.

Csóka, G., G. N. Stone, and G. Melika. 2005. The biology, ecology and evolution of gallwasps. Pp. 573–642 *in* C. Raman, W. Schaefer, and T. M. Withers, eds. Biology, ecology and evolution of gall inducing insects. Science Publishers, Enfield, NH.

Culling, M. A., K. Janko, A. Boron, V. P. Vasilév, I. M. Côté, and G. M. Hewitt. 2006. European colonization by the spined loach (*Cobitis taenia*) from Ponto-Caspian refugia based on mitochondrial DNA variation. Mol. Ecol. 15:173–190.

Degnan, J. H., and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Degnan, J. H., and L. A. Salter. 1995. Gene tree distributions under the coalescent process. Evolution 59:24–37.

Din, W., R. Anand, P. Boursot, D. Darviche, B. Dod, E. Jouvin-Marche, A. Orth, G. P. Talwar, P.-A. Cazenave, and F. Bonhomme. 1996. Origin and radiation of the house mouse: clues from nuclear genes. J. Evol. Biol. 9:519–539.

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

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

Folmer, O., M. Black, W. Hoeh, R. Lutz, and R. Vrijenhoek. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit 1 from diverse metazoan invertebrates. Mol. Marine Biol. Biotechnol. 3:294–299.

Galtier, N., F. Depaulis, and N. H. Barton. 2000. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics 155:981–987.

Griffiths, R. C. 1981. Transient distribution of the number of segrating sites in a neutral infinite-sites model with no recombination. J. Appl. Probab. 18:42–51.

Hayward, A., and G. N. Stone. 2005. Oak gall wasp communities: evolution and ecology. Basic Appl. Ecol. 6:435–443.

———. 2006. Comparative phylogeography across two trophic levels: the oak gall wasp *Andricus kollari* and its chalcid parasitoid *Megastigmus stigmatizans*. Mol. Ecol. 15:479–489.

Hedrick, P. W., and J. D. Parker. 2003. Evolutionary genetics and genetic variation of haplodiploids and X-linked genes. Annu. Rev. Ecol. Syst. 28:55–83.

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

Hey, J. 2010. Isolation with migration models for more than two populations. Mol. Biol. Evol. 27:905–920.

Hillis, D. M., and J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Zool. 42:182–192.

Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37:203–217.

Jennings, W. B., and S. V. Edwards. 2005. Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. Evolution 59:2033–2047.

Juste, J., C. Ibáñez, J. Muñoz, D. Trujillo, P. Benda, A. Karatas, and M. Ruedi. 2004. Mitochondrial phylogeography of the long-eared bats (*Plecotus*) in the Mediterranean Palaearctic and Atlantic Islands. Mol. Phylogenet. Evol. 31:1114–1126.

Kass, R. E., and A. E. Raftery. 1995. Bayes factors. J. Am. Stat. Assoc. 90:773–795.

Koch, M. A., C. Kiefer, and D. Ehrlich. 2006. Three times out of Asia Minor: the phylogeography of *Arabis alpina* L. (Brassicaceae). Mol. Ecol. 15:825–839.

Kubatko, L. S., B. C. Carstens, and L. L. Knowles. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971–973.

Lennartsson, T. 2002. Extinction thresholds and disrupted plant-pollinator interactions in fragmented plant populations. Ecology 83:3060–3072.

Li, Y.-J., Y. Satta, and N. Takahata. 1999. Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method. Genes Genet. Syst. 74:117–127.

Liu, L., and D. K. Pearl. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.

Lohse, K. 2009. Can mtDNA barcodes be used to delimit species? A response to Pons et al. (2006). Syst. Biol. 58:439–442.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.

Michaux, J. R., R. Libois, E. Paradis, and M.-G. Filippucci. 2004. Phylogeographic history of the yellow-necked fieldmouse (*Apodemus flavicollis*) in Europe and in the Near and Middle East. Mol. Phylogenet. Evol. 32:788–798.

Nei, M., and S. Kumar. 2000. Molecular evolution and phylogenetics. Oxford Univ. Press, New York.

Nicholls, J. A., S. Preuss, A. Hayward, G. Melika, G. Csóka, J.-L. Nieves-Aldrey, R. R. Askew, M. Tavakoli, K. Schönrogge, and G. N. Stone. 2010. Concordant phylogeography and cryptic speciation in two Western Palaearctic oak gall parasitoid species complexes. Mol. Ecol. 19:592–609.

