Effects of Environmental Change on the Genetic Diversity and Distribution of *Phlebotomus ariasi*, a Vector of Visceral Leishmaniasis in Southwest Europe

by

Shazia Sophie Mahamdallie

Thesis submitted for the degree of Doctor of Philosophy to the London School of Hygiene and Tropical Medicine

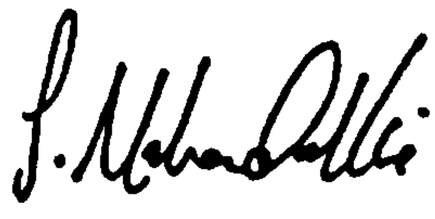
2010

University of London

Department of Infectious and Tropical Diseases
London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
and
Department of Entomology
Natural History Museum
Cromwell Road
London SW7 5BD

DECLARATION BY CANDIDATE

I have read and understood the School's definition of plagiarism and cheating given in the Research Degrees Handbook. I declare that this thesis is my own work, and that I have acknowledged all results and quotations from the published or unpublished work of other people. I have read and understood the School's definition and policy on the use of third parties (either paid or unpaid) who have contributed to the preparation of this thesis by providing copy editing and, or, proof reading services. I declare that no changes to the intellectual content or substance of this thesis were made as a result of this advice, and, that I have fully acknowledged all such contributions.

Signed: 

Date: August 2010

Full name Shazia Sophie Mahamdallie

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

This thesis is dedicated to my Mother, and the rest of the family

ACKNOWLEDGEMENTS

I am indebted to many who have supported and guided me before I embarked on this doctorate, through to its completion. I express unquantifiable thanks to my principal supervisor, Paul Ready, for discovering and encouraging the scientist within me - a full hand salute to you. In memory of my supervisor Clive Davies, I will always adopt his approach throughout every research project: "...so, what is the question?" And thanks to Jonathan Cox, my third supervisor, for stepping forward to offer support and reasoned perspective. Weeks in the field in search of the best apple slippers, ten-star toilets and the odd sandfly, would not have been as enjoyable and successful without the help of Bernard Pesson - thank you for everything.

For adding a little extra diversity to this thesis, thanks to colleagues who donated specimens: Dr Ana Aransay, Dr Samia Boussaa, Prof. Robert Farkas, Prof. Montse Gállego, Dr Parviz Parvizi and Dr Carlos Pires. At the Natural History Museum, London, a huge thank you to Julia Llewellyn-Hughes and Claire Griffin for letting me loose in their DNA sequencing facility and always meeting my last minute demands. To those poor souls in the Molecular Sytematics Laboratory who were exposed to my O.C.D. tendencies, yet still provided friendly smiles and help: Andrea Hall, Alex Martin, April Wardhana, Alison Cownie, Fenella Halstead and Jo Ingle. To those lovers of beetles and mosquitoes, thank you for sharing long laboratory hours, lunches and knowledge. Thank you to Anna Papadopoulou and Stephen Ansell for providing invaluable help and stimulating discussions.

As a dedication is simply not enough, thanks to my mother, Yvonne Mahamdallie, whose daily support made my every goal achievable. She has guided me through every decision, and without her, this simply would never have been possible. For supplying sustenance and laughter Leela, Tara, Yasmin and Zahra, the four pillars that support me. Sharif here is a different kind of school report to read! Kamal, Salim, Taz, Hassan, Zelica and Karim, I now promise to get a proper job. And the last Mahamdallie-Pritchard, Hugh, thank you for showing me the fighting spirit. My great friends Claire, Sarah, Saskia and Shamila the best listening ears, company and motivators one could wish for. Finally, I bestow the medal of patience upon my partner, for all your encouragement and understanding - thank you.

This research was funded by EU grant GOCE-2003-010284 EDEN and is catalogued as publication EDEN0235 (http://www.eden-fp6project.net/). It does not necessarily reflect the views of the European Commission.

ABSTRACT

Leishmania infantum is the causative agent of zoonotic visceral leishmaniasis (ZVL) in the Mediterranean region, with the domestic dog as the main reservoir host. *Phlebotomus (Larroussius) ariasi* is the principal vector in cooler, forested ecotopes in southwest Europe, which suggests that it might be subject to environmental and geographical isolation. However, the population genetics of *P. ariasi* had been little studied before this thesis, which investigated how the population differentiation of this vector might affect its ability to spread northwards, or persist in the Mediterranean region, in response to climate and habitat change. Thirty-six spatio-temporal populations of *P. ariasi* were molecularly characterized across its range, predominantly from southwest France but including geographical outgroups from Spain, Portugal and North Africa. Phylogenetic and population genetic assessments were made based on five DNA sequences: mitochondrial cytochrome b, nuclear elongation factor-1α and apyrase, plus two anonymous nuclear loci, AAm20 and AAm24. The results demonstrated the absence of cryptic sibling species of *P. ariasi* and the selective neutrality of each locus. Mitochondrial DNA revealed a historical phylogeographic structure, which was consistent with Pleistocene climate change driving multiple haplogroup divergences within glacial refuges and phalanx-like population expansions in interglacial periods. Nuclear loci mostly showed isolation by distance, but some supported restricted gene flow between the Pyrenees and the Massif Central, France, as indicated by cytochrome b. A glacial refuge may have existed north of the Pyrenees. The genetic diversity observed in the northeast Pyrenees, France, permitted an assessment of the effects of broadleaf forest fragmentation on the differentiation of *P. ariasi*. No conclusive evidence was found to support contemporary genetic sub-structuring or impoverishment associated with a recent increase in forest fragmentation. The salivary peptide apyrase revealed a geographical pattern of polymorphism consistent with the other selectively neutral loci. A range of selection tests indicated that apyrase was not evolving under positive directional or balancing selection and, therefore, a genetic arms race with the mammalian host and/or *Leishmania* parasite was not supported. The approach taken provides a proof of principle for helping to assess apyrase and other salivary peptides as vaccine candidates against leishmaniasis.

CHAPTER 1

General introduction

1.1 Overview

In the western Mediterranean *Phlebotomus (Larroussius) ariasi* Tonnoir, 1921 (Diptera: Psychodidae) is one of the two main incriminated sandfly vectors of *Leishmania infantum* Nicolle, 1908 (Kinetoplastida: Trypanosomatidae), the causative agent of both cutaneous leishmaniasis (CL) and visceral leishmaniasis (VL) in humans and canine reservoirs. Disease distribution is geographically limited, determined by the combined environmental requirements of all three components of the transmission cycle - parasite, vector and mammalian host. This thesis resulted from a co-ordinated European research project (Emerging Diseases in a changing European eNvironment), whose aims were in part to use molecular genetics to resolve vector population structure to assist epidemiological risk modelling of the leishmaniases. Between 2005 and 2008 populations of sandflies were captured in southwest France using a systematic sampling strategy. Investigations specific to this thesis were to examine the monophyly of *P. ariasi* and the effects of past climate change on its population genetic structure (Chapter 2), to determine the natural polymorphism and the processes of molecular evolution on the salivary peptide apyrase of *Phlebotomus* including *P. ariasi* (Chapter 3), and to identify the local landscape features that affect the distribution of *P. ariasi* (Chapter 4). This introduction summarizes the literature on the components of the disease transmission cycle, provides a background to the techniques implemented and gives the thesis' aims.

1.2 The leishmaniases and their sandfly vectors

1.2.1 Sandflies in relation to the distribution of the leishmaniases

Approximately 1200 species of phlebotomine are known, of which the females of ca. 30 *Phlebotomus* species (predominantly belonging to four subgenera) are suspected or proven vectors of *Leishmania* species causing anthropozoonotic transmission in the Old World (Killick-Kendrick, 1990; Lane and Crosskey, 1993). Transmission among mammalian reservoirs and hosts typically occurs when an infective female takes a blood meal in order to acquire the necessary proteins for the development of her eggs (Lane and Crosskey, 1993). Vectorial competence of the sandfly initially requires its ability to support the development of the ingested parasite

in the gut (Bates, 2007) and its capacity to transmit relies on the close proximity of both a sustained vector and reservoir host population. Directly limiting the distribution of the various leishmaniases is their co-association, co-evolution or co-speciation with specific *Phlebotomus* (reviewed Ready, 2000). It follows that an understanding of *Phlebotomus* speciation (i.e. presence of a species complex) will have implications for both disease distribution and its targeted control or intervention.

The leishmaniases are a globally widespread group of diseases, whose first clinical symptoms were given in 1756 and the causative parasite formally identified and named *Leishmania* Ross, 1903 (Kinetoplastida: Trypanosomatidae). Endemic in 88 countries on five continents, leishmaniasis affects 12 million people with 1.5 to 2 million new cases arising annually and a further 350 million people at risk (WHO, 2010a). The leishmaniases of humans can be classified into four main forms that have a range of clinical descriptions from localized cutaneous manifestations (cutaneous leishmaniasis, CL), disseminated and chronic skin lesions (diffuse cutaneous leishmaniasis, DCL), the destruction of mucous membranes (mucocutaneous leishmaniasis, MCL) to a visceralizing disease with 100% mortality if left untreated (Kala azar or visceral leishmaniasis, VL). Global VL mortality is estimated at 59,000 per annum, (WHO, 2002) but as leishmaniasis is only notifiable in 32 of the 88 countries affected, actual mortality is likely to be higher. Relevant to this thesis, VL is caused by the zoonotic *Leishmania infantum* Nicolle, 1908, distributed in most of the Mediterranean Basin both in Europe and north Africa, through to Iran and China. The same species (*L. i. chagasi*) is a greater health problem in the Neotropics. VL also can take an anthroponotic form, when caused by *L. donovani sensu lato* in northeast Africa and the Indian subcontinent. The dependence of *L. infantum* on a restricted number of related sandflies leads to a strong association between leishmaniasis and environmental features (Ashford, 2000). For example, proven Old World sandfly vectors transmitting *L. infantum* are all classified in the subgenus *Larroussius* (Killick-Kendrick, 1990), which is mostly restricted to Mediterranean bioclimates, including the sub-humid, humid, semi-arid and arid (Ready, 2008).

1.2.2 Choice of transmission cycle: *L. infantum* and *P. ariasi* in southwest Europe

Leishmaniasis is caused by a wider range of parasites than many other parasitic diseases of man, with at least 23 species implicated globally (Killick-Kendrick, 1990). In Europe, two species or species complexes are found: *L. infantum* and *L. tropica*

(reviewed Ready, 2008). The causative agent of VL in Europe is solely *L. infantum*, and it is transmitted by five *Larroussius* species whose distributions are sympatric or parapatric: in western Europe, *P. (L.) ariasi* and *P. (L). perniciosus* Newstead, 1911; in Italy, additionally *P. (L). perfiliewi* Parrot, 1930 and *P. (L). neglectus* Tonnoir 1921; and in eastern Europe, *P. ariasi* is replaced by *P. neglectus* and *P. perniciosus* by *P. (L). tobbi* Adler, Theodor and Lourie, 1930.

P. ariasi and *P. perniciosus* are sympatric vectors in southern France, Portugal, Spain, Italy, Morocco and Tunisia (Esseghir *et al.*, 2000; Rioux and Golvan, 1969). The current study region of southern France comprised the Massif Central (which in this thesis includes also the Lot and Rhone valleys) and southwestern Mediterranean France bordered to the west by the central Pyrenees. This region lies at the edge of leishmaniasis endemicity (Rispail *et al.*, 2002; Trotz-Williams and Trees, 2003) and vector distribution (Rioux and Golvan, 1969; P.D. Ready and B. Pesson, unpublished), so poses a significant risk for range expansion of both vector and disease with environmental change, including climate warming.

The ecological distributions of these two vectorial sandflies are overlapping, which leads to a disjunct geographical distribution in some regions. The distribution of *P. ariasi* is predominantly related to the humid and sub-humid bioclimatic belts (Sauvage and Brignon, 1963), with Mediterranean oaks (*Quercus ilex, Q. pubescens* and *Q. coccifera*) as indicators (Emberger, 1936; 1939; Rioux *et al.*, 1984; Rispail *et al.*, 2002), whereas *P. perniciosus* favours semi-arid and arid areas (Rioux *et al.*, 1984). Grandes *et al.* (1988) reviewed the reports and concluded that *P. perniciosus* predominates over *P. ariasi* in Marseilles, Italy and semi-arid Mediterranean Spain, whereas the reverse is true in the cool French Cévennes (Rioux *et al.*, 1967). The latter is also true for the cooler bioclimatic zones in Spain (Aransay *et al.*, 2004). The species' bioclimatic preferences can lead to altitudinal separation in a given region. With the average air temperature decreasing by 0.6°C with each 100 m increase in altitude (International Standard Atmosphere; ISO 2533:1975), *P. ariasi* has been collected at altitudes > 750 m.a.s.l. in the cooler, more humid supra-Mediterranean bioclimatic zone of Spain, shifting upwards at more southerly latitudes (Aransay *et al.*, 2004). Similar patterns are reported from France, where in the cooler and more temperate Cévennes mountains (southern Massif Central) and Pyrenees, *P. perniciosus* occurs in comparatively low numbers at altitudes up to 600 m, whereas *P. ariasi* peaks in abundance at 300-500 m.a.s.l. (Rioux and Golvan, 1969). In the eastern Pyrenees,

France *P. ariasi* has been recorded from 120 to 1415 m.a.s.l., whereas the *P. perniciosus* was limited to below 600 m.a.s.l. (P.D. Ready, S.S. Mahamdallie and B. Pesson, unpublished).

In France and Iberia the two sandflies' habitats differ, with *P. perniciosus* often being found peri-domestically, whereas *P. ariasi* is more frequently associated with hillsides and forests (Rioux and Golvan, 1969; P.D. Ready, S.S. Mahamdallie and B. Pesson, unpublished). The distribution of *P. perniciosus* has been described based on phylogeographic and population-based inferences, in addition to knowledge of its ecology – requirement for Mediterranean-like environment for adult activity and overwinter survival of diapausing larvae (Rioux and Golvan, 1969; Ready and Croset, 1980). This fly's distribution has been limited by historical changes in climate, with Pleistocene refugia during the glacials in southern Spain and Italy (Esseghir *et al.*, 2000; Pesson *et al.*, 2004), and post-glacial secondary contact in southern France (Perrotey *et al.*, 2005). However, population studies of *P. perniciosus* in France at its northern limit have shown predominantly low levels of genetic differentiation (P.D. Ready, S.S. Mahamdallie and B. Pesson, unpublished) and putative genetic introgression, both hindering further population genetic inferences (Perrotey *et al.*, 2005).

Several biological characteristics of *P. ariasi* make it a suitable and important choice for population studies. Firstly, occurrence in relatively cool environments not only suggests greater genetic diversity, but also identifies it as the vector most likely to expand northwards with climate warming. Indeed the few studies which have been conducted support a greater genetic diversity of this species compared to *P. perniciosus* in the study region (B. Pesson, unpublished data). Secondly, the virtual absence of *P. ariasi* in continental northern Italy (*P. perniciosus* abundant) (Maroli *et al.*, 2008) makes this re-colonization route into France unlikely, limiting the possibility of bias by the genetic introgression of diverged lineages that may mask the detection of historical events. Thirdly, the many eco-epidemiological reports from France make *P. ariasi* a good vector to study. In the 1960s to 1980s, J.-A. Rioux, R. Killick-Kendrick and colleagues carried out an extensive body of work on the biology and ecology of *P. ariasi* and the epidemiology of leishmaniasis in southern France, concluding that specifically in the Cévennes, around the northern limit of *Leishmania*, this species was the predominant vector of human leishmaniasis (HumL) and canine leishmaniasis (CanL). This result was matched by Grandes *et al.* (1988), outside the study region in Salamanca, Spain, who demonstrated a linear relationship between *P. ariasi* density and

CanL prevalence and, therefore, HumL prevalence. As *P. ariasi* and disease distribution are associated, it is therefore relevant to understand the extrinsic variables that shape this sandfly's historical and contemporary population structure, in order to inform predictions of disease emergence and spread.

The final component of this zoonotic transmission cycle, the reservoir host, poses both a significant public and veterinary problem. In most parts of the range of *L. infantum*, including southwest Europe, domestic dogs (*Canis lupus familiaris* Linnaeus, 1758) are considered the main reservoirs of infection for zoonotic VL (Ashford, 1996), with both symptomatic and asymptomatic dogs as sources for transmittable parasites (Molina *et al.*, 1994). CanL is of significant epidemiological concern in southwest Europe from two perspectives. Firstly, like most human leishmaniases, those caused by *L. infantum* are actively zoonotic or have recent zoonotic origins (Ashford, 2000) and HumL is endemic when suitable reservoir hosts and vectors co-occur. Secondly, CanL is of major veterinary concern *per se*, endemic in all countries bordering the Mediterranean Sea and Portugal (reviewed Dereure *et al.*, 1999). In the Mediterranean Basin prevalence rates are as high as 80% (reviewed Trotz-Williams and Trees, 2003), with high prevalances in both Spain (up to 34%) and France (4-20%; 5000 clinical canine cases) (Dujardin *et al.*, 2008), where in the latter all rural dogs will become infected at some stage in their lifetime.