Nichols, R. 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16:358–364.

Norborg, M. 1998. On the probability of Neanderthal ancestry. Am. J. Hum. Genet. 63:1237–1240.

Oliveira, D. C. S. G., R. Raychoudhury, D. V. Lavrov, and J. H. Werren. 2008. Rapidly evolving mitochondrial genome and directional selection in mitochondrial genes in the parasitic wasp *Nasonia* (Hymenoptera: Pteromalidae). Mol. Biol. Evol. 25:2167–2180.

Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

Patterson, N., D. J. Richter, S. Gnerre, E. S. Lander, and D. Reich. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature 441:1103–1108.

Pauw, A. 2007. Collapse of a pollination web in small conservation areas. Ecology 88:1759–1769.

Rannala, B., and Z. Yang. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.

Raychoudhury, R., L. Baldo, D. C. S. G. Oliveira, J. H. Werren, and M. Wayne. 2009. Modes of acquisition of *Wolbachia*: horizontal transfer, hybrid introgression and codivergence in the *Nasonia* complex. Evolution 63:165–183.

Rokas, A., R. Atkinson, G. 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 87:294–305.

Rokas, A., R. J. Atkinson, L. Webster, G. Csóka, and G. N. Stone. 2003. Out of Anatolia: longitudinal gradients in genetic diversity support an eastern origin for a circum-Mediterranean oak gallwasp *Andricus quercustozae*. Mol. Ecol. 12:2153–2174.

Rosenberg, N. 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61:225–247.

Rozas, J., and R. Rozas. 1995. DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. Comput. Appl. Biosci. 11:621–625.

Saitou, N., and M. Nei. 1986. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. J. Mol. Evol. 24:189–204.

Satta, Y., J. Klein, and N. Takahata. 2000. DNA archives and our nearest relative: the trichotomy problem revisited. Mol. Phylogenet. Evol. 14:259–275.

Schönrogge, K., G. N. Stone, and M. J. Crawley. 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.

———. 1996a. Abundance patterns and species richness of the parasitoids and inquilines of the Alien Gall- Former *Andricus quercuscalicis* (Hymenoptera: Cynipidae). Oikos 77:507–518.

———. 1996b. Alien herbivores and native parasitoids: rapid development of guild structure in an invading gall wasp, *Andricus quercuscalicis* (Hymenoptera: Cynipidae). Ecol. Entomol. 21:71–80.

Schönrogge, K., P. Walker, and M. J. Crawley. 1998. Invaders on the move: parasitism in the galls of four alien gall wasps in Britain (Hymenoptera, Cynipidae). Proc. R. Soc. Lond. B 256:1643–1650.

Sharanowski, B. J., B. Robbertse, J. Walker, S. R. Voss, S. R. Yoder, J. Spatafora, and M. J. Sharkey. 2010. Expressed sequence tags reveal Proctotrupomorpha (minus Chalcidoidea) as sister to Aculeata(Hymenoptera: Insecta). Molecular Phylogenetics and Evolution. *In press*.

Simon, C., F. Frati, A. Beckenbach, B. Crespi, H. Liu, and P. Flook. 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.

Slatkin, M., and J. L. Pollack. 2008. Subdivision in an ancestral species creates asymmetry in gene trees. Mol. Biol. Evol. 25:2241–2246.

Stone, G. N., and P. Sunnucks. 1993. Genetic consequences of an invasion through a patchy environment – the cynipid gallwasp *Andricus quercuscalicis* (Hymenoptera: Cynipidae). Mol. Ecol. 2:251–268.

Stone, G. N., K. Schönrogge, M. J. Crawley, and S. Fraser. 1995. Geographic and between-generation variation in the parasitoid communities associated with an invading gallwasp, *Andricus quercuscalicis* (Hymenoptera: Cynipidae). Oecologia 104:207–217.

Stone, G. N., R. 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 lifecycles. Mol. Ecol. 10:761–778.

Stone, G. N., R. J. Challis, R. J. Atkinson, G. Csóka, A. Hayward, G. Melika, S. Mutun, S. Preuss, A. Rokas, E. Sadeghi, and K. Schönrogge. 2007. The phylogeographical clade trade: tracing the impact of human-mediated dispersal on the colonization of northern Europe by the oak gallwasp *Andricus kollari*. Mol. Ecol. 16:2768–2781.

Stone, G. N., R. W. J. M. Van Der Ham, and J. G. Brewer. 2008. Fossil oak galls preserve ancient multitrophic interactions. Proc. R. Soc. Lond. B 275:2213–2219.

Stone, G. N., A. Hernandez-Lopez, J. A. Nicholls, E. d. Pierro, J. Pujade-Villar, G. Melika, J. M. Cook, and P. Abbot. 2009. Extreme host plant conservatism during at least 20 million years of host plant pursuit by oak gallwasps. Evolution 63:854–869.

Swofford, D. L. 2001. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4.1 Sinauer Associates, Sunderland, MA.

Taberlet, P., L. Fumagalli, A. G. Wust-Saucy, and J. F. Cosson. 1998. Comparative phylogeography and postglacial colonization routes in Europe. Mol. Ecol. 7:453–464.

Tajima, F. 1983. Evolutionary relationships of DNA sequences in finite populations. Genetics 105:437–460.

———. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599–607.

Takahata, N., Y. Satta, and J. Klein. 1995. Divergence time and population size in the lineage leading to modern humans. Theor. Popul. Biol. 48:198–221.

Tamura, K., S. Subramanian, and S. Kumar. 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol. Biol. Evol. 21:36–44.

Tamura, K., J. Dudley, M. Nei, and S. Kumar. 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24:1596–1599.

Tavaré, S. 1984. Line-of-descent and genealogical processes, and their application in population genetic models. Theor. Popul. Biol. 26:1984.

Van Der Ham, R. W. J. M., W. J. Kuijper, M. J. H. Kortselius, J. van der Burgh, G. N. Stone, and J. G. Brewer. 2008. Plant remains from the Kreftenheye formation (Eemian) at Raalte, The Netherlands. Veg. Hist. Archaeobotany 17:127–144.

Velichko, A. A., E. Y. Novenko, V. V. Pisareva, E. M. Zelikson, T. Boettger, and F. W. Junge. 2005. Vegetation and climate changes during the Eemian interglacial in Central and Eastern Europe: comparative analysis of pollen data. Boreas 34:207–219.

Villablanca, F. X., G. K. Roderick, and S. R. Palumbi. 1998. Invasion genetics of the Mediterranean fruit fly: variation in multiple nuclear introns. Mol. Ecol. 7:547–560.

Wakeley, J. 2004. Recent trends in population genetics: More data! More math! Simple models? Heredity 95:397–405.

Wall, J. D. 2003. Estimating ancestral population sizes and divergence times. Genetics 163:395–404.

Weinert, L. A., J. H. Werren, A. Aebi, G. N. Stone, and F. M. Jiggins. 2009. Evolution and diversity of Rickettsia bacteria. BMC Biol. 7:6.

Wilkinson-Herbots, H. M. 2008. The distribution of the coalescence time and the number of pairwise nucleotide differences in the "isolation with migration" model. Theor. Popul. Biol. 73:277–288.

Yang, Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 162:1811–1823.

Zhou, R., K. Zeng, W. Wu, X. Chen, Z. Yang, S. Shi, and C.-I. Wu. 2007. Population genetics of speciation in nonmodel organisms: I. ancestral polymorphism in mangroves. Mol. Biol. Evol. 24:2746–2754.

Associate Editor: L. L. Knowles

Supporting Information

The following supporting information is available for this article:

Figure S1. Prior and posterior distributions of model parameters under the "Out of the East" scenario of population history obtained for the extended sampling (20 loci, 13 sampled for three individuals per population).

Table S1. Primer, sequence, annealing temperature (°C), degeneracy (*De*) for 20 nuclear loci (CG identifier) and *Cox1* used in this study.

Table S2. Rearing information and sampling locations of individuals used for sequencing.

Table S3. Prior and posterior means and 95% confidence intervals for divergence times and ancestral population sizes in Bayesian analyses of extended sampling (20 loci, 13 sampled for three individuals per population) assuming an "Out of the East" population tree topology (E, (C, W)).

Supporting Information may be found in the online version of this article.

Syst. Biol. 58(4):439–442, 2009
Copyright © Society of Systematic Biologists
DOI:10.1093/sysbio/syp039
Advance Access publication on July 15, 2009

Can mtDNA Barcodes Be Used to Delimit Species? A Response to Pons et al. (2006)

KONRAD LOHSE *

Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3JT, UK;
**Correspondence to be sent to: Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3JT, UK;*
E-mail: K.R.Lohse@sms.ed.ac.uk.