1.2.3 Leishmaniasis as a (re-)emerging zoonosis in western Europe and disease modelling in relation to environmental change

Classically circum-Mediterranean, VL took an 'infantile' form that was a public health problem until the 1950s-1960s, when it declined following the introduction of DDT for malaria control, and improved nutritional status and housing of children in western Europe post-World War II (Ashford, 2000). However, as the disease has remained a serious problem in dogs, actual transmission was not greatly affected. Dujardin *et al.* (2008) estimated approximately 700 new human cases per year are reported in southern Europe (3,950 if Turkey is included), confirming the general recognition of leishmaniasis (re-)emergence (e.g. Ashford, 2000; Dujardin *et al.*, 2008; Ready, 2008). Several conditions, some new while others previously known, are responsible for this (re)-emergence (reviewed in Desjeux, 2001; Dujardin, 2006). Two of the major risk factors proposed are relevant to European VL: (i) host immune status; and (ii) anthropogenic or natural environmental changes. Addressing the first,

Leishmania-HIV co-infections in southern Europe account for 70% of adult VL cases where symptoms can be fatal (WHO, 2010b), which in addition to transmission by the sharing of syringes among infected intravenous drug users (Alvar and Jimenez, 1994) make these populations most at risk. Patients with *Leishmania*-HIV co-infections can act as reservoirs for VL and syringes potentially by-pass the need for transmission by the sandfly. These factors have changed the traditional epidemiology of the disease in Europe, where conditions for epidemics are now favoured in urban areas (WHO, 2010b).

The combined effects of the alteration of global climate and other environmental changes (i.e. local land use) disrupt the natural ecosystem and can increase the risk of disease emergence (Patz *et al.*, 2000), and/or expand the range of arthropod-borne diseases (Dujardin *et al.*, 2008). Indeed in Europe, autochthonous cases of leishmaniasis are no longer limited to the Mediterranean region, with northerly reports of VL in Germany (Naucke and Schmitt, 2004: Ready, 2008), and focal endemics in continental Italy (Maroli *et al.*, 2008).

Climate change is popularly discussed as a risk factor for the spread of vector-borne diseases in general, which has spurred the development of models to predict disease emergence and spread. It is accepted that to correctly predict future patterns of disease emergence and spread, temperature should be included as a single variable in a multivariate climate model, and this estimate in turn must be incorporated into a more comprehensive model of the transmission cycle as a whole. Moreover, caution should be taken when extrapolating results outside of the region or the particular transmission cycle studied (e.g. Dye and Reiter, 2000; Rogers and Randolph, 2000; Ready, 2008). Explanatory and predictive risk modelling of zoonotic diseases generally use the same set of tools, often involving variants of linear or logistic regression and discriminant analysis, frequently with a Geographical Information System (GIS) where multiple environmental information (data layers) can be combined for a single location. These data layers can integrate field, remote sensed and molecular data, which in concert with statistical tools hold great promise to understand disease epidemiology. However, such models are not without their drawbacks. Detailed knowledge of the ecology and biology of the transmission cycle are required, but are not always available or sufficient.

The environmental risk factors implicated for the leishmaniases include: temperature, average rainfall, soil type, re-/de-forestation and other vegetation changes, immigration and travel of humans and reservoir hosts, vector and host population

density changes and shifts, urbanization and malnutrition (reviewed in Ashford, 2000; Patz *et al.*, 2000; Ready, 2008). In Europe, HumL and CanL occur in regions where vector population and vegetation type distributions have been shown to be correlated as outlined above (e.g. oak or broadleaf forest), and it is these preferred habitats which are associated with the greatest risk of transmission of VL (Rioux *et al.*, 1980). As new foci of CanL have been reported over the last 20 years in southern France (Dereure *et al.*, 1999), spatial modelling is likely to prove a valuable tool for mapping the risk of emergence of *Phlebotomus* species and leishmaniasis.

1.3 Chosen molecular markers

1.3.1 Characteristics of molecular markers to investigate speciation and neutral population structure

The development of the polymerase chain reaction (PCR) to amplify targeted DNA fragments and associated analytical tools have revealed information at all levels of biotic hierarchy, from population to phylum (Avise, 1994). Molecular markers are mainly classified into two types, mitochondrial (from organelles present in the cell cytoplasm) and nuclear loci (contained in a cell nucleus), and are recommended to be used in concert to avoid evolutionary inferences based on a single genealogy (Ballard and Whitlock, 2004). A plethora of genetic techniques are available to infer the evolution of a genetic system, where the appropriate marker depends on the given question to be studied and ideally the availability of a comparative database for the same gene region if outgroups of species or populations are required.

The mitochondrial (mtDNA) genome has been characterized for about 20 years. Kocher *et al.* (1989) published the first highly conserved primers for PCR. MtDNA has become a mainstay of phylogenetics and intra-specific genealogies, specifically phylogeography, due to its technical convenience. This includes: (i) mtDNA is present in high copy number in most eukaryotes, making it relatively easy to isolate in the laboratory; (ii) mtDNA is usually maternally inherited which, along with its frequent lack of recombination, allows the reconstruction on a single genealogy, and this is certainly true for many Diptera; (iii) mtDNA can have an evolutionary rate up to 10-fold higher than a single copy nuclear genome (see review by Ballard and Whitlock, 2004), allowing for shallower phylogenetic inferences than some nuclear genomes.

Lack of recombination can be a drawback of mtDNA through its inheritance as a single linkage group, where independent population history estimates cannot be gained

from other mtDNA gene regions in the same sample (Moore, 1995). The biology of mtDNA and nuclear DNA (nDNA) differ in a number of innate characteristics: their ploidy (haploid vs. diploid), mode of inheritance (maternal vs. biparental), mutation rates (mtDNA > nDNA), number of introns, number of copies (mtDNA present in hundreds to thousands of copies, whereas many nDNAs are single-copy genes). The effective population size (N_e) is one of the main differences between mtDNA and nDNA. All things being equal, N_e causes mtDNA to fix mutations through random genetic drift four times faster than nDNA (Ballard and Whitlock, 2004). These characteristics allow mtDNA to resolve shallower phylogenies or population processes, whereas nDNA tends to be too invariant (Sunnucks, 2000), often being more usefully applied to construct deeper phylogenies (Cho *et al.*, 1995).

Nuclear genomes are popularly characterized in two ways, directly sequenced or genotyped, with analyses of mutations in the former providing inferences at the longest time-scale in phylogeography or phylogenetic studies as discussed above (Avise, 2000; Moore, 1995). Genotype markers often provide information on allele or genotype frequencies only, which are appropriate in aiding our understanding of population processes. These markers amongst others include single-locus microsatellites which, although their isolation can be laborious, once developed can provide sensitive, connectible data from individual identification through to shallow phylogeny (Sunnucks, 2000). These markers can reflect genetic variation at two levels. Firstly, as the process of sexual reproduction reorganizes genotypes each generation, the concatenation of several independent genotype datasets allows individual level or within-population inferences at this shortest time-scale (current generation), e.g. parentage, individual relatedness and migration (Queller and Goodnight, 1989; Rannala and Mountain, 1997). Whereas, the application of single-locus allele frequency data can be used to infer relatively long timescale population processes such as population size changes, gene flow and genetic drift (Bossart and Prowell, 1998).

1.3.2 Molecular markers chosen to investigate the effects of environmental change on *P. ariasi* distribution

The metazoan mtDNAs range from ca. 11.5 kilobases (kb) to 32 kb, are double-stranded and most frequently consist of 37 genes (Gissi *et al.*, 2008): 24 encode the translational machinery of the mtDNA itself, the additional 13 encode subunits for the electron transport chain that metabolises substrates for ATP production. Part of this

machinery includes the one gene of mitochondrial cytochrome b (cyt b). Cyt b is one of the least conserved of the protein coding subunits, second to the A-T rich control region, making it a useful molecular tool in the systematics of closely related genera rather than deep divergences (Simmons and Weller, 2001). Direct sequencing of cyt b has proved useful in sandfly systematics. Successful amplification of the 3′ terminus (449 to ca. 720 bp,) across the genus *Phlebotomus*, has utilized sequences for species level comparisons, further to its use in discerning species complexes and demographic histories associated with historical climate change of individual sandfly species (Esseghir *et al.*, 1997; 2000; Pesson *et al.*, 2004). However, comparative cyt b direct sequencing has been unsuccessful at resolving population structures in *P. perniciosus* of Spain (Aransay *et al.*, 2001; 2003) and limited for *P. papatasi* of Iran (Parvizi *et al.*, 2003) and across the Mediterranean (Hamarsheh *et al.*, 2007).

Heteroplasmy (carrying more than one mtDNA haplotype, often in somatic tissues only) and recombination of the mtDNA genome can pose problems for molecular inferences i.e. assessment of multiple histories instead of a single genealogy. For recombination to be important in terms of changing the patterns of descent it is necessary that some individuals be heteroplasmic (Ballard and Whitlock, 2004). The studies at the *Phlebotomus* cyt b locus described above have not reported evidence of heteroplasmy in their direct sequences. As cyt b has an appropriate biology to study sandfly molecular ecology, phylogeography, and phylogenetics, it was used as a marker in this thesis for such purposes.

Elongation factor-1α (EF-1α) is a conserved nuclear protein coding gene (ca. 1,300 bp) involved in the GTP-dependent binding of charged tRNAs to the acceptor sites of the ribosome during translation of mRNA to proteins (Hovemann *et al.*, 1988). EF-1α is widely applicable in insect systematics to resolve both deep and derived phylogenetic relationships (e.g. Cho *et al.*, 1995; Esseghir *et al.*, 2000; Kandul *et al.*, 2004). Attributes include an absence of internal repeats, highly conserved amino acid sequence, and a moderate synonymous substitution rate. One drawback in characterizing EF-1α is the existence of two paralogous copies in a diverse array of insects (including flies and other holometabolous insects), which may confound phylogenetic studies if paralogous copies are confused (Danforth and Ji, 1998).

Esseghir *et al.* (2000) demonstrated that EF-1α was not conserved across *Phlebotomus*. Their primers successfully amplified and re-constructed its gene tree for species of the subgenus *Larroussius* (in which *P. ariasi* is classified), but not for species

of the subgenera *Phlebotomus* and *Paraphlebotomus*. They concluded that their primers targeted a single-copy sequence from an orthologous locus in *Larroussius*, evidenced by intronless PCR products, occurrence of only synonymous substitutions, and homogeneity of nucleotide base composition of 10 *Larroussius* species. In this thesis, I shall use the same primers as Esseghir *et al.* (2000) to target a fragment of the EF-1α from *P. ariasi*, to determine whether it is a suitable marker to resolve inter-specific relationships, identify intra-specific lineages, and/or study the population structure of this species.

Published nucleotide sequences in GenBank are few for *P. ariasi*. These include mitochondrial cyt b and NADH1, and nuclear 5.8S/2S/28S ribosomal RNA, EF-1α and various salivary peptides. No hypervariable markers are known - no single locus microsatellites, which are considered the most sensitive and informative markers for shallow phylogenetic inferences i.e. in population genetics. As the scope of this thesis was not only to understand current population structure but also the demographic history and evolutionary processes driving genetic variation, it was not considered time effective to isolate microsatellites for *P. ariasi*. Polymorphic microsatellites have been isolated from *P. perniciosus* (Aransay *et al.*, 2001), a sympatric species of *Larroussius* (Esseghir *et al.*, 1997; 2000; Di Muccio *et al.*, 2000). Of its six microsatellites, only three amplified consistently in *P. ariasi*, but based on a limited data set of 100 flies originating from the Massif Central and Pyrenean France, only a single size variant was recorded over all loci (S. Mahamdallie and F. Halstead, unpublished data). Of these three loci, two were shown to be single locus and polymorphic at the nucleotide sequence level – loci AAm20 and AAm24 (Aransay *et al.*, 2001) – and were used in this thesis as two anonymous nuclear loci.

1.3.3 Investigation of one salivary peptide gene putatively under selection in sandflies

The four aforementioned markers were characterized to investigate the phylogeography and population genetics of *P. ariasi*, because it was probable that they were evolving neutrally or under purifying selection, not under positive directional or balancing selection resulting from interactions with the environment, mammalian hosts or *Leishmania*. In contrast, the fifth marker characterized in this thesis was chosen for its potential as a marker under positive or balancing selection and its relevance in the *Leishmania* transmission cycle. During blood feeding female sandflies counteract their host's protective haemostatic, inflammatory and immune responses, by secreting a suite

of potent pharmacological substances into their saliva (Ribeiro and Francischetti, 2003). Salivary peptides' relevance to the disease transmission cycle have been demonstrated in mouse models, where co-inoculation of sandfly homogenised salivary glands with *Leishmania* parasites has been shown to exacerbate parasite load and thus the course of infection (e.g. Belkaid *et al.*, 1998). Conversely, pre-exposure to sandfly bites (Kamhawi *et al.*, 2000) or saliva (Belkaid *et al.*, 1998) is associated with protection against *Leishmania* development, through either cell-mediated (CM) immunity or anti-saliva antibody production of the vertebrate host.

Salivary gland apyrase has been studied in a diverse range of haematophagous arthropods e.g. sandflies (Ribeiro *et al.*, 1989), blackflies (Cupp *et al.*, 1993), tsetse flies (Mants and Parker, 1981), and mosquitoes (Ribeiro *et al.*, 1984). Sandfly (both Old and New World) apyrase is homologous to the *Cimex* apyrase family of proteins (Valenzuela *et al.*, 1998). Binding to Ca^{2+} activates the apyrase to function as a potent anti-platelet factor, by the hydrolysis of platelet activator ATP and ADP, and it inhibits the host's inflammatory and vasodilation responses (Riberio *et al.*, 1986; 1987a; Valenzuela *et al.*, 1996; 1998). Apyrase has the most abundant transcript in the salivary gland cDNA library of *P. ariasi* (Oliveira *et al.*, 2006). Pre-sensitisation of mice by injection of a DNA plasmid expressing apyrase of *P. ariasi* was shown to produce the second strongest CM delayed-type hypersensitivity (DTH) response, accompanied by a no antibody response, after subsequent exposure of mice with salivary gland homogenate (SGH) (Oliveira *et al.*, 2006). Reverse antigen screening revealed that the DTH response induced by inoculation using apyrase plasmids was consistent with a CM recall response associated with protection against *Leishmania* infection (Kamhawi *et al.*, 2000). Natural variation in the apyrase of *P. ariasi* may therefore influence the ZVL transmission cycle in Mediterranean Europe, and this variation should be considered if apyrase is selected for use in an anti-*Leishmania* vaccine. Sandfly species salivary peptides have been used experimentally as vaccine candidates in other transmission cycles (Morris *et al.*, 2001; Valenzuela *et al.*, 2001a; Collin *et al.*, 2009). However, few studies have aimed to understand the evolution of the salivary genes in (natural) sandfly populations (Milleron *et al.*, 2004a; Elnaiem *et al.*, 2005). For example, this evolution may be driven by an arms race as often observed in endoparasite-host immunity gene systems (Endo *et al.*, 1996), which could hinder vaccine success. Apyrase was characterized in this thesis both among related sandfly species and populations of *P.*

ariasi to investigate genetic variation driven by positive directional or balancing selection, which might result from sandfly peptide-host-parasite interactions.

1.4 Theory of speciation and statistical methods for identifying species and intra-specific lineages using DNA sequences

The species is considered a fundamental unit in biology, whose delimitation has real purpose in vector-borne disease transmission cycles, because correct identification of vectors is important for targeted interventions (Curtis, 1999). Two of the dominant speciation concepts, which are referred to in this thesis, include: (i) the Biological Species Concept (BSC) (Mayr, 1942; 1963) and (ii) the Phylogenetic species concept (PSC) (Eldredge and Cracraft, 1980; Nelson and Platnick, 1981). BSC defines a species as a group of interbreeding natural populations that are reproductively isolated from other such groups, and not based on phenotypic similarly. The BSC still remains the most widely accepted species concept 60 years after its formulation, and has not been rejected even though it is inapplicable to asexual organisms, and its premise of inbreeding in terms of gene flow introduces many caveats in delimiting speciation with respect to geographical proximity in nature i.e. in the case of ring species (Donoghue, 1985), or temporal separation (Willmann and Meier, 2000). The most frequent alternative to the BSC, the PSC, has been defined as a character-based concept "...the smallest aggregation of (sexual) populations or (asexual) lineages diagnosable by a unique combination of character states" (Wheeler and Platnick, 2000), or as a lineage-based concept "...a basal group of organisms all of whose genes coalesce more recently with each other than with those of any organism outside the group" (Baum and Donoghue, 1995). PSC applied to DNA sequences also has its caveats, for example any number of natural events (e.g. hybridization with inter-specific introgression) can result in the non-monophyly of a species (Avise, 2000).