The question of how DNA barcodes can and should be used in taxonomy has been debated for some time (Lipscomb et al. 2003; Tautz et al. 2003; Blaxter 2004; Vogler and Monaghan 2007; Wiens 2007). Although few doubt that they are a valuable molecular tool for matching unidentified specimens to described taxa, this has little to do with the question of whether barcodes can be used to delimit species in the first place. The most radical turn in this debate has been the plea for a DNA-based taxonomy (Tautz et al. 2003; Blaxter 2004; Pons et al. 2006; Vogler and Monaghan 2007). Its proponents argue that "the vast majority of sequence variation in nature is partitioned into clearly defined clusters" (Vogler and Monaghan 2007, p. 4), which " [. . .] broadly mirror the species category" (Papadopoulou et al. 2008, p. 1) and could thus serve as basic taxonomic units. Initial attempts to employ this "barcoding gap" have relied on defining cutoff values of sequence divergence a priori (e.g., Blaxter 2004). Considering that the amount of genetic diversity within species can vary by orders of magnitude, it is clear that such an approach is arbitrary at best. Pons et al. (2006) have recently proposed a likelihood method that circumvents this problem by testing for clustering in ultrametric trees. They argue that "these new quantitative approaches can infer the elusive species boundary directly from the transition in branching rate and constitute an exciting possibility to define species from sequence variation [. . .]" (Vogler and Monaghan 2007, p. 6). Given such claims, it is not surprising that this method enjoys increasing popularity, having been applied to a number of mitochondrial DNA (mtDNA) data sets (e.g., Pons et al. 2006; Ahrens et al. 2007; Fontaneto et al. 2007; Papadopoulou et al. 2008).

MODELS AND METHODS

In essence, the Mixed-Yule-Coalescent model (MYC) of Pons et al. (2006) splices together the classical null models of macroevolution and microevolution. Unlike standard models of divergence that view the genealogical process as nested within the species tree, the MYC model assumes a single transition time T at which lineage sorting happens instantaneously and the branching of species clades is replaced by multiple independent coalescences occurring within them (Pons et al. 2006). Assuming T to be a particular node in the tree, Pons et al. (2006) use the internode intervals to find the maximum likelihood solution for T under the MYC model and compare this to the likelihood under a null model of a single neutral coalescent process. Although it has been pointed out that this and similar schemes relying on single locus data cannot deal with lineage sorting and thus necessarily fail to detect recently diverged lineages (Hudson and Coyne 2002; Pons et al. 2006), the potential problems arising from population structure have so far largely been ignored.

In a recent paper, Papadopoulou et al. (2008) have tested the MYC method on genealogies simulated under a symmetric island model, which assumes a population divided into multiple demes or subpopulations that are connected to all other such demes through migration occurring at rate m (Wright 1931). Such population structure tends to produce clustering, very similar to that expected under the MYC model, simply because lineages residing in the same deme coalesce more rapidly on average than those in different demes (Fig. 1). In this setting, the genealogy of a sample may be related to the demic structure in 3 different ways:

1. If gene flow is very low, clusters may correspond well to demes and may thus constitute meaningful taxonomic units in the broadest sense (leaving aside the question how species should be defined).
2. If gene flow is high, clustering may be weak or nonexistent.
3. Clusters may be essentially random, that is, only partially corresponding to demes.

In their simulations, Papadopoulou et al. (2008) assume an extreme sampling scheme where samples are taken from all demes. They find that clustering under the MYC model is only significant when migration rates are extremely low ($Nm < 10^{-3}$) in which case clusters correspond very well to demes (Case 1) (Papadopoulou et al. 2008, figure 3). Once migration is above a certain threshold value, clustering disappears rapidly and is not