A phylogenetic tree describes the evolutionary ancestor-descendant relationships between DNA sequences (or organisms) showing timing and direction of mutations and the position of shared characters. These trees can be used in both species delimitation and to identify amino acid residues showing evidence of being shaped by natural selection (e.g. location of excessive non-synonymous substitutions) (Holder and Lewis, 2003). Traditional methods of tree reconstruction include distance matrix methods (Neighbour-Joining (NJ), not further discussed) and tree searching methods that use an optimal criterion to search for the best tree (Maximum Parsimony (MP) or Maximum

Likelihood (ML)), and then assess the confidence of this optimal tree (e.g. by bootstrapping). Whereas, the newer Bayesian approach simultaneously produces both a tree estimate and a measure of uncertainty for group nodal support (Lewis, 2001). MP, ML and Bayesian reconstructions are all discrete character-based methods, but differ in their ability to incorporate models of character changes, how they construct the 'tree space' to find the optimal/true tree, and how they assess the statistical confidence of a given tree. The principles of each of these algorithms have been reviewed comprehensively elsewhere (see Lewis, 2001; Holder and Lewis, 2003), so the next section briefly outlines the advantages and disadvantages of these character-based methods, the genetic content they utilize and appropriate application.

MP is often used to construct trees for large datasets and is considered robust for closely related species or for dense datasets (which can avoid long-branch attraction). The MP optimal tree is that requiring the least number of character changes to explain the data. Although rapid to compute, MP's drawbacks mainly stem from its inability to include nucleotide substitution models (Hall, 2004). ML improves on MP in its ability to correct for multiple hits at a single base position and, therefore, is appropriately implemented to reconstruct the relationships between sequences that have been separated for a long time or are evolving rapidly (Holder and Lewis, 2003). However, ML considers all probable mutation pathways that are compatible with the data which, along with the bootstrap to assess the statistical confidence of a grouping, makes the computation of this algorithm a burden and an obstacle for its application. Both ML (limited options) and Bayesian (extensive options) can incorporate nucleotide substitution models, which culminate in the most complex model the General Time Reversible model (GTR). This allows unequal nucleotide frequencies and all six changes between nucleotide states to occur at different rates (Rodríguez *et al.*, 1990). Moreover, information can be included that allows various levels of substitution rate heterogeneity across sites (Lewis, 2001). Such breadth of substitution-model options makes Bayesian modelling an attractive alternative to ML, especially as it uses a relatively fast algorithm (Markov Chain Monte Carlo (MCMC)) to generate the tree space and a posterior probability approach to support a given hypothesis (i.e. tree). In Bayesian reconstruction, the true tree is one that maximises the posterior probability density – an estimate proportional to the product of the prior probability and the likelihood (Lewis, 2001) – which is conditional on the model, the priors, and the data. The reliability of the method rests, therefore, on the model and parameter priors that are

assumed by the user (Huelsenbeck *et al.*, 2002). Erixon *et al.* (2003) showed that Bayesian modelling is more sensitive to under-parameterization. Therefore, when applied in this thesis, all models will be compared to the most complex GTR+I+G model, and Bayes factors calculated to support one model over the other.

As mentioned, support for a group, node or phylogenetic species is estimated either through bootstrapping or posterior probabilities, but there can be debate over the value deemed as a reasonable cut-off. A general consensus concludes that a bootstrap value of 70% is an indication of strong group support (Hillis and Bull, 1993). Bayesian posterior probabilities are used more conservatively than bootstrap values (Huelsenbeck *et al.*, 2002), e.g. Mar *et al.* (2005) found a posterior probability of 100% corresponds to about an 80% bootstrap proportion. However, whether the posterior probability is too trusted for estimating group support, or whether these two estimators measure something qualitatively different, is an area of debate (Erixon *et al.*, 2003).

Phylogenetic trees are ideally used to investigate relationships among species, but are also utilised to distinguish between inter-specific diversification and intra-specific coalescence. Several statistical methods offering operational criterion for delimiting species based on DNA sequence clusters have been proposed. The Birky 4x rule (Birky *et al.*, 2005), delimits different species when two monophyletic groups have a mean sequence difference between them greater than four times $\theta=2N_e\mu$, where N_e is the effective population size and μ the mutation rate/base/generation. The Mixed Yule Coalescent (MYC) method (Pons *et al.*, 2006) uses a clock-constrained phylogram and ML to determine the point of transition from slow to faster branching rates expected at the boundary between species-level and population-level evolutionary processes. Alternatively, Hart and Sunday (2007) use the 95% parsimony connection limit of a TCS network to provide a simple quantitative standard for phylogenetic species. In this thesis, one aim is to delimit intra-specific lineages within morphologically identified *P. ariasi*, so a genealogical network approach was implemented both for species delimitation, to define intra-specific lineages/haplogroups (Avise, 2000) and to reconstruct evolutionary relationships. Intra-specific data can be subject to processes such as parallel mutation, hybridization, recombination and gene-conversion, and such evolutionary histories can not be modelled by a bifurcating tree (Posada and Crandall, 2001).

1.5 Theory of genetic selection in relation to the thesis' aims

Initial studies on the mechanisms of evolution were based on the principles of Darwin's (1859) evolution by natural selection, Mendel's theoretical and practical studies on the laws of hereditary and the population genetic theorems of Fisher (1918), Wright (1921) and Haldane (1932). Applying Darwin's theory of natural selection on a genetic (as opposed to phenotypic) level, adaptation arises by the transmission of genotypes that promote survival in their current environment. The pressures of selection are differentially experienced depending on how a variant allele or genotype frequency is correlated to the fitness of an individual. This variation in a natural population drives the fittest or 'improved' genotypes (as a function of its alleles) to become present in disproportional excess and thus contribute more to the next generation. Ultimately a population becomes 'adapted' to its environment and diverges genetically from those individuals inhabiting environments with differing selection pressures and allelic fitness (Hartl, 1981).

1.5.1 Types and genetic signals of selection

The terminology used to describe the various modes of selection pressures on molecular evolution can vary within different scientific communities, and therefore this thesis follows the definitions given by Nielsen (2005). Purifying (or negative) selection describes any type of selection against new deleterious mutations, eliminating them from a population due to their negative fitness effect. Gene regions often under purifying selection reflect their functional importance, for example in proteins where mutations cause disruption to structure and consequently function (Zhao *et al.*, 2003). Signals of purifying selection include lower diversity in coding versus non-coding regions, a deficiency of rare and intermediate frequency alleles, and a low level of nonsynonymous compared to synonymous divergence (among species). Along with purifying selection, positive selection – where new mutations are advantageous – is an example of directional selection. Both eliminate variation within populations lowering its heterozygosity. However, positive directional selection is accompanied by raised nonsynonymous (compared to synonymous) divergence (Hurst and Smith, 1999). When a mutation is driven to fixation by positive selection, neighbouring sites that are neutral but linked through short genomic distances can experience a loss in their variability in a process termed a selective sweep (Nielsen, 2005): a within population force distinguished from positive directional selection by no accompanying change in

divergence. Opposing the process of directional selection, which erodes genetic variability, balancing selection maintains multiple alleles above the rate of neutral mutations within a population and between species, which is evidenced by high levels of heterozygosity. Mechanisms by which balancing selection maintains variability include heterozygote advantage (over-dominance) and frequency-dependent selection. The former describes heterozygotes that have greater fitness than homozygotes, whereas the latter concerns the fitness of a genotype that is dependent (negatively or positively) on its frequency relative to the other genotypes in the population (Gilbert *et al.*, 1998).

1.5.2 Selection and its relevance to this study

The appropriate application of a locus in genetic studies requires knowledge of the processes of molecular evolution to which it is subject. Population processes affecting species' geographic spatial structure are commonly investigated through population genetic tests that assume neutral evolution of a marker e.g. F_{ST} as an estimator of genetic differentiation, AMOVA testing for population sub-division, allele frequency spectrum based neutrality tests identifying demographic events (Chapter 2). For this reason mtDNA (or chloroplast DNA of plants) has historically been an informative molecular marker (Hewitt, 1999; Avise, 2000), for which selection has been assumed to be absent. However, a growing body of evidence expresses caution in using mtDNA as a neutral marker, because its haplotypes are shown to be under pressure of direct selection (e.g. Mishmar *et al.*, 2003; Ballard and Kreitman, 1994) or indirect selection (reviewed Ballard and Whitlock, 2004). Bazin *et al.* (2006) analyzed an extensive range of animal sequences and found that mtDNA nucleotide diversity was not correlated with effective population size, showing that in fact mtDNA diversity distribution is explained by recurrent adaptive evolution - selective sweeps. One example of indirect selection that might affect *P. ariasi* is mitochondrial cytoplasmic hitchhiking with *Wolbachia* transmission (Benlarbi *et al.*, 2003). Acknowledging the possibility of selection at any locus, mitochondrial or nuclear, this thesis will test the assumption of neutrality for all loci, in addition to seeking evidence of positive directional or balancing selection on the salivary peptide apyrase of *Phlebotomus*.

1.6 Population structure: population demographics and genetics

1.6.1 Theory of population structure and some population genetic parameters

In natural environments, members of a species are rarely distributed homogeneously in space. Sub-division of a species into "populations" is often caused by environmental patchiness – a mosaic of areas with favourable and unfavourable habitats – a result of both past historical and contemporary population processes. Even in landscapes where species' habitats are continuous, populations can become sub-divided to some extent if migration is smaller than the habitat range, constituting a metapopulation *sensu lato* (Hanski and Gilpin, 1997). Accordingly, in the genetic sense, a population is an interbreeding group of individuals sharing a common geographical area (Hartl, 1981), and it is this spatial structure that has important consequences in determining the genetic structure of natural populations (Slatkin, 1973).

Population structure is composed of two distinct yet interrelated parts: demographic structure determined by the processes associated with birth, death, extinction, colonization, population density and migration distances (gene flow); and genetic structure determined by genetic drift, mutation, selection and recombination (Slatkin, 1995). Mutation, a heritable change in genetic material, is considered the "ultimate source of genetic variation" (Hartl, 1981), and the neutral theory of molecular evolution (Kimura, 1968; 1983) proposes that in most natural populations the high level of polymorphisms observed and their changes in frequency are driven by the fixation of neutral mutations by random genetic drift not Darwinian selection.

Population genetic theory allows the prediction of actual genetic structure from knowledge of observed genetic structure, which in turn allows conclusions to be drawn about demographic structure and its processes (Slatkin, 1995). Within-species genetic diversity is thought to reflect population size, history, ecology, and ability to adapt. The effective population size, N_e, is an example of a core population genetic parameter (Wright, 1931). It is defined as the number of individuals that have descendents at the next generation, which is approximately equal to one-half the number of mating individuals. Neutral theory predicts that a positive relationship should exist between N_e and the extent of genetic variation (allelic diversity and heterozygosity) at loci not subject to strong selection (Kimura, 1983). Based in this premise, small populations formed by bottle-neck, vicariant or other population contraction events, should be less polymorphic than large populations. This was demonstrated by Spielman *et al.* (2004), who showed that lowered heterozygosity in small populations was associated with

lower evolutionary potential, compromised reproductive fitness, and elevated extinction risk.

The stochastic process of genetic drift is a corollary of neutral population structure. In finite populations, chance natural sampling from the ancestral population and an inbreeding-like effect of population sub-division cause the loss of some alleles and the accumulation and eventual fixation of others (and thus a heterozygosity decline). This process is known as genetic drift. Accordingly, each sub-population can have its own genetic trajectory (Hartl, 1981), and therefore genetic drift can result in neutral evolutionary divergence between sub-populations (i.e. populations isolated by landscape fragmentation). In ideal (diploid) populations (Wright-Fisher model) the rate at which genetic drift causes an increase in divergence between isolated populations is given by 1/(2N) where N is number of mating individuals. Therefore, as the rate of change of gene frequency by random drift depends on the size of the population, N_e can be thought of in terms of a measure of the strength of the stochastic process of genetic drift in a finite population (Wang and Caballero, 1999).

Gene flow is a further important component of neutral population structure, where it determines the extent to which each local population evolves as an independent evolutionary unit. Gene flow opposes mutation and random genetic drift, allowing genetic exchange which limits genetic divergence, and results in homogenization and thus sub-population connectedness (Hedrick, 2000). One generation of complete gene flow (accompanied by random mating) should cause differentiation among sub-populations to disappear completely.

As the geographical structure of natural populations can be complex, patterns of allele frequencies attributed to gene flow and drift between sub-populations have been simplified by models such as the "stepping stone model" (Kimura, 1953) or the "neighbourhood model" (Wright, 1943; 1946). Discontinuous sub-population distribution causes patterns of allele frequencies to exhibit large changes over short distances ("step" or discontinuous changes), where barriers to gene flow (discontinuous habitat) or putative adaptive hotspots can be inferred. Alternatively, individuals in a natural population can be continuously distributed, but the continuum is formed by random mating units (neighbourhoods) (migratory distance is larger than distance separating populations, but smaller than the entire species range). As individuals are more likely to reproduce locally, allele frequency patterns will follow a gradual (clinal) pattern, showing patterns according to an isolation-by-distance model (Wright, 1943).

Such patterns could be seen in sub-populations arranged along a linear axis, e.g. sampling along a latitudinal, longitudinal or altitudinal gradient.

1.6.2 Estimating population genetic structure using the principles of neutral theory: inferring selection and demographics

One of the important outcomes of neutral theory lies in its power to make statistical predictions of the mutation and allele distribution within populations and between species, by providing a null hypothesis for studying molecular evolution. Neutrality tests are categorized into two groups (Nielsen, 2001), those assessing molecular evolution through polymorphism and divergence between different classes of mutations to detect selection, and those based on the haplotype/allele frequency spectrum.

Neutrality tests based on DNA sequence evolution estimate polymorphism (within a species) and divergence (between species). They are amongst the most powerful tests for selection, in part explained by their general robustness to demographic alternatives (Garrigan and Hedrick, 2003), and so population size is not required to be at statistical equilibrium. Tests based on the evolution of sequences (mutations) are appropriate to detect long-term selection. In population genetic studies, the most widely used test for protein coding data is the McDonald-Kreitman (1991) (MK), which compares the relative counts of nonsynonymous and synonymous substitutions, with a null (neutral) hypothesis predicting the ratio of nonsynonymous to synonymous substitutions to be the same within populations and between a closely related outgroup if driven by mutation and genetic drift. Deviations from this null hypothesis indicate either directional or balancing selection.

Statistical tests modelling neutrality based on the allele frequency spectrum make demographic assumptions (e.g. constant population size, no population structure) and genetic assumptions of neutral mutations (do not affect fitness), which along with genetic drift are the only forces driving genetic variation. Based on these assumptions, such statistics as Tajima's (1989) or Fu and Li's (1993) *D*, are appropriate to distinguish population growth or decrease from constant size, and population sub-division. However, as noted, the null hypothesis is a composite hypothesis, so violations of these assumptions can be explained by the occurrence of selection (Nielsen, 2001). Consequently, these alternative explanations will be considered for the test results obtained in this thesis.

1.7 Thesis aims

This study of *P. ariasi*, the vector of *L. infantum* in southwest Europe, aims:

1. To confirm that *P. ariasi* is a single species over the geographical range investigated, so that any natural variation can be attributed to neutral or adaptive evolution rather than reproductive barriers.
2. To determine the neutral genetic differentiation of *P. ariasi* across the geographical range investigated.
3. To use these results to infer the historical demographic events and identify the landscape features that affect the distribution of this species.
4. To design a molecular protocol to score the genotypes of the salivary peptide apyrase in individual sandflies, in order to investigate whether apyrase is subject to selection that might be driven by sandfly peptide-host-parasite antagonism or environment.
5. To evaluate the implications of the findings for the emergence of *P. ariasi* (and the transmission of *L. infantum*) in western Europe.

CHAPTER 2

Multiple genetic divergences and population expansions of a Mediterranean sandfly, *Phlebotomus ariasi*, in Europe during the Quaternary glacial cycles

2.1 Introduction

The oscillating climatic extremes of the Quaternary (Pleistocene and Holocene epochs) have produced repeated shifts in species' distribution limits across Europe that are highly variable in space and time (e.g. Coope, 1994; Hewitt, 1996; 1999; 2000; 2001; 2004a; Taberlet *et al.*, 1998; Petit *et al.*, 2003; Gómez and Lunt, 2006). Evaluations based on insect subfossil distribution show extant species to have responded to Pleistocene climate oscillations, by evolving out and/or moving out of trouble (Coope, 1994). It is widely accepted that species' geographical distribution limits are locally not globally determined, dependent on individual ecological requirements, dispersal ability, presence of pre-colonizers and barriers to gene flow, factors that are closely correlated with climatic variables and biogeographic barriers (e.g. review Huntley, 2001; Schmitt, 2007). *Phlebotomus ariasi* is a vector of *Leishmania infantum* in the Mediterranean bioclimates of Iberia and France (Ready, 2008). The distribution range of this species extends into northwest Italy where its low population densities (Maroli *et al.*, 2008) suggest re-colonization into France only from Iberia. *P. ariasi* is endemic in the Iberian Peninsula and shows a current distribution in the previously glaciated regions of France (Pyrenees and the Massif Central), abundant and widespread in southern France up to latitude 45^0 N (Rioux and Golvan, 1969). The aim of this chapter is to investigate whether the opportunities for its current populations, and therefore the *Leishmania* it transmits, to spread northward have been constrained by the effects of past climate change and the role played by biogoegraphical barriers (e.g. the Pyrenees mountains) in limiting the re-colonization of France by a Mediterranean species.

The distribution shifts of temperate European species, including insects, in response to Quaternary climate changes have been well studied, where periods of climate cooling forced their contraction into warmer refugia of southern latitudes, followed by subsequent expansion during climate warming: refugia were commonly limited to between 30^0 N and 40^0 N, restricted at northern latitudes by the Fennoscandian ice sheet and permafrost, and the Mediterranean Sea in the south

(Hewitt, 2004b; Taberlet *et al.*, 1998). There is strong evidence that the cold glacial climates isolated temperate species into three independent Mediterranean refugia: (i) Atlantic-Mediterranean (Iberia, Maghreb); (ii) Adriatic-Mediterranean (Italian Peninsula); (iii) Pontic-Mediterranean (Balkan Peninsula) (Hewitt, 1999; Taberlet *et al.*, 1998; Schmitt, 2007). Little or no genetic exchange occurred between these refugia as species would have had to migrate over several hundred kilometres of open sea. Such demographic processes can observably shape species' genetic structure, for example fragmented niches tended to sub-divide species into independently evolving genetic groups (lineages/haplogroups) each containing a large proportion of unique haplotypes (Taberlet *et al.*, 1998; Hewitt, 2000; Schmitt, 2007).

In periods of climate warming, species re-colonized previously unsuitable cold landscapes by northward expansion from their southern refugia. Most reports on the four paradigms of post-glacial re-colonization in Europe – categorized by their refugia and the mountain ranges that act as barriers to their dispersal – have focused on the dispersal of temperate not Mediterranean species (Hewitt, 1999; 2004a; Habel *et al.*, 2005; Schmitt, 2007). Temperate species often show isolation in more than one allopatric glacial refugium. Examples of this are common in species of the Iberian Peninsula (e.g. Martínez-Solano *et al.*, 2006; Gómez *et al.*, 2007), the mountain ranges of which offer high microclimatic scope to create heterogeneous landscapes of diverse microhabitats (Hewitt, 1996). The punctuation of large refugial regions is well supported, described by the "refugia within refugia" paradigm of Gómez and Lunt (2006).

The mountain ranges of Mediterranean Europe (including the Pyrenees) offered some of the main refugia for retreating northern temperate species and for the interglacial survival of montane species in the same region that tracked vertical shifts in their habitats (Hewitt, 1996; 2004a). In addition these high mountain systems shaped the post-glacial expansions from Mediterranean refugia, often cited as hybrid (suture) or secondary contact zones, the latter evidenced by the presence of parapatric lineages (Taberlet *et al.*, 1998; Hewitt, 2004b; Schmitt, 2007). Species persistence in the fragmented yet stable mountain environments is evidenced by their harbouring relict populations (bank vole *Myodes glareolis* in the Pyrenees; Deffontaine *et al.*, 2009), endemic species (Varga and Schmitt, 2008), deeper haplogroups, (Hofman *et al.*, 2007), and high genetic diversity (Gugerli *et al.*, 2001).

This chapter provides the first genetic study of *P. ariasi*, including both phylogenetics and population differentiation, taken from across its South-North range. This species' geographically sympatric vector, *P. perniciosus* does not provide a comparable population distribution model, because it is found peri-domestically, at lower altitudes and in southern France comprises two independent lineages originating from glacial refugia in Iberia or Italy/north Africa/Malta (Esseghir *et al.*, 1997; 2000; Pesson *et al.*, 2004). Assuming *P. ariasi* is a single species (phylogenetic and biological), which has yet to be determined, the simplest model should consider this species to be unable to survive in its current position north of the Pyrenees. My first hypothesis is that the species constitutes a single continuous population, with northward post-glacial expansion from a refugium most likely to be in southern Iberia or north Africa. Alternatively, following the lineage distribution of *P. perniciosus*, north African flies may be of an independent lineage to Iberian/French *P. ariasi*.

In France *Quercus spp.* (*Q. pubescens*, *Q. ilex*) are considered biological indictors for the presence and abundance of *P. ariasi* (Rioux and Golvan, 1969; Riou, 2004). Based on this premise phylogeographic studies of cpDNA might allow for the inference of alternative Pleistocene glacial and post-glacial population responses of *P. ariasi*. Lumaret *et al.* (2002) showed genetic support for two Iberian *Q. ilex* refugia, one eastern and the other in the south and (north) west: a cpDNA distribution that is approximately coincident to two recognised morphs. Furthermore, chlorotype phylogeography suggests a post-glacial migration route into France following the Mediterranean clime by crossing the Pyrenees exclusively in the East.

An alternative demographic scenario accepts the model that mesophilous trees e.g. deciduous *Quercus* (Beaudouin *et al.*, 2007), as well as temperate mammals (Deffontaine *et al.*, 2009) and insects (Kidd and Ritchie, 2006) were present outside of southern refugia, surviving during the Last Glacial Maximum (LGM) in the protective microclimate of valleys of southern France. *P. ariasi* prefers cooler environments, the adults most abundant and active on hillsides in wooded rural regions during the dry Mediterranean summer. In Languedoc-Roussillon region of southeast (SE) France, *P. ariasi* can be found up to 1,400 m.a.s.l. (Rioux and Golvan, 1969; P.D. Ready, S.S. Mahamdallie and B. Pesson, unpublished), so may have had the capability to track limited altitudinal shifts in this region to find suitable mircoclimates and persist *in situ*. If true, then the Pleistocene climate oscillations could have accentuated the

fragmentation of *P. ariasi* habitat quality over space and time, creating multiple isolated buffered microclimates/refugia across Iberia and France.

Standard criteria for defining phylogeographic lineages have been quantitatively defined using the distribution of pairwise sequence differences within and between putative haplogroups (e.g. Naderi *et al.*, 2007). Population histories can be complex, i.e. sequential divergence with migration rather than divergence by bifurcation, where their discrimination requires a composite of summary statistics. In this way genealogical samplers use molecular genetic data (i.e. allelic diversity) and their estimated gene networks to attempt to disentangle the contributions of demographic histories and recurrent gene flow, to identify supported alternative explanations for observed variation in spatial structure (Kuhner, 2009). Implementation of these often coalescence based samplers should be taken with caution as populations and/or data do not always meet the model assumptions or parameter demands (e.g. Hey, 2010), or confidence limits of hypotheses are not assessed (e.g. Nested Clade Analysis as discussed in Knowles and Maddison, 2002).

As detailed, the responses of temperate species to Quaternary climate changes are well evidenced. However, the response of true Mediterranean and named subtropical species are less well documented. The Mediterranean regions supported isolated patches of multiple refugia. Each may have been associated with individual demographic histories, leading to no single model of response to Quaternary changes (Canestrelli *et al.*, 2007; Pinho *et al.*, 2007). *P. ariasi* is an appropriate species to investigate the correlation between the Quaternary climates and species' distributions through biogeographic patterns of genetic architecture, as it has a sufficiently high dispersal ability to spread rapidly into newly emerging suitable habitats, yet single individuals are mostly sedentary so a phylogeographical pattern is not blurred by high migration (Schmitt, 2007). Moreover, isoenzyme studies record *P. ariasi* as showing greater local geographical variation than sympatric *P. perniciosus* (Pesson *et al.*, 2004; B. Pesson, unpublished), making it easier to study the effects of past demographic events.

Considering *P. ariasi* has a preferred ecological niche (Rioux and Golvan, 1969), the presence of multiple refugia and secondary contact zones in its western Mediterranean distribution range is a plausible scenario. Moreover, it is important to identify any refugia in the northern Pyrenees as they might have given *P. ariasi* a springboard for post-glacial re-colonization northwards. Alternatively, however, such refugial populations might have blocked (*sensu* Hewitt, 2004a) the dispersal of Spanish

populations containing flies better adapted to northern environments or disease transmission. This chapter characterizes the genetic variability in morphologically identified *P. ariasi*, based on the nucleotide sequences of mitochondrial cytochrome b (cyt b) (19 populations) and three nuclear loci (18 populations). The latter are elongation factor-1 alpha (EF-1α) and two anonymous loci (AAm20 and AAm24) originally reported as microsatellites of *P. perniciosus* (Aransay *et al.*, 2001; 2003). A population was defined by being distinct either in space or time (capture year).

This chapter's aims were:

1. To confirm that morphologically identified *P. ariasi* is a single phylogenetic and biological species, to guard against demographic analyses being confounded by the presence of cryptic sibling species.
2. To test the assumption of neutrality at each of the four loci characterized, justifying their use for inferring neutral population structure.
3. To determine the genetic structure of *P. ariasi*, to explore if its distribution has been restricted by past environmental change.
4. To assess the population structure of *P. ariasi*, to identify the roles played by the Pyrenees mountains and local environmental barriers on its postglacial re-colonization of southwest France.

The findings of this study should be informative for predicting the risk of spread of zoonotic visceral leishmaniasis (ZVL) in response to climate and other environmental change.

2.2 Materials and methods

2.2.1 Sampling of *P. ariasi* and pre-molecular preparation

19 rural populations, 464 individual *P. ariasi*, were sampled along the South-North axis of its range, from Morocco through the Iberian Peninsula to southern France (Figure 2.1; Table 2.1). 15 populations originated from France: eight populations sampled from within and bordering the Massif Central region, a single population within the Massif Central (ROQ), three at the northern distribution 'leading-edge' (SAM13 and Lot LNP, RME), two at its southern foothills (CTU, SPV) and one in the Rhone valley (DRAz4); six populations from the eastern Pyrenees (PAS, TUL, IRL07, ARQ06, ARQ08, CAT); and two populations from the central Pyrenees (HP1, HP2). Four populations were outgroups to France; northeastern (TRJ) and northwestern (CSP) Spain; northern Portugal (CHR) and Marrakech Morocco (AGH). Numbers of individual *P. ariasi* per population ranged between 13 and 54, sample sizes appropriate to confer statistical support for population genetic tests and to be comparable - between 22 to 27 individuals in 12 to 14 populations (locus dependent).

Collections of adult flies were made using either Centers for Disease Control (CDC) miniature light traps (Sudia and Chamberland, 1962) placed overnight in peri-domestic locations, usually 1-2 m above the ground near farm-animal shelters, or by sticky traps (A4 papers covered in castor oil) placed in road-side walls, retrieving after four nights. In a field laboratory, flies from light traps were immobilized at -20°C and stored in 80% analytical grade ethanol or dry in liquid nitrogen. Flies on sticky papers were removed with fine brushes wetted with 96% (v/v) ethanol and stored in 80% analytical grade ethanol at 4°C. Longer-term storage was in ethanol at -20°C or frozen dry at -80°C or -196 °C.

All *P. ariasi* were identified (by the author or P.D. Ready) based on external form, colour and size (P.D. Ready, unpublished) and on internal morphological characters of the head and genitalia (Gállego *et al.*, 1992). *P. ariasi* and other *Phlebotomus* (Table 2.1) used for molecular characterization were dissected according to the sterile procedures of Testa *et al.* (2002): flame sterilizing dissection forceps and microneedles between preparations; dissections carried out in a room away from the molecular biology laboratory to minimize polymerase chain reaction (PCR) carry-over risk. Voucher specimens of slide-mounted heads and abdominal terminalia in Berlese fluid, were placed in the phlebotomine collection of the Department of Entomology, Natural History Museum, London.

Figure 2.1 Digital Elevation Map of the western Mediterranean showing locations where 19 *P. ariasi* populations were sampled for molecular characterization. Additional information on location environment given in Table 2.1.

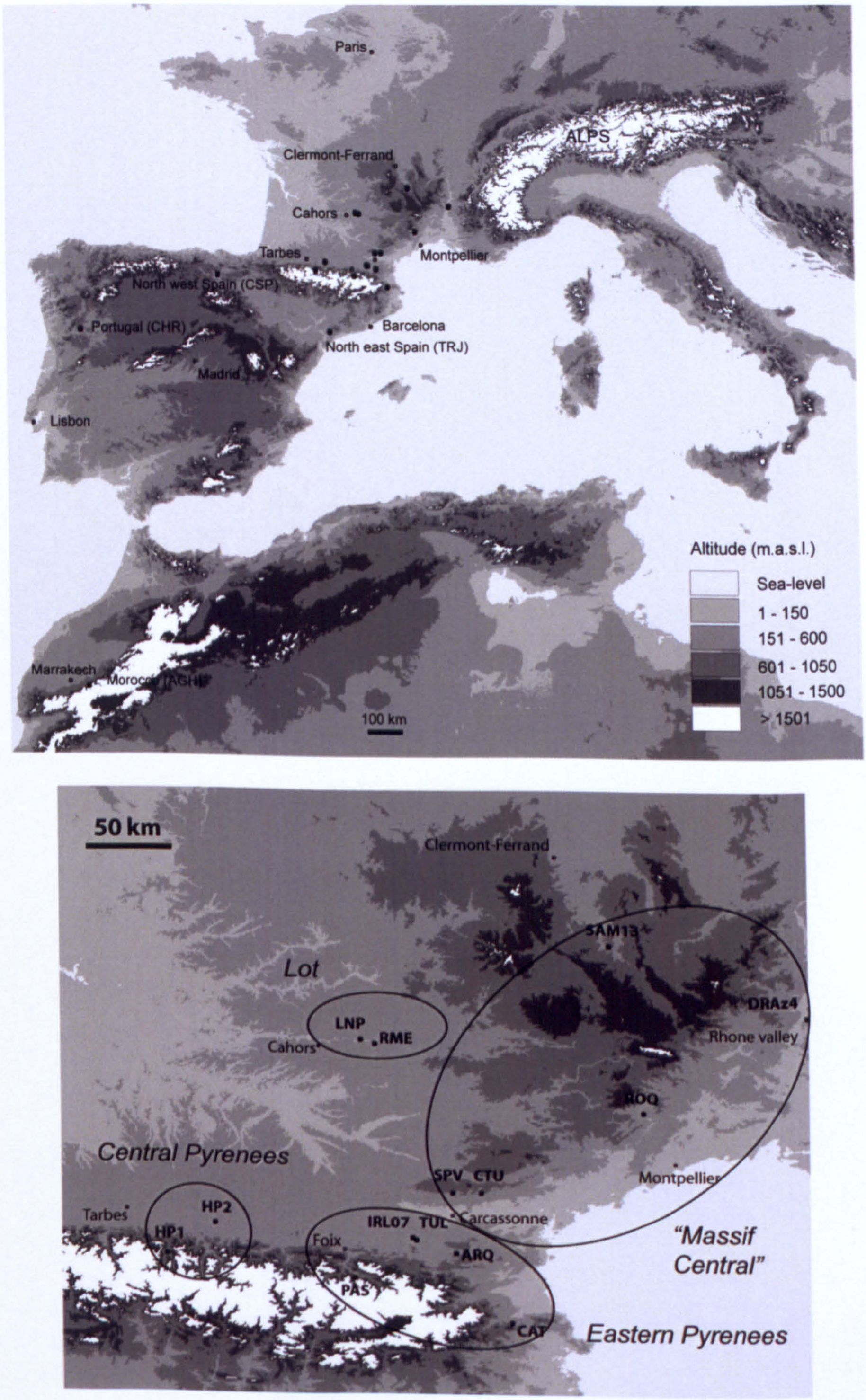


Table 2.1 Origins of the molecularly characterized populations of *P. ariasi* and other sandflies.