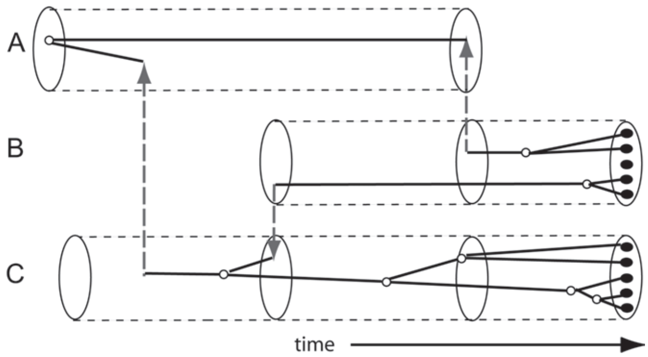

FIGURE 1. Genealogy of a sample taken from two demes *B* and *C* in an island model. Coalescence within demes happens rapidly compared with coalescence of lineages from different demes, which have to be preceded by migration events (dashed arrows). In this case, three "clusters" are produced because a lineage from *B* escapes within-deme coalescence through migration into an unsampled deme *A* and has to wait a long time until it finds itself in the same deme as the remaining lineage.

detected by the MYC method (Case 2). The authors conclude that "the MYC approach appears to be conservative, only detecting the products of population isolation when the levels of gene flow are much lower than those traditionally regarded as sufficient for neutral population divergence" (Papadopoulou et al. 2008, p. 8).

It is worthwhile to recall some basic properties of the coalescent for samples in an island population here. Going backwards in time, lineages have a probability m per generation of escaping coalescence in their local deme. The number of sampled demes over the total number of demes, d/D, is crucial in determining the fate of such escaping lineages. The first migrating lineage has probability d/D of landing in a sampled deme in which it may coalesce. Alternatively, with probability $1 - d/D$, it lands in an unsampled deme and its coalescence has to be preceded by at least one additional migration event. Realizing the pivotal role of the sampling scheme, Wakeley (1998, 2008) has developed an elegant approximation for the coalescent in an island model. If the number of unsampled demes is large, d/D tends to zero and the ancestral process can be split into two phases occurring on different timescales (large D-approximation). Initially, lineages may either coalesce in their local deme or spread out into unsampled demes (scattering phase) (Fig.1). Once every lineage resides in a separate deme, the ancestral process is a neutral coalescent with a rate dependent on the total number of demes, D, their size, N, and m (collecting phase). This separation of timescales and the strong pattern of sequence clusters resulting from it may superficially resemble the two phases in the MYC model. However, there are two important differences. First, there is no branching process in the structured coalescent. Instead, the collecting phase is another, although much slower, neutral coalescent. Thus, theoretically, one could extend the likelihood approach of Pons et al. (2006) to distinguish between the two models. Second and more importantly, clusters in the structured coalescent may be the result of migration events into unsampled demes during the scattering phase and are thus fundamentally random (Case 3). One would therefore expect the sampling scheme to have a major impact on the performance of the MYC method.

To investigate this, I repeated the simulations of Papadopoulou et al. (2008) for varying d/D. Genealogies were simulated in *MS* (Hudson 2002). The effect of the mutational variance on tree reconstruction was ignored, that is, the method was applied directly to simulated genealogies. Likelihoods under both the MYC and a single neutral coalescent were calculated using the genealogy package in *Mathematica* (available from www.biology.ed.ac.uk/research/institutes/ evolution/software/barton/index.html). For each replicate, the two models were compared in a likelihood ratio test and the number of inferred clusters recorded (Papadopoulou et al. 2008). To investigate the region of the parameter space for which the MYC method breaks down, the following sampling scheme was used. Genealogies were simulated for a total of 100 samples taken evenly from 10 demes. Both Nm (0.001, 0.002, 0.004, 0.008, 0.016, 0.032, 0.064, 0.128) and d/D (1, 0.5, 0.2, 0.1, 0.05) were varied and 100 replicates simulated for each parameter combination.

RESULTS

The results agree with those of Papadopoulou et al. (2008) in general, in that the chance of detecting significant clustering under the MYC model declines with increasing migration rates. However, inspection of Figure 2a shows that the robustness of the MYC method depends significantly on the sampling scheme. With decreasing d/D, the chance of detecting significant clustering in the face of high migration rates increases drastically (Fig. 2a). The main effect of migration at the beginning of the coalescent process is then to randomly move lineages into unsampled demes, thereby creating additional clusters and increasing the support of the MYC model. For instance, if only every 20th deme is sampled and $Nm = 0.064$, the chance of detecting significant clustering is still >0.8 (Fig. 2a). The overall excess of clusters detected by the MYC method matches the theoretical prediction for the number of lineages escaping coalescence in their local deme (Fig. 2b) (Wakeley 1998, equation 32). As expected, the fit to the prediction (which neglects the chance of migration to a sampled deme during the scattering phase) increases with decreasing d/D. In the extreme case of complete sampling ($d/D = 1$), the number of inferred clusters is slightly lower than the number of demes (the gray dashed line in Fig. 2b) because escaping lineages necessarily land in sampled demes. The results agree both with intuition gained from the separation-of-timescales arguments as well as earlier simulations (Wakeley 1998) in that d/D does not have to be very small for strong clustering to emerge in the face of migration.

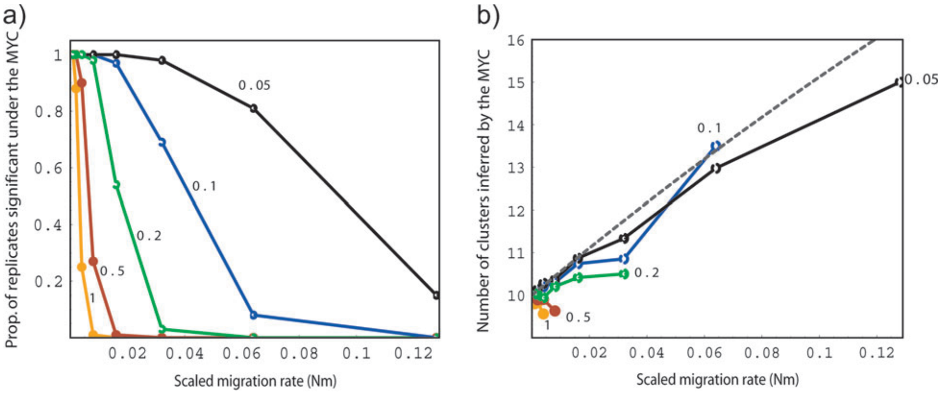

FIGURE 2. a) The proportion of genealogies with significant ($P < 0.05$) clustering under the MYC model plotted against the scaled migration rate. Different colors correspond to different sampling schemes, that is, proportions of sampled demes, d/D: Orange = 1, red = 0.5, green = 0.2, blue = 0.1, and black = 0.05. In each case, genealogies were simulated for a total of 100 sequences. 10 samples were taken from each of 10 demes. Each point is based on 100 replicates. The orange line corresponds to the complete sampling scheme assumed by Papadopoulou et al. (2008). b) The average number of clusters inferred by the MYC method for different sampling schemes. The upper dotted line is the theoretical prediction for the number of lineages at the end of the scattering phase in the limit when d/D tends to zero.

DISCUSSION

Given the large effect of the sampling scheme, how realistic is the assumption of incomplete sampling? First, geographic sampling is hardly ever complete in practice. This is true in particular for most barcoding data which are rarely collected with a particular sampling scheme in mind (but see Pons et al. 2006; Papadopoulou et al. 2008), and it has been argued before that the "barcoding gap" may in part result from incomplete spatial sampling (Moritz and Cicero 2004). Second, there are biological reasons why the kind of completeness required for the MYC method to be reliable may be impossible to achieve in practice. What governs the formation of clusters is not the population structure at the time of sampling but rather the sum of population structures that have affected the ancestral process of the sample in the past. The symmetric island model considered here is the simplest possible model of structure. In more realistic metapopulation models, demes are transient so that lineages may spend the majority of their history in demes that have subsequently gone extinct and can therefore not be sampled. Thus, increasing the geographic scale of sampling does not necessarily get around the problem. Considering that separation of timescales have been applied to a variety of models of structures (Wakeley 2004; Wilkins 2004; Matsen and Wakeley 2006), the main result is likely to hold in general. For instance, an analogous argument can be made for samples from a population in a continuous 2-dimensional habitat (Wilkins 2004). In this model, there is no discrete underlying structure at all so any observed clustering must be spurious. However, if a sample is taken from a set of random locations, one would expect a pattern similar to that observed in the island model. At the beginning, lineages either coalesce quickly in their neighborhood or escape by chance, in which case coalescence takes a much longer time on average. Again, the resulting clusters would only partly correspond to sampling locations with additional clusters being created by migration during the scattering phase (see Wilkins 2004, figure 4).