Population	N (code)	Location	Country	Geographical region	Latitude (DD)	Longitude (DD)	Altitude (m.a.s.l.)	Date	Collected by	Trap
P. (Larroussius) ariasi										
AGH	17	Aghbalou	Morocco	Western Atlas, Wilaya de Marrakech	31.00000	-7.73330	1250	27-28/08/2006	S. Boussaa & A. Boumezzough	Sticky
CHR	24	Cheires	Portugal	Northern Portugal, Distrito de Vila Real	41.26667	-7.53333	499	24-25/07/1996	B. Pesson & C. Pires	CDC
CSP	24	Caicedo Sopeña	Spain	North-west Spain, Pais Vasco, Alava	42.78333	-2.900	585	22-23/07/2000	A.M. Aransay & J.M. Testa	CDC
HP1	27	Haute-Pyrénées 1	France	Central Pyrenees, Haute-Pyrénées	42.91427	0.35974	607 - 789	July 2006	Authors	Sticky
HP2	19	Haute-Pyrénées 2	France	Central Pyrenees, Haute-Pyrénées	43.13901	0.69808	307 - 428	July 2006	Authors	Sticky
PAS	54	Aston	France	Eastern Pyrenees, Ariège	42.76164	1.66310	647	20-21/07/2005	Authors	CDC
IRL07	22	Irlat07	France	Eastern Pyrenees, Aude	43.04949	2.06878	421	19-20/07/2007	Authors	CDC
TUL	24	La Tuilerie	France	Eastern Pyrenees, Aude	43.03310	2.09572	467	19-20/07/2007	Authors	CDC
ARQ06	44	Arques06	France	Eastern Pyrenees, Aude	42.94381	2.36861	382	10-15/07/2006	Authors	CDC
ARQ08	23	Arques08	France	Eastern Pyrenees, Aude	42.94381	2.36861	382	09-10/07/2008	Authors	CDC
CAT	16	Céret	France	Eastern Pyrenees, Pyrénées-Orientales	42.450	2.750	600	05-06/08/2001	Authors	CDC
TRJ	23	Torroja del Priorat	Spain	North-east, Spain, Catalonia, Tarragona	41.21667	0.81667	332	27-28/08/1997	B. Pesson & M. Gállego	CDC
CTU	24	Citou	France	South Massif Central, Aude	43.37426	2.54141	358	07-08/07/2008	Authors	CDC
SPV	24	St Pierre en Vals	France	South Massif Central, Aude	43.37356	2.34766	351	11-12/07/2008	Authors	CDC
ROQ	15	Roquedur	France	Massif Central, Hérault	43.96670	3.66670	337	22-24/07/1998	B. Pesson & R. Killick-Kendrick	CDC
SAM13	24	Chilhac	France	Massif Central, Auvergne	45.157	3.44160	510	06-10/07/2004	P.D. Ready & S. Riou	Sticky
DRAz4	23	Les Tourrettes	France	Rhone Valley, Drôme	44.650	4.80170	100	15-17/07/2004	B. Pesson & F. Paroissien	CDC
LNP	24	Pourcel	France	Saint-Martin-Labouval, Lot valley	44.47967	1.72450	316	22-23/07/2007	Authors	CDC
RME	13	Rames	France	Carjarc, Lot valley	44.44948	1.82130	270	24-25/07/2007	Authors	CDC
P. (Larroussius) major	1 (MSH0861)	Niaz	Iran					30-31/04/1985	P. Parvizi	Sticky
P. (Larroussius) neglectus	1 (MAG05)		Hungary	Near Serbian Border				July 2006	P.D. Ready & R. Farkas	Sticky
P. (Transphlebotomus) mascittii	1 (STG02)	Haute-Pyrénées 1	France	Central Pyrenees, Haute-Pyrénées				July 2006	Authors	Sticky
P. (Adlerius) brevis	1 (MSH0545)	Ghurt Tappeh	Iran					17-18/04/1985	P. Parvizi	Sticky
P. (Adlerius) halepensis	1 (AEZ51)	Goyogaj	Iran	Kaleybar, East Azerbaijan				10-11/09/2005	P. Parvizi	CDC
P. (Adlerius) halepensis	1 (AEZ151)	Aslanbaghloo	Iran	Kaleybar, East Azerbaijan				20-21/07/2005	P. Parvizi	CDC
P. (Adlerius) halepensis	1 (AEZ127)	Otokandi	Iran	Kaleybar, East Azerbaijan				26-27/07/2005	P. Parvizi	CDC

CDC light-traps = Centers for Disease Control (CDC) miniature light traps (John W. Hock Company, Gainsville, Florida and Hausherr's Machine Works, Toms River, New Jersey); Sticky traps = A4 sheets of paper both sides covered with castor oil; Sample number. N = sample size. DD = decimal degrees.

2.2.2 Molecular characterization

DNA extraction

Genomic DNA was extracted from each sandfly thorax and/or anterior abdomen, according to the ethanol precipitation based protocol of Ish-Horowicz (1982) and described for phlebotomine sandflies by Ready *et al.* (1991) (Appendix 2.1).

Polymerase Chain Reaction (PCR) amplification

PCR amplifications and sequencing reactions were performed using a 0.2 ml 96-well format in one of two thermocyclers (Techne Genius Thermal Cycler or Applied Biosystems Perkin Elmer model 9700). A single PCR reaction for loci cyt b and EF-1α gave a final volume of 25 µl, that included: 1 µl of DNA extract; 1x Colorless GoTaq® Flexi buffer (Promega Corporation); 100µM each dNTP (Applied Biosystems Inc); 1.5mM $MgCl_2$ (standard concentration unless otherwise stated) (Promega Corporation); 500ng of each forward and reverse primers (Sigma-Genosys); 1.5U *Taq* (GoTaq® Flexi DNA polymerase, Promega Corporation). A single PCR reaction for loci AAm20 and AAm24 gave a final volume of 20 µl, where concentrations were the same as above but modifying each forward and reverse primer (Sigma-Genosys) to 0.5µM. Volumes were made-up to total using PCR grade water (Sigma). To minimize potential of contamination, all PCRs were carried out in a laminar flow hood with DNA free pipettes using filtered tips.

All primers are quoted in base pairs (bp) from 5′ to 3′ on the DNA sense strand and fragment lengths include primers unless otherwise stated.

Locus cytochrome b (cyt b)

A 796 bp fragment was amplified that included the 3′ terminus of the cyt b gene in addition to the immediate downstream Intergenic Spacer (IgS) and transfer RNA ($tRNA^{ser}$ (UCN)), and was targeted by primer pair CB1-SE (Testa *et al.*, 2002): TATGTACTACC[C]TGAGGACAAATATC [C to A nucleotide substitution in the modified sense strand primer CB1 of Simon *et al.* (1994) used for sequencing], and CB-R06: TATCTAATGGTTTCAAAACAATTGC (Parvizi and Ready, 2006). PCR cycling parameters (adapted from Parvizi and Ready, 2006) used a 'hot start' at 80°C; an initial 3 min denaturation step at 94°C; 35 cycles of denaturation 94°C for 30 sec, annealing at 51°C for 30 sec and extension at 72°C for 90 sec; a final extension step of 72°C for 10 min; terminating by holding at 4°C. If PCR failed because of DNA degradation, two

short overlapping fragments were amplified. 5′ 488 bp were targeted by primer pair CB1-SE with CB3-R3A: GCTATTACTCCYCCTAACTTRTT (Esseghir *et al.*, 2000). 3′ 388 bp were targeted by primer pair CB3-FC with CB-R06: CAYATTCAACCWGAATGATA (Esseghir *et al.*, 2000). PCR annealing temperatures: the first five cycles at 40 or 44°C, and the final 35 cycles at 44 or 48°C, for CB3-R3A or CB3-FC, respectively.

Locus elongation factor-1α (EF-1α)

A 856 bp fragment of EF-1α was targeted by the conserved primers designed for *Larroussius* (Esseghir *et al.*, 2000); EF-FSE: TGAGCGTCAGCGTGGTATC and EF-SE2: CGGGTGGTTCAGTACGATGA. PCR thermocycling conditions were as quoted for cyt b with the first 5 cycles annealing at 51°C, and the final 35 cycles annealing at 55°C (optimized for *P. ariasi* based on Esseghir *et al.*, 2000). Direct sequencing of this product in *P. ariasi* revealed superimposed nucleotide peaks, often of equal amplitude, at single base positions. The method of PCR Amplification of Specific Alleles (PASA) was used to directly resolve genotypes where two or more heterozygous base positions occurred (Sommer *et al.*, 1992). 6 novel allele-specific reverse primers were designed to discriminate ambiguous genotypes, by pairing with the conserved EF-FSE forward primer. In the following PASA primer names, the number denotes the variable 3′ nucleotide (underlined) which conferred specificity by targeting one of the two nucleotides present at the heterozygous base position; parentheses give optimized PCR annealing temperatures (35 cycles); and all amplifications utilized a final $MgCl_2$ concentration of 1mM, otherwise PCR cycling conditions were standard. EFRSM-817G (61°C): CTGAGCGGTAAAGTCAGAG; EFRSM-709C (62°C): ATTGTCACAGGGA ACGGCC; EFRSM-643T (64°C): GAGATTGGCCGGGGCGAAT; EFRSM-631G (62°C): GGCGAAAGTCACGACAGTG; EFRSM-619C (62°C): GACAGTTCCTGGC TTCAGC; EFRSM-496C (62°C): CAGAATGGCGTCCAGAGCC.

Loci AAm20 and AAm24

Non-fluorescent primers were adapted from Aransay *et al.* (2001) that sized the *P. perniciosus* microsatellites AAm20 and AAm24. Both loci showed little or no size variation in *P. ariasi*, and so were directly sequenced as anonymous nuclear DNA loci in *P. ariasi* and *P. mascittii*. PCR cycling conditions included (Aransay *et al.*, 2001): a 'hot start' at 80°C; an initial 5 min denaturation step at 94°C; first 5 cycles of denaturation 94°C for 30 sec, annealing 57°C for 40 sec, extension 72°C for 60 sec; 30

cycles annealing at 55°C; a final extension step of 72°C for 10 min; cooled and held at 4°C. For locus AAm20 a ca. 187 bp product was amplified by primers AAm20F2: CTGGTGGAGGGTGAGTTGAG and AAm20R2: ACAAGCGAGTCATAG AGTCCG. Two novel PASA primers were designed to resolve the allele composition of ambiguous genotypes, paired with conserved reverse primer AAm20R2 using 1mM $MgCl_2$, and one annealing temperature for 35 cycles; AAm20F-33G (66°C): AGTTGAGGCTTGCGTATCCG, and AAm20F-51C (66°C): CCCAGAGAGCGACG ACTC. For locus AAm24 a 170 bp product was amplified by primers AAm24F1: TCAATCGACATTCGGACAGGC, with AAm24R1: CTATTCCCGCCCCACTTGG. PCR cycling conditions were as stated for locus AAm20. PASA primers were designed to resolve ambiguous genotypes: conserved forward primer AAm24F1 paired with AAM24R-151C TATTCCCGCCCCACTTGGC (66°C 35 cycles; 0.7mM $MgCl_2$); and AAM24F-79G AGTTCAGCCGTCGCAGCAG (64°C 35 cycles; 1mM $MgCl_2$) paired with forward conserved primer AAm24R1.

PCR product purification

Two methods of PCR product purification were utilized, if PCR generated non-specific bands the targeted DNA fragments were fractionated by submerged agarose gel horizontal electrophoresis, excised and purified using GENECLEAN® II (BIOL 101 Qbiogene, Inc.). Millipore MultiScreen® PCR_{96} Filter Plates were used for higher throughput when PCR amplified specific products. (Protocols in Appendix 2.2).

Direct sequencing

Following purification nucleotide concentrations were estimated, for sequencing, using a photometric Nanodrop apparatus (Labtech International). Cycle sequencing was carried out on both strands using conserved primers and the single strand of PASA primers. 1/8 sequencing reactions were set-up on ice and carried out using the BigDye® Terminator v1.1 Cycle Sequencing Kit. A total 10 µl volume per reaction included: 2ng DNA per 100 bp of purified PCR product; 1pMol of a single sequencing primer (one direction only); 0.75x Big Dye Dilution Buffer (from kit); 1 µl Big Dye Terminator Mix (from kit). Thermal cycling at: 1 cycle of 96°C for 5 min; 25 cycles of 96°C for 10 sec, 50°C for 5 sec, 60°C for 4 min; cooled and held at 4°C. Dye terminators were removed by ethanol precipitation, and sequences were read using a 3730 capillary sequencer (Applied Biosystems).

2.2.3 Sequence editing and alignment

Sequence chromatograms of nucleotides were edited in SEQUENCHER™ v4.6 for Macintosh (Gene Codes Corporation), by manually correcting for errors introduced by the automatic processing of the ABI software. Ambiguity codes were scored where appropriate, these identifying genotypes whose allele composition required a PASA system for their resolution. Primers were removed for analyses and a consensus sequence (labelled with specimen and locus name) exported. Fully resolved (no nucleotide ambiguities) consensus sequences per locus were aligned in SEQUENCHER™ to permit the identification of unique sequences (haplotypes and alleles) for phylogenetic or population genetic analyses of *P. ariasi*. Composite sequence files for analyses were exported from SEQUENCHER™ in Nexus sequential format.

Where alignment required the insertion of gaps, nucleotide sequences were 'contigged' in SEQUENCHER™ and gaps manually inserted (by the author or P.D. Ready). Gap placement was either based on alignments from the literature, or following a rule to retain the locus' Open-Reading Frame(s) (preserving codons/amino acid units). BIOEDIT Sequence alignment Editor v7.0.9.0 for Windows (Hall, 1999) was used to translate nucleotide sequences to amino acids.

2.2.4 Methodology for allele inference

Most alleles in heterozygous genotypes could be deduced directly (for one dimorphic site) or by PASA (for > 1 dimorphic sites). However, it was not resource effective to resolve by PASA genotypes present in only 1-3 specimens. A compromise was made and in some cases deductive and inferred reasoning was applied to score alleles from ambiguous DNA direct sequences.

Alleles were inferred manually based on estimating the number of alternatives at a single locus (r^n, where r = number of alleles observed at each polymorphic site and n = number of observed polymorphic sites) and by using the following algorithm. Step 1 For each population, the allele and genotype frequencies were calculated based only on alleles read directly (homozygotes) or deduced (one dimorphic site). Step 2 These known alleles were aligned with each ambiguous sequence to identify any allele pairs that could constitute the latter. Step 3 The likelihood that each of these genotypes formed the ambiguous sequence was then ranked, based on the frequency of their alleles in the population. Often, the selected genotype contained one allele with high frequency

in the population and nearby (geographical regional bias) and one unrecorded allele. Step 4 If alternative genotypes were equally likely, then a TCS network of known alleles was used to identify by statistical parsimony the least derived unrecorded allele.

2.2.5 Data analyses

2.2.5.1 Phylogenetic reconstruction

Phylogenies were reconstructed for each locus to evaluate the phylogenetic species status of the *P. ariasi* under investigation, to identify discrete intra-specific *P. ariasi* lineages and their relative branching order, and to identify appropriate outgroups for selection tests. Outgroups to *P. ariasi* were molecularly characterized (Table 2.1) or were downloaded from GenBank among sequences available for *Phlebotomus* (listed in Appendix 2.3). The nomenclature of Esseghir *et al.* (2000) was followed: the *P. perniciosus* complex as *P. perniciosus*, *P. tobbi*, *P. orientalis*, *P. longicuspis*, *P. langeroni*; the *P. major* complex as *P. major*, *P. neglectus*, near *P. neglectus* (probably *P. syriacus*); the *P. ariasi* complex as morphologically identified *P. ariasi* caught in France and near *P. ariasi* from Tunisia. Near *P. ariasi* are likely to be *P. chadlii* where only males can be morphologically identified through a qualitative character of their genitalia, whereas females are morphologically indistinguishable or nearly so (Chamkhi *et al.*, 2006) (GenBank near *P. ariasi* accession number AF161196 was a female).

Phylogenetic trees were reconstructed by Bayesian estimation using MRBAYES v3.1.2 (Ronquist and Huelsenbeck, 2003; submitted online to http://cbsuapps.tc.cornell.edu/mrbayes.aspx). Nucleotide substitution models given the data were selected using the Akaike Information Criterion (AIC) approach in MRMODELTEST (v2.3; Nylander, 2004). Each analysis was run for 10 million generations with two parallel searches, using one cold and three heated Markov chains. Trees were sampled from each chain every 1000th generation and the first 5000 tree samples were discarded as burn-in. All other parameters of the Markov chain Monte Carlo (MCMC) run were left as default. Convergence of the two MCMCs onto a stationary distribution was assessed in the sump file (Convergence diagnostics: split frequency approaching zero, Potential Scale Reduction Factor (PSRF) of each model parameter approaching one). Frequent mixing of the two runs was assessed using the plot of the log likelihood values against generation. TRACER (v1.4.1; Rambaut and Drummond, 2007) was used to plot log likelihood values of the cold chain against generation to visualize the suitable burn-in point from the stationary phase at which the logarithm of the harmonic mean was

estimated. Bayes factors were then estimated and used as indicators of evidence for favouring the better of two models; the Bayes factor = 2 x (harmonic mean 1 - harmonic mean 2), where harmonic mean 1 and 2 are the more and less restrictive model, respectively. Higher log likelihoods (closer to zero) indicate a better model fit, significantly so when the Bayes factor is six units or greater (Kass and Raftery, 1995). Phylogenetic trees were viewed and edited in FIGTREE (v1.2.2; Rambaut, 2009).

Effects of alternative parameters on tree topology were tested using MRBAYES (v3.1.2) including: partitioning the data by codon position, 1st, 2nd and 3rd position independently, versus 1st + 2nd apart from 3rd position, versus no partitioning and each partition having independent preset commands for model priors; outgroup choice, where a probability value of 100 was given when multiple taxa were constrained as the outgroup; substitution model, as chosen by MRMODELTEST (v2.3) compared against the most parameterized model of GTR+I+G.

To compare topologies generated by alternative tree building algorithms, maximum likelihood (ML) and maximum parsimony (MP branch-and-bound search) were implemented on sequence datasets which were considered to reconstruct the 'best' Bayesian topologies. Rapid bootstrapping heuristics for ML (1,000 replicates) were conducted using RAxML (v7.0.4; Stamatakis *et al.*, 2008) through the CIPRES portal (v1.15; http://www.phylo.org/portal/Home), which has the advantage over other ML methods i.e. PHYML and GARLI, by not only having faster processing capabilities but also allows data partitioning, however, this is limited to a single nucleotide substitution model (GTR). MP groups taxa in the absence of a substitution model, where groups are formed by assuming that shared characters result from common descent. In PAUP* (v4.0b10; Swofford, 2002) each locus was partitioned by 1st, 2nd, 3rd codon positions weighted as 2:5:1, respectively. MP search parameters included: max trees set to 100; initial upper bound computed heuristically; furthest additional sequence; MulTrees in effect. Statistical support for trees generated was obtained by resampling using 1,000 bootstrap replicates.

2.2.5.2 Genealogical network reconstruction for *P. ariasi*

Haplotype (or allele) networks were constructed to represent the intra-specific gene genealogy for *P. ariasi* per locus. Networks are preferred over bifurcating phylogenetic trees to represent intra-specific data as they take into account population phenomena such as persistent ancestral nodes, multifurcations and reticulations (review

Posada and Crandall, 2001). Networks were reconstructed using statistical parsimony in TCS (v1.21 for Macintosh; Clement *et al.*, 2000), by inclusion of nucleotide sequences from all individuals that were connected based on single step-wise substitutions between haplotypes. A 95% parsimony connection limit was set to test whether *P. ariasi* formed a single haplotype network expected of a phylogenetic species (Hart and Sunday, 2007). Patterns of network reconstructions were used to examine the genealogical relationships between haplotypes, including signals of demographic processes.

2.2.5.3 Testing for reproductive isolation, panmixia and independent gene assortment

Random association of alleles within gametes were investigated to conclude that *P. ariasi* originates from a single random-mating population, not reproductively isolated groups (biological species). Adherence to Hardy-Weinberg equilibrium (HWE) at a single locus (ARLEQUIN v3.11; Excoffier *et al.*, 2005) and linkage disequilibrium (LD) across multiple unlinked loci (GENEPOP v4.0; Raymond and Rousset, 1995) were tested: assuming no evolutionary factors (e.g. selection, migration etc) influencing gametic frequencies. LD in GENEPOP tested for cyto-nuclear and nuclear-nuclear disequilibria. For estimating the standard error (Raymond and Rousset, 1995) and the probability of rejecting the null (no allele differentiation), Markov chain parameters included: dememorization = 10,000; batches = 10,000; iterations per batch = 5,000.

2.2.5.4 Testing for positive selection on molecular markers of *P. ariasi*

Based on the sequence information used, two classes of tests were implemented to detect positive directional or balancing selection. The first assessed the ratios or numbers of nonsynonymous to synonymous substitutions, which are powerful statistical methods for detecting molecular natural selection in protein-coding regions as they are often robust against demographic population processes. Whilst the third considered the allele frequency spectrum.

A species divergence approach was implemented in the CODEML program of Phylogenetic Analysis by Maximum Likelihood (PAML v4.2; Yang, 2007) that used the nonsynonymous/synonymous substitution rate ratio (d_N/d_S, denoted ω) as a measure of selective pressure at the protein level. Selection on the *P. ariasi* branch was tested for using a one-ratio null model which assumed a single ω across all branches versus a two-ratio model assuming a different ω_a for the *P. ariasi* branch free from the background

ω_o. Positive selection was inferred when $\omega_a > 1$, and model comparison showed significant heterogeneity in selection pressure by the likelihood (l) ratio test (LRT) where $2\Delta l = 2(l\omega_a - l\omega_o)$, compared to a χ^2 distribution with df = 1 at $P < 0.05$. Practically, the branch lengths of input gene trees (Bayesian derived) were re-estimated under ML in CODEML (model = 0; NSsites = 0, for the number of nucleotide substitutions per codon), and then used as initial values in further PAML analyses. In the control file transition/transversion rate ratio (k) was estimated; alpha was fixed at a constant rate. Anonymous nuclear markers AAm20 and AAm24 were not tested.

Divergence with polymorphism information was combined to conclude against selection within the *P. ariasi* branch. Here the McDonald-Kreitman (MK) (1991) population test for selection was implemented in DNASP (v4.90.1; Rozas *et al.*, 2003), whose neutral model predicts that the proportions of nonsynonymous (P_n) to synonymous (P_s) polymorphisms within a species are linearly related to the proportions of nonsynonymous (D_n) to synonymous (D_s) divergence between two species. Selection was inferred by significant departure from neutrality using a 2x2 contingency table of a two-tailed Fisher's exact test, and direction of selection indicated using the Neutrality Index (NI) (Rand and Kann, 1996); under neutrality NI = 1, where $D_n/D_s = P_n/P_s$; positive selection elevates D_n so that NI < 1; weak purifying and balancing selection suppress D_n but allow deleterious mutations to be found as polymorphisms so that NI > 1. Sensitivity of the MK test relies on the correct choice of outgroup (Wayne and Simonsen, 1998; Bellgard and Gojobori, 1999; Garrigan and Hedrick, 2003), which were selected based on d_N and d_S saturation levels estimated by an approximate per site model of Nei and Gojobori (1986) with a Jukes-Cantor correction (DNASP v4.90.1), and a more accurate (when substitution rate is close to saturation) ML method which incorporates an evolution model of substitution rates between codons (Goldman and Yang, 1994) [PAML CODEML: parameter settings: runmode -2; seqtype = 1; CodonFreq = 2; icode = 4 (insect mitochondrial) or icode = 0 (universal code); fix_kappa = 0; fix_omega = 0]. Per locus, an outgroup representative of each subgenus and species complex was compared with two *P. ariasi* alleles, the geographically most widespread and one distant from this.

Selection within *P. ariasi* at all loci was also investigated by neutrality tests (in ARLEQUIN v3.11) based on deviations from neutral expectation of the allele frequency spectrum using nucleotide data. Presence and direction of selection were detected using Tajima's *D* statistic (1989) based on the difference between estimators of θ_π and θ_S.

However, caution should be taken as demographic processes can reject the null hypothesis of population equilibrium and mimic selection. Balancing selection, population decrease, a recent bottle-neck and sub-division generate a positive *D*, by increasing θ_π (maintaining intermediate alleles) relative to θ_S. Directional selection (positive or purifying), selective sweep, expanding populations and a less recent bottle-neck generate a negative *D*, by raising the level of singletons (excess of low frequency alleles) which inflates θ_S relative to θ_π (Schmidt and Pool, 2002). Fu's F_s test (Fu, 1997), is more sensitive than Tajima's *D* for detecting population expansion and so was used to help distinguish between alternative conclusions. It uses a neutral coalescence model to estimate θ_π, and then calculates the probability of the number of haplotypes or alleles being greater than that observed in a sample of n. Negative F_s values arise from recent population expansions (or genetic hitchhiking) that produce an excess of low-frequency alleles. For all neutrality tests, significance from a null of neutrality was calculated using 16,000 coalescence simulations, and significant *P*-values of multiple tests (Rice, 1989) were manually corrected for familywise Type 1 errors by applying a sequential Bonferroni correction ($\alpha = 0.05$) (Holm, 1979). As neutrality statistics can be affected not only by selection but also recombination, the latter was estimated as the minimum number of recombination events (Rm) using the four gamete model (Hudson and Kaplan, 1985) in DNASP (v4.90.1). Mismatch distributions based on information from the distribution of the pairwise sequence differences was also implemented to detect the alternative of population expansion (see next section for full description).

2.2.5.5 Population genetic analyses

Descriptive population statistics

ARLEQUIN (v3.11) was used to estimate number of segregating sites (S), number of haplotypes (*h*), haplotype diversity (H_d; Nei, 1987), nucleotide diversity (π; Nei, 1987), and departure of genotype distributions from HWE (exact test; method of Guo and Thomson, 1992). Demographic processes e.g. population bottle-necks and expansions, can be inferred from such statistics.

Population structure

Pairwise F_{ST}. Estimates of the pairwise population parameter F_{ST} (Wright, 1951) were used to measure the extent of genetic differentiation, by deviations in observed heterozygosity, between populations based only on haplotype or allele frequencies: obtained for each nuclear locus according to the exact test of Weir and Cockerham

(1984) in FSTAT v2.9.3.2 (Goudet, 2002) which is unaffected by sampling scheme, deriving significance levels through 1,000 permutations and a sequential Bonferroni correction; for cyt b conventional F_{ST} was estimated in ARLEQUIN (v3.11) significance generated using 1,000 permutations. Among interbreeding populations F_{ST} reflects the opposing processes of random genetic drift (population differentiation) and gene flow (population homogenization): F_{ST} values close to zero support migration between populations, whereas near to one indicates no migration and the increasing divergence effects of drift. Crudely according to Wright (1978) extent of genetic differentiation between population gene pools can be categorized into four F_{ST} value classes: F_{ST} 0-0.05, "little"; F_{ST} 0.05-0.15, "moderate"; F_{ST} 0.15-0.25, "great"; and F_{ST} > 0.25 "very great".

Analysis of Molecular Variance. Support for *a priori* and *post-hoc* population sub-division was tested using hierarchical AMOVA (ARELQUIN v3.11; Excoffier *et al.*, 1992), AMOVA estimates Φ-statistics and variance components, which reflect the proportion of molecular variability of haplotypes at different levels of sub-division (hierarchies): among regions, among populations within regions, and within populations of regions. Probability of having more extreme Φ-statistics and variance component than observed by chance alone (a null of global panmixia at the different hierarchical levels) was tested under 16,000 random permutations.

Isolation-by-distance (IBD). As the dispersal capability of *P. ariasi* is limited (Killick-Kendrick *et al.*, 1984) dependence between genetic distances with geographical proximity between population pairs per locus was sought, using the principle of isolation-by-distance (IBD) (Wright, 1943). Regression of genetic distance was fitted to estimates of geographical distances according to the method of Rousset (1997), and non-parametric significance of association between the distance matrices was implemented within the ISOLDE suboption of GENEPOP (v4.0): 1,000 permutations for a Mantel test where the null states a regression line of zero or independence between the two distance matrices. Genetic distance was based on F_{ST} values (calculated as detailed above), and straight-line geographical distances between populations were measured from a digital map using the Distance and Azimuth Matrix v2.1 extension (Jenness, 2005) within ARCVIEW (v3.2; ESRI). No assumptions on the dimension of dispersal were made which can affect the correlation between genetic and geographical distance (Kimura and Weiss, 1964). Results were therefore reported for both $F_{ST}/(1-F_{ST})$ against geographical distance (one-dimensional habitats) or logarithm of geographical

distance (two-dimensional habitats) (Rousset, 1997). Regression outliers were identified, using a z-test, as those falling more than three standard deviations from the mean (PASW Statistics v18).

Distance-based redundancy analysis (dbRDA). Multiple regression analysis was implemented in DISTLM (v5.0; Anderson, 2004) to test the affect on genetic distance of geographical distance and of geographical region. Marginal tests were implemented for each of the predictor variables, and conditional tests were performed in which geographical distance was included as a covariate in the model. The latter allowed the examination of the extent to which regionality explained the variation in genetic distance beyond that of IBD, to identify the presence of barriers to dispersal or fragmentation. Significance was obtained using 999 unrestricted permutations of rows and columns of the matrices for all variables.

Identifying and dating demographic events

Mismatch distribution of cyt b haplogroups. Evidence of sudden demographic population expansion of haplogroups, and haplogroups by geographical regions, were investigated using mismatch distributions; the plot of the number of pairwise differences between haplotypes based on estimated inter-haplotypic distances as calculated in ARLEQUIN (v3.11) using the pairwise distance option. The mismatch distribution is unimodel or multimodal during demographic expansion or population size equilibrium, respectively, the latter reflecting the highly stochastic shape of a haplotype gene tree (Slatkin and Hudson, 1991; Rogers and Harpending, 1992). The Raggedness index (Harpending, 1994) was used to test for statistical support of sudden demographic expansion using 10,000 bootstrap replicates under the null model of expansion. This study is aware that mismatch distribution can be a conservative method to detect sudden population expansion (Ramos-Onsins and Rozas, 2002), however, it was considered useful to implement as expansion events are not only detected but dated. Estimates of time since the beginning of sudden demographic expansion events used the mode of an observed Poisson mismatch distribution as expressed by the parameter $\tau = 2ut$. For DNA sequence data u is the mutation rate per generation for the whole sequence and t is the number of generations elapsed since the beginning of the expansion event (Rogers, 1995). 95% confidence intervals of τ were estimated around two mutation rates, 2.3% and 1%, at $\alpha = 0.05$ (see below).

MDIV analysis of cyt b haplogroups. One aim was to date the demographic events that produced the founders of current cyt b haplogroups of *P. ariasi*. The basic isolation with migration coalescence model in MDIV (Nielsen and Wakely, 2001; http://cbsuapps.tc.cornell.edu/mdiv.aspx) implements both likelihood and Bayesian methods that were used to jointly estimate posterior distributions of θ (scaled parameter for nucleotide heterozygosity, $2N_{fe}\mu$), M (scaled migration rate), T (scaled gene divergence time, t/N_{fe}) and TMRCA (estimated Time to the Most Recent Common Ancestor or gene coalescence time, tMRCA/N_{fe}) among pairs of cyt b haplogroups. As MDIV can estimate migration rates, it was also used as an indirect method to confirm demographically independent/genetically (not reproductively) isolated cyt b haplogroups (e.g. Smith and Farrell, 2005), inferred when the mode of the posterior distribution of migration rate intersected the Y-axis (was zero).

A Markov chain length of $3x10^6$ with a burn-in of 10% was used. Model priors included: nucleotide substitution according to the finite site model to account for multiple hits (HKY, Hasegawa *et al.*, 1985), and preliminary M_{max} and T_{max} were set at various values to select the final optimal priors for each haplogroup pairwise comparison. Optimal values of the priors M_{max} and T_{max}, and the shortest credibility intervals were determined as those values that generated a bell-shaped posterior distribution with the minimum number of estimators on the right-hand tail of the distribution. A minimum of three replicate Markov chains using different random seeds were run with the optimal values of M_{max} and T_{max}, to check for convergence and consistency of the parameter estimates, and their outputs were averaged and plotted, from which estimators of the parameters θ, M and T were determined based on the maximum posterior probability/highest likelihood values - read as the mode of the estimator's posterior probability distribution. The estimators of T and TMRCA (given by MDIV) were converted to years (for a haploid genome) according to Nielsen and Wakely (2001); $t = (T\theta/2\mu)$, and tMRCA = TMRCA $\theta/2\mu$); where μ is the mutation rate for the whole cyt b sequence per year per generation.

Rates of cyt b divergence. Dating used two rates of pairwise divergence of cyt b: 1% per million years (p.m.y.) upper limit (Esseghir *et al.*, 2000) and 2.3% p.m.y. lower limit (Brower, 1994). Two generation times for *P. ariasi* were used: 1 generation per annum (p.a.) lower limit and 3 generations p.a. upper limit (Ready and Croset, 1980).

2.3 Results

2.3.1 Phylogenetic reconstruction

2.3.1.1 Cyt b

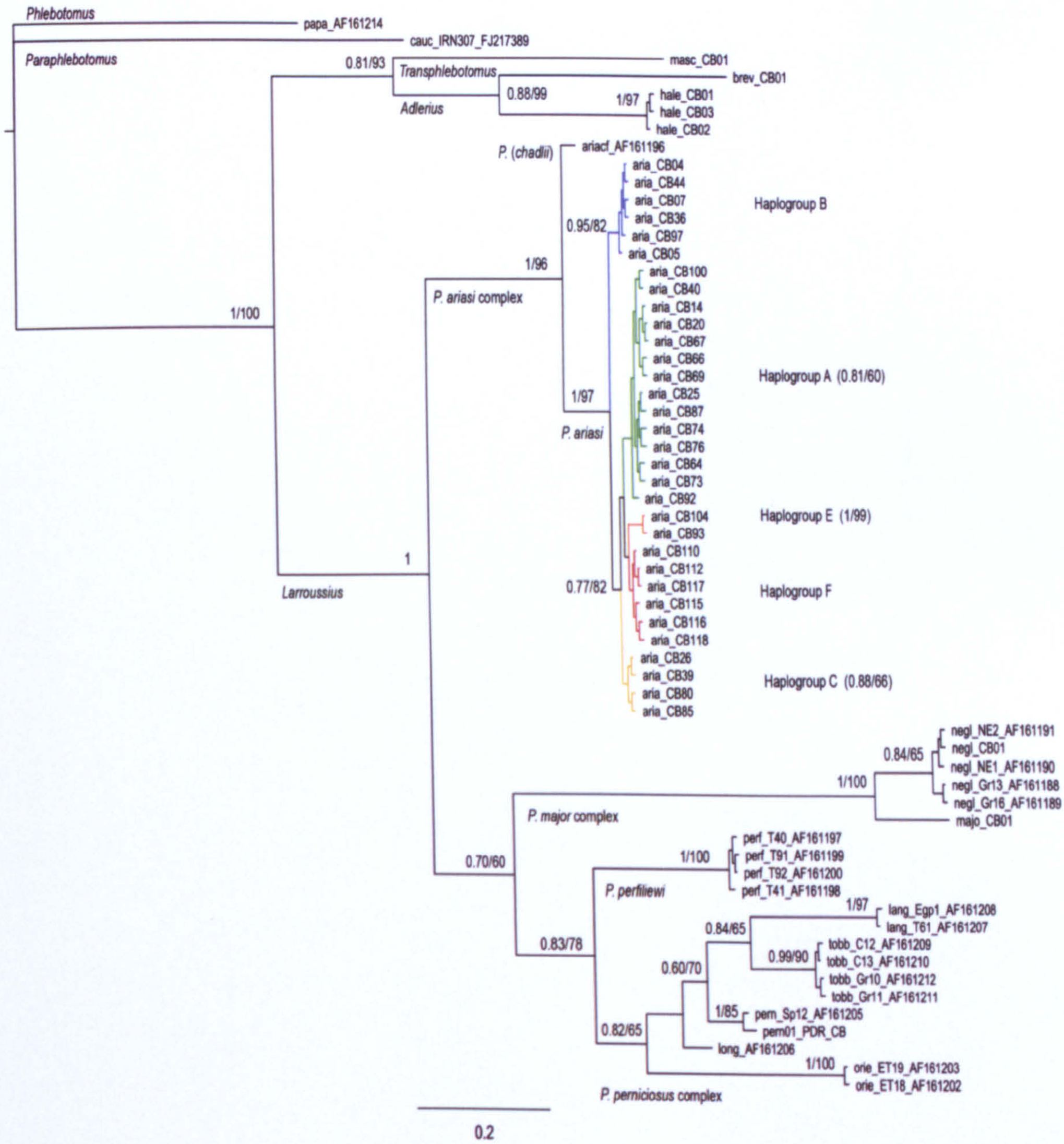
In addition to *P. ariasi,* haplotypes were isolated from other species: 1 *Phlebotomus (Transphlebotomus) mascittii*, 1 *Phlebotomus (Adlerius) brevis*, 3 *Phlebotomus (Adlerius) halepensis*, 1 *Phlebotomus (Larroussius) neglectus* and 1 *Phlebotomus (Larroussius) major*. Phylogenies used the cyt b coding region only; 714 bp (no indels) starting on base position 10,918 of *Drosophila melanogaster* (NCBI Reference Sequence: NC_001709.1).

With single species of the subgenera *Phlebotomus* and *Paraphlebotomus* as the outgroup, the Bayesian phylogeny (Figure 2.2) of both new and GenBank haplotypes of cyt b gave strong support (posterior probability, pp, 0.7-1) for: the monophyly of all taxonomic species; two subgenera *Transphlebotomus* and *Adlerius* as outgroups to *Larroussius*, where the former shared a more recent common ancestor to *Larroussius*; the monophyly of subgenus *Larroussius* with a branching order of *P. ariasi* complex, *P. major* complex, and *P. perfiliewi* sister to the *P. perniciosus* complex. Within the *P. perniciosus* complex *P. orientalis*, *P. perniciosus* and *P. longicuspis* formed a basal polytomy, support only being given to *P. tobbi* and *P. langeroni* as sister taxa (pp 0.84). *P. ariasi* was monophyletic (pp 1) with the female near *P. ariasi* – GenBank accession AF161196, likely to be *P. (Lr.) chadlii* based on the location in Tunisia (Esseghir *et al.*, 2000) and the large genetic distance between its cyt b haplotype and those of both males and females of *P. ariasi* – as its sister species which branched first in *Larroussius.*

Phylogeny reconstruction using ML was not strictly concordant with the Bayesian topology, discrepancies identified as: the *P. major* complex unresolved position within *Larroussius* (60%), and the unresolved phylogenetic relationship of individual species of the *P. perniciosus* complex (node bootstrap ≤ 70%; ML bootstrap values are given after Bayesian pp, on Figure 2.2). For MP a pruned dataset of 15 species, a single haplotype per species containing 201 parsimoniously informative sites, was concordant with the Bayesian phylogeny except for low support for the monophyly of *Larroussius* (60%) and not resolving the *P. perniciosus* complex within it.

Evidence supporting cryptic speciation was recorded within *P. ariasi*, for which a pruned dataset of haplotypes showed a well supported primary bifurcation of a basal

Figure 2.2 Bayesian phylogeny of the haplotypes of the 3′ end of cyt b (714 bp) from *Phlebotomus* species. Branches for subgenera, species complexes, some species, and the haplogroups of *P. ariasi* (aria) are labelled. Haplotypes obtained from GenBank contain the accession number in their code. Codes for unlabelled species: papa: *P. papatasi*; cauc: *P. caucasicus*; masc: *P. mascittii*; brev: *P. brevis*; hale: *P. halepensis*; ariacf: *P. chadlii*; negl: *P. neglectus*; majo: *P. major*; lang: *P. langeroni*; tobb: *P. tobbi*; pern: *P. perniciosus*; long: *P. longicuspis*; orie: *P. orientalis*; perf: *P. perfiliewi*. Cyt b was partitioned by each codon position, each with an independent substitution model selected by MRMODELTEST v2.3. Node values and to the right of haplogroups A, C, E, represent posterior probabilities/ML % bootstrap values, support for node when > 0.7/70%. Scale bar = substitutions per site.



European haplotype B (pp 0.95) from a macrohaplogroup[1] A (pp 0.77). The latter contained three supported European haplogroups, C (pp 0.88), A (pp 0.81) and E (pp 1) but a poorly supported branching order (pp < 0.7) and one unsupported haplogroup F (pp 0.28), which represented the entire population from Morocco (AGH). ML showed concordant primary grouping within *P. ariasi*.

The above described Bayesian phylogeny was statistically supported as using the 'best' model to fit the data based on Bayes factor values > 6 units (Appendix 2.4): when partitioning the data by 1st, 2nd and 3rd position (Cyt b_bayes1; Appendix 2.4) versus 1st + 2nd ≠ 3rd (Cyt b_bayes2) or no partition (Cyt b_bayes3) [harmonic means, -4262.10, -4349.48 and -4553.92, respectively]; and model selection by MRMODELTEST (Cyt b_bayes1) favoured against overparameterizing with the GTR+I+G model (Cyt b_bayes6), Bayes factor 21.78 units. Outgroup choice affected topology, but only with respect to the sister subgenus to *Larroussius*: inclusion of both *Phlebotomus* and *Paraphlebotomus* is described above; *P. caucasicus* (*Paraphlebotomus*) only, supported (pp 0.99) *Transphlebotomus* as the sister to *Larroussius* (Cyt b_bayes5); *P. papatasi* (*Phlebotomus*) only, supported neither *Transphlebotomus* nor *Adlerius* (Cyt b_bayes4). Within *Larroussius* all outgroup combinations listed above supported the basal position of *P. ariasi* complex in *Larroussius*, followed by the *P. major* complex in the cyt b gene tree.

2.3.1.2 EF-1α

A Bayesian phylogeny based on a short fragment of the nuclear EF-1α gene (453 bp; 3 subgenera; 14 species), showed (genealogical) discordance with the mitochondrial cyt b gene tree: EF-1α did not support either *Transphlebotomus* or *Adlerius* as sister to *Larroussius* (Figure 2.3a); *P. major* complex as the basal group within the *Larroussius*; no support for the sister status of *P. perfiliewi* to the *P. perniciosus* complex, instead nesting the former within the latter (pp 1); branching order within the *P. perniciosus* complex differed. The EF-1α short fragment did maximally support (pp 1) the monophyly of the *P. ariasi* complex, but failed to resolve *P. ariasi* from near *P. ariasi* or any intra-specific groups (pp < 0.7). To partition the data by 1st, 2nd and 3rd position using different nucleotide substitution models given by MRMODELTEST for each position was given as the best model to fit the data through

---

[1] "Macrohaplogroup: a group of haplogroups that are closely related and share a recent common ancestor". Shriver and Kittles, 2004.

the Bayes factor approach. ML (bootstrap values given in Figure 2.3a) and MP (82 parsimoniously informative sites) were less resolved than the Bayesian topology but not discordant, and neither supported discrete intra-specific grouping within *P. ariasi*. With respect to outgroup choice: when *Adlerius* was the outgroup to *Larroussius* the *P. major* complex was not supported within *Larroussius*, also observed for both ML and MP methods. Where *P. mascittii* was the only outgroup, *P. major* complex grouped within *Larroussius* but as a basal polytomy (not *P. ariasi*, as consistently seen in cyt b). The *P. major* complex was supported as the outgroup to all other *Larroussius* when it was so designated (pp 0.97), within which the *P. ariasi* and *P. perniciosus* complexes were monophyletic (pp 1) and sister to one another, however, *P. perfiliewi* nested within the latter.

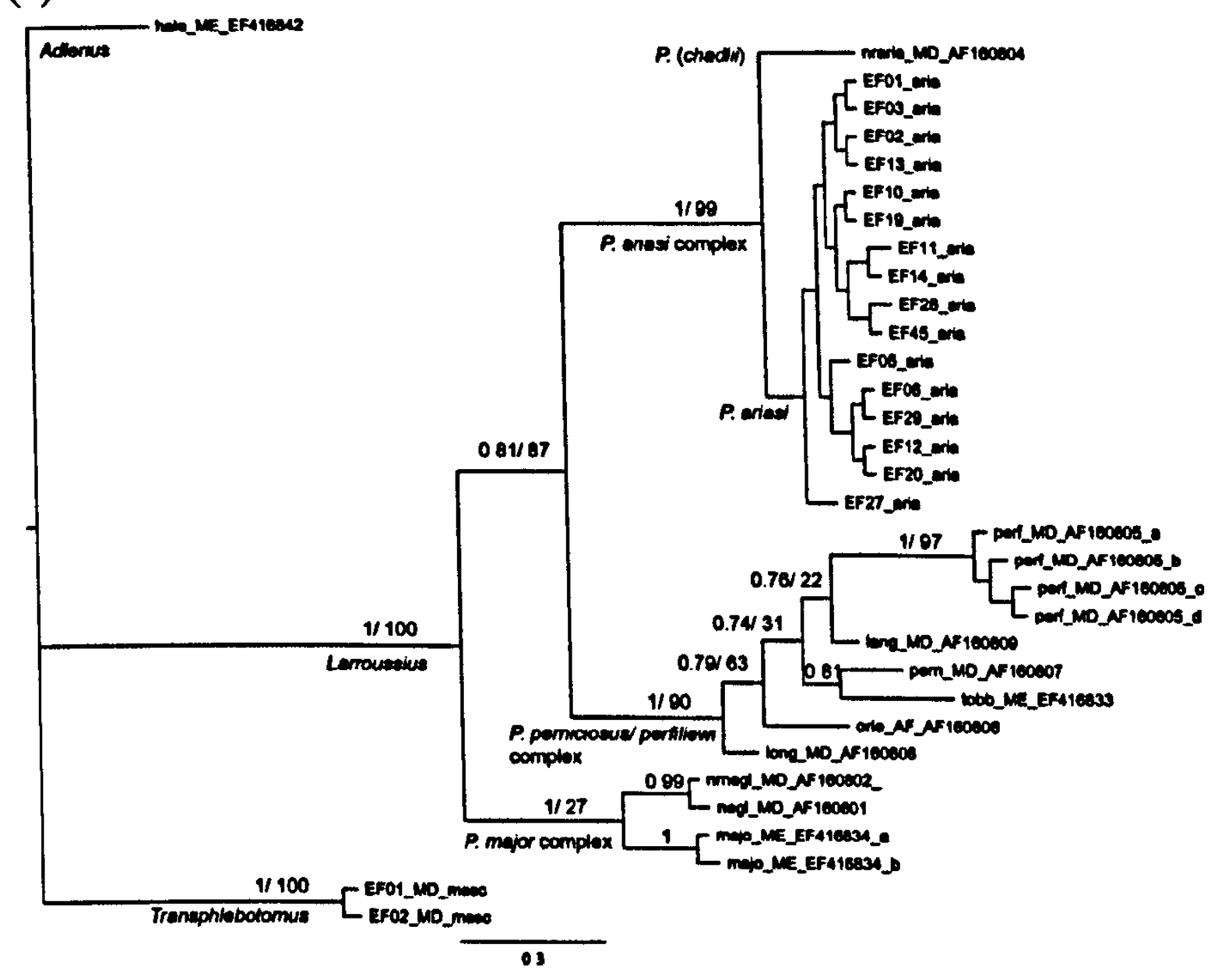
Phylogenetic reconstruction using the long EF-1α fragment (720 bp), was obtained for 9 *Larroussius* species (Figure 2.3b), where *P. neglectus* and near *P. neglectus* (likely to be *P. (Lr.) syriacus*) were not rejected as the outgroup to all other *Larroussius* (pp 1). This study showed that the longer EF-1α sequence is necessary to resolved the monophyly of the *P. perniciosus* complex with *P. perfiliewi* as its sister group (pp 1), but no intra-specific groups for *P. ariasi* were observed, and near *P. ariasi* remained unsolved from *P. ariasi*. ML and MP gave concordant support to this Bayesian tree.

2.3.1.3 AAm20 and AAm24

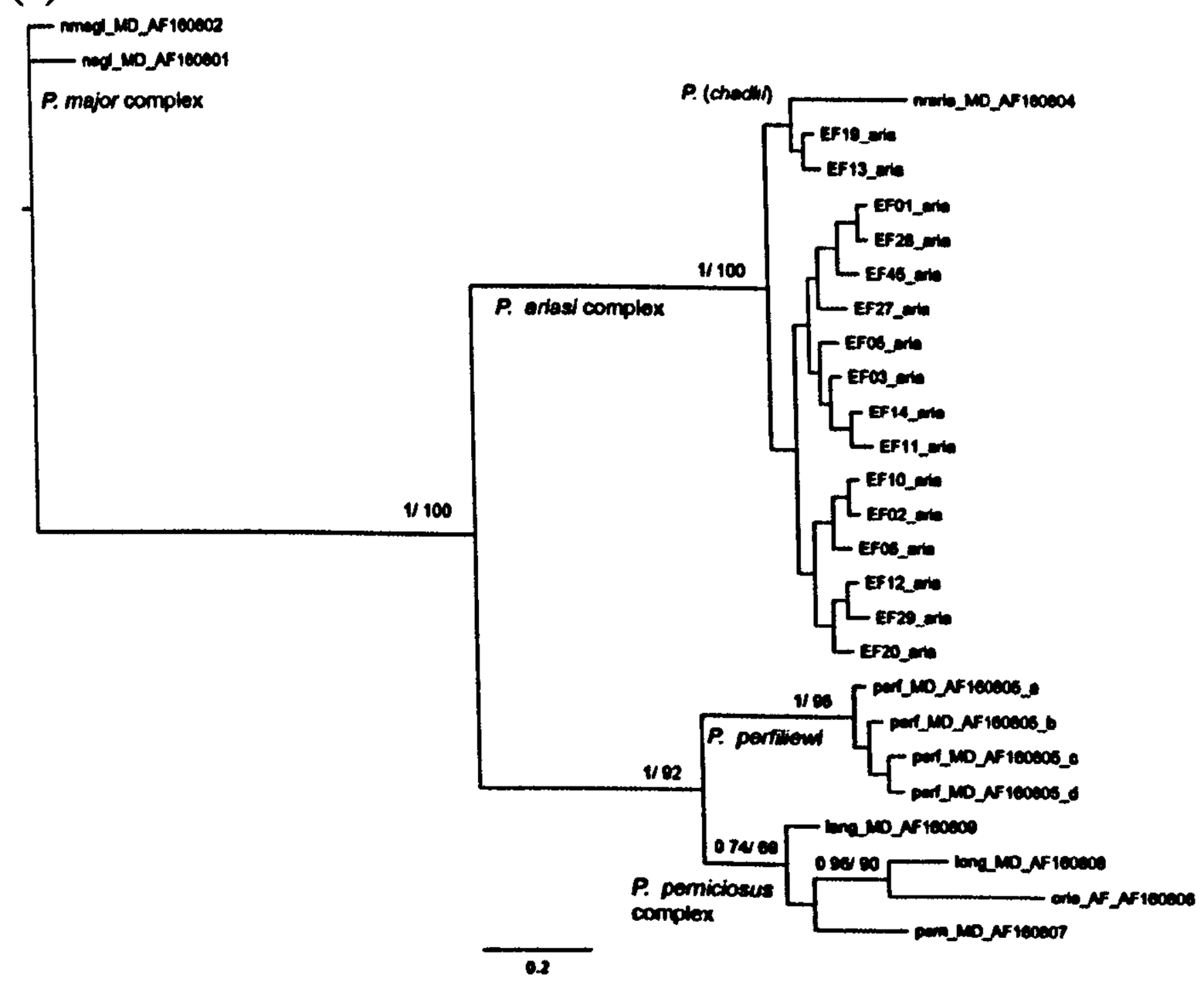
For both anonymous nucleotide phylogenies, *P. mascittii* was designated as the outgroup, where phylogenies were based on manually inferred inter-species alignments: AAm20 149 bp Appendix 2.5 AAm24 175 bp Appendix 2.6). Both loci were phylogenetically uninformative inter-specifically and intra-specifically, a consequence of either limited outgroup choice or low sequence variation in their short sequences (phylogenies not shown). For locus AAm20 Bayesian estimation using a non-partitioned codon model supported *P. ariasi* as a single monophyletic group (pp 0.82), but not *Larroussius*, and ML and MP were unresolved phylogenies (all nodes < 70%). For locus AAm24 only MP (not Bayesian or ML) supported the monophyly of *P. ariasi* (bootstrap 80%), but it failed to support *Larroussius* as an ingroup to *P. mascittii*.

Figure 2.3 Bayesian phylogeny from *Phlebotomus* species of the haplotypes of elongation factor-1 alpha (a) short (453 bp) and (b) long (720 bp) fragment. Branches for subgenera, species complexes, some species, and the haplogroups of *P. ariasi* (aria) are labelled. Haplotypes obtained from GenBank contain the accession number in their code. Codes for unlabelled species: hale: *P. halepensis*; masc: *P. mascittii*; nraria: *P. chadlii*; negl: *P. neglectus*; nrnegl: *P. syriacus*; majo: *P. major*; long: *P. longicuspis*; orie: *P. orientalis*; tobb: *P. tobbi*; pern: *P. perniciosus*; lang: *P. langeroni*; perf: *P. perfiliewi*. EF-1α was partitioned by 1st, 2nd, 3rd codon position, each with an independent substitution model selected by MRMODELTEST (v2.3). Node values represent posterior probabilities/ML % bootstrap values, support for node when > 0.7/70%. Scale bar = substitutions per site.

(a)



(b)



2.3.1.4 Phylogenetic inference using statistical parsimony networks

Species-specific TCS networks were reconstructed based in the 95% parsimony connection limit for cyt b (714 bp), EF-1α (long 720 bp), AAm20 (149 bp) and AAm24 (175 bp), which supported the monophyly of *P. ariasi* and the absence of cryptic speciation. The EF-1α short fragment (453 bp) failed to distinguish near *P. ariasi* (*P. chadlii*) from *P. ariasi* these connected by a minimum of nine mutational steps in a single network, confirming this marker as unsuitable to resolve cryptic species of *Phlebotomus*.

2.3.2 Intra-specific locus description

2.3.2.1 Cyt b

The 745 bp (excluding primers) fragment was sequenced for 452 *P. ariasi* from all 19 populations: 01-715 bp cyt b; 716-718 bp stop TAA; 719-725 bp IgS; 726-745 bp tRNA. For population genetic analyses, missing data at the 5′ terminus putatively lacking segregating sites) were excluded reducing the sequence length analyzed to 738 bp. Pairwise-distance analysis identified 94 unique haplotypes defined by 89 segregating sites (Appendix 2.7); 76 transitions and 13 transversions and an overall relative nucleotide composition of C: 16.96%, T: 41.26%, A: 31.93%, and G: 9.85%. Two base positions (one in each of the IgS and tRNA) showed degeneracy of more than two alternative nucleotides; a 'D' ambiguity (G, T or A nucleotide) among two or more flies. The uninterrupted ORF of cyt b and the lack of heteroplasmy indicated the absence of pseudogenes.

2.3.2.2 EF-1α

The 817 bp (excluding primers) fragment of EF-1α was sequenced for 403 *P. ariasi* from 18 (out of 19) populations (not ROQ). PCR amplification was not successful for *P. (Tr.) mascittii* therefore to reconstruct an EF-1α short fragment phylogeny a nested PCR amplified a 454 bp fragment using primers EF-F05/EF-R08 (Parvizi and Assmar, 2007). For population analyses missing data were excluded (5′ 22 bp and 3′ 18 bp), reducing the long fragment analyzed to 777 bp with the putative loss of no segregating sites. Based on 29 segregating sites, 22 transitions and 7 transversions, 45 alleles (Appendix 2.8) and 65 genotypes (Appendix 2.9) were scored. 39 genotypes were resolved through the PASA method, and only 4 alleles [Morocco (2), Pyrenean France (2)] and 7 individuals' genotypes [Morocco (4), Pyrenean France (3);

1.7%], were inferred using the allele scoring algorithm described in section 2.2.4. Relative nucleotide compositions were C: 24.53%; T: 22.38%; A: 22.54%; G: 30.55%. In *P. ariasi* no nonsynonymous changes were observed for EF-1α in a single intronless ORF, confirming that the conserved primer pair amplified a single-copy orthologue gene sequence.

2.3.2.3 AAm20

A ca. 146 bp (excluding primers) fragment was sequenced from 396 *P. ariasi* from 18 populations (not ROQ), and 2 *P. mascittii*. Reading from the conserved forward primer, size variation was recorded in only 12 *P. ariasi* that were always either within [France, Pyrenees 8 flies and Massif Central 2 flies] or flanking the microsatellite region [1 fly each from outgroup population Portugal and Morocco] identified for *P. perniciosus* (Aransay *et al.*, 2001). Using the entire sequence fragment 13 *P. ariasi* alleles were scored – 146 bp not including size variants – whose alignment with the *P. mascittii*, showed their nesting within the 395 bp clone isolated from *P. perniciosus* (AJ303377) starting at 110 bp. This inter-species alignment (Appendix 2.5) required gap insertions in all three species to maintain a single ORF. BLAST searches (BLASTp, BLASTn and BLASTx; http://blast.ncbi.nlm. nih.gov/Blast.cgi) with *P. ariasi* and *P. mascittii* alleles found only GenBank sequence AJ303377 *P. perniciosus* with significant homology (E-value > e^{-5}).

90 bp of locus AAm20 was used for population genetic analyses of *P. ariasi*; 5′ 44 bp containing size variant indels (eliminating two low frequency segregating sites in three *P. ariasi* only) and 12 bp of missing data at the invariable 3′ terminus were excluded. 13 sites were segregating, 10 transitions and 3 transversions, and a relative nucleotide composition of C: 44.36%, T: 20.07%, A: 14.95%, and G: 20.61%. 14 alleles (Appendix 2.10.) paired to score 19 genotypes, 4 alleles [NW Spain 1, Pyrenean France 2, S Massif Central 1] and 5 genotypes had to be inferred for 7 individual's (1.8%) (Appendix 2.11).

2.3.2.4 AAm24

The 130 bp (excluding primers) fragment of locus AAm24 was sequenced from all 398 *P. ariasi* from 18 populations (not ROQ) and 2 *P. mascittii*. Manual gap insertion permitted the alignment of *P. ariasi* alleles with a new one of *P. mascittii* and that of *P. perniciosus* (GenBank sequence AJ303378), to produce a 175 bp ORF

(Appendix 2.6). Four nonsynonymous and three synonymous changes upstream of the microsatellite and 15 synonymous sites downstream (nucleotides 110-175) were observed. A BLAST search of GenBank detected 13 continuous amino acids matched 100% the "Jumonji domain containing 1B" of *Nasonia vitripennis* (LOC100120387), and up to 15 amino acids (with indels) matching the "Jumonji domain containing 1B" of both *Nasonia vitripennis* and *Apis mellifera* (LOC408944). The Jumonji protein belongs to a family of transcription factors with homologues in mouse and *Drosophila* (Jung *et al.*, 2005; Sasai *et al.*, 2007).

For *P. ariasi* size variation of alleles was not recorded. Paired combinations of 11 deduced alleles (Appendix 2.12) scored the 21 genotypes recorded (Appendix 2.13), of which only a single genotype required inference constituted by two known alleles [3 *P. ariasi* from Pyrenean France and 1 fly from NW Spain; 1%]. Exclusion of missing data, invariable 5′ 9 bp, reduced the fragment analyzed to 121 bp in population genetic analyses of *P. ariasi*. 5 segregating sites were recorded, 4 transitions, 1 transversion and a relative nucleotide composition of, C: 26.43%, T: 17.36%, A: 32.04%, G: 24.17%.

2.3.3 Haplogroups, gene networks and geographical variation of *P. ariasi*

The cyt b parsimony network (Figure 2.4) showed five clusters of haplotypes, four of which matched those phylogenetically supported (A-C, E pp > 0.81), and the remaining haplotypes were those of unsupported haplogroup F (Morocco *P. ariasi*). Only haplogroup B had no reticulate loops with any other haplogroup, the others had multiple most parsimonious pathways connecting them confirming the polytomy observed in the Bayesian phylogeny of macrohaplogroup A. The designation of the five haplogroups was strengthened by pairwise distance values, estimated using the Maximum Composite Likelihood approach in MEGA (v4.0; Tamura *et al.*, 2007); all means of within-haplogroup distances (0.000952-0.002918) were less than those of between-haplogroup distances (0.008758-0.021846).

To maximize intra-specific evaluation, networks for nuclear genes used population datasets: EF-1α 777 bp; AAm20 90 bp; AAm24 121 bp (Figures 2.5 to 2.7). No nuclear marker showed discrete lineages in their genealogical networks. EF-1α showed a web-like network linking all alleles, with numerous reticulate loops around two high frequency central modes in Europe (alleles 01, 03) and others in Morocco (allele 29). Loci AAm20 and AAm24 each had fewer haplotypes, simpler networks (0-1 loops), and two modes with shallow radiations except in Portugal (AAm20) or Morocco

Figure 2.4 Parsimony network (TCS v1.21) showing the genealogical relationships between the 92 cyt b (length 714 bp) haplotypes from 452 *P. ariasi*, with a 11 step 95% connection limit. These haplotypes are shown as coloured circles with sizes proportional to their frequency of occurrence, which is given if > 5. Black filled circles denote missing haplotypes. The six lettered haplogroups or sub-haplogroups (B) are followed by the code of their modal haplotype (CBNN) along with their geographical distributions. All most parsimonious pathways are shown.

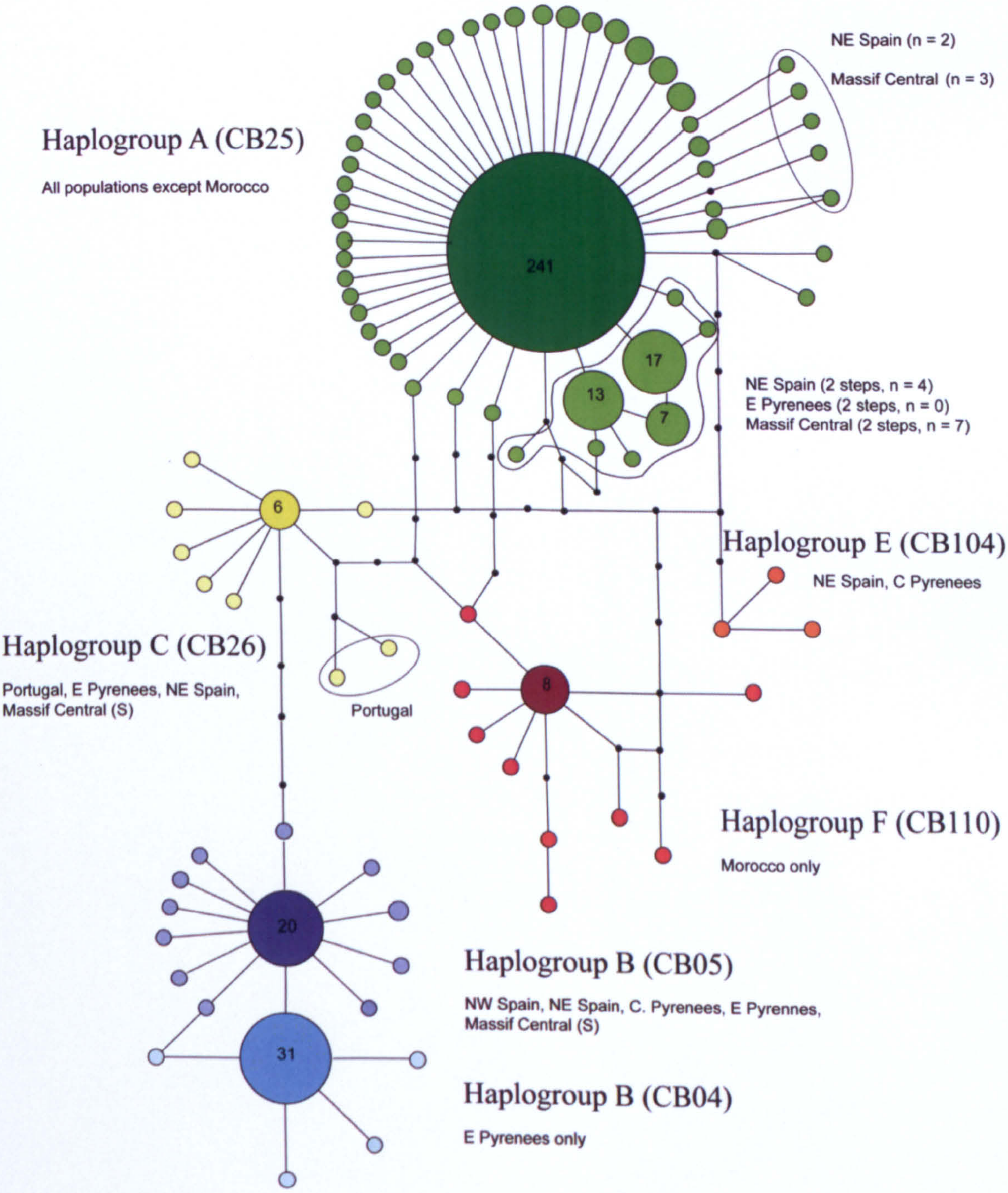