In conclusion, the method of Pons et al. (2006) delimits essentially random clusters when applied to samples from a single island model population if d/D is low. Similar behavior is expected under any model of geographic structure as long as there is a considerable fraction of unsampled space and a separation-of-timescales exists. This is particularly worrisome considering the envisioned application of the MYC method to high-throughput mtDNA profiles (Pons et al. 2006). Such mass samples are likely to contain both individuals from truly isolated clades or species and structured populations connected by gene flow, making it even harder to distinguish between the two types of clusters.

Taken together, the results cast serious doubts on the usefulness of mtDNA barcodes as a scaffold for an automated DNA taxonomy (Pons et al. 2006). The stochastic nature of both migration and lineage sorting requires multilocus data, exhaustive geographic sampling, and realistic models, which can deal with the expected incongruence between gene genealogies to delimit meaningful taxonomic units from sequence data (Edwards 2009). However, this remains a difficult task even for a very modest number of taxa (e.g., Knowles and Carstens 2007) and is incompatible with the notion of a DNA taxonomy based on a single locus. Given the ubiquity of population structure in nature, the number of potentially detectable clusters in mitochondrial barcode data is likely to vastly exceed that of meaningful taxonomic units.

FUNDING

K.L. is funded by the Biotechnology and Biological Sciences Research Council.

Acknowledgments

Many thanks to Nick Barton for much helpful advice and encouragement. Thanks to Jerome Kelleher for modifying *MS*; Marianne Elias, Brian O'Meara, and one anonymous reviewer for useful comments; and Anna Papadopoulou, Alfried Vogler, and Tim Barraclough for taking the time to discuss this problem.

References

Ahrens D., Monaghan M.T., Vogler A.P. 2007. DNA-based taxonomy for associating adults and larvae in multi-species assemblages of chafers (Coleoptera: Scarabaeidae). Mol. Phylogenet. Evol. 44: 436–449.

Blaxter M. 2004. The promise of a DNA taxonomy. Phil. Trans. Roy. Soc. B 29:669–679.

Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63:1–19.

Fontaneto D., Herniou E.A., Boschetti C., Caprioli M., Melone G., Ricci C., Barraclough T.G. 2007. Independently evolving species in asexual bdelloid rotifers. PLoS Biol. 5:914–921.

Hudson R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 18:337–338.

Hudson R.R., Coyne J.A . 2002. Mathematical consequences of the genalogical species concept. Evolution. 56:1557–1565.

Knowles L.L., Carstens B. 2007. Delimiting species without monophyletic gene trees. Syst. Biol. 56:887–896.

Lipscomb D., Platnick N., Wheeler Q. 2003. The intellectual content of taxonomy: a comment on DNA taxonomy. Trends Ecol. Evol. 18: 65–66.

Matsen F.A., Wakeley J. 2006. Convergence to the island-model coalescent process in populations with restricted migration. Genetics. 172:701–708.

Moritz C., Cicero C. 2004. DNA barcoding: promise and pitfalls. PLoS Biol. 2:1529–1531.

Papadopoulou A., Bergsten J., Fujisawa T., Monaghan M.T., Barraclough T.G., Vogler A.P. 2008. Speciation and DNA barcodes: testing the effects of dispersal on the formation of discrete sequence clusters. Phil. Trans. Roy. Soc. B 363:2987–2996.

Pons J., Barraclough T., Gomez-Zurita J., Cardoso A., Duran D., Hazell S., Kamoun S., Sumlin W., Vogler A. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55:595–610.

Tautz D., Arctander P., Minelli A., Thomas R.H., Vogler A.P. 2003. A plea for DNA taxonomy. Trends Ecol. Evol. 18:70–74.

Vogler A.P., Monaghan M.T. 2007. Recent advances in DNA taxonomy. J. Zool. Syst. Evol. Res. 45:1–10.

Wakeley J. 1998. Segregating sites in Wright's Island Model. Theor. Pop. Biol. 53:166–174.

Wakeley J. 2004. Metapopulation models for historical inference. Mol. Ecol. 13:865–875.

Wakeley J. 2008. Coalescent theory an introduction. Greenwood Village (CO): Roberts and Company.

Wiens J. 2007. Species delimitation: new approaches for discovering diversity. Syst. Biol. 56:875–879.

Wilkins J.F. 2004. A separation-of-timescales approach to the coalescent in a continuous population. Genetics. 168:2227–2244.

Wright S. 1931. Evolution in Mendelian populations. Genetics. 16: 97–159.

Received 8 October 2008; reviews returned 25 February 2009; accepted 9 June 2009
Associate Editor: Marshal Hedin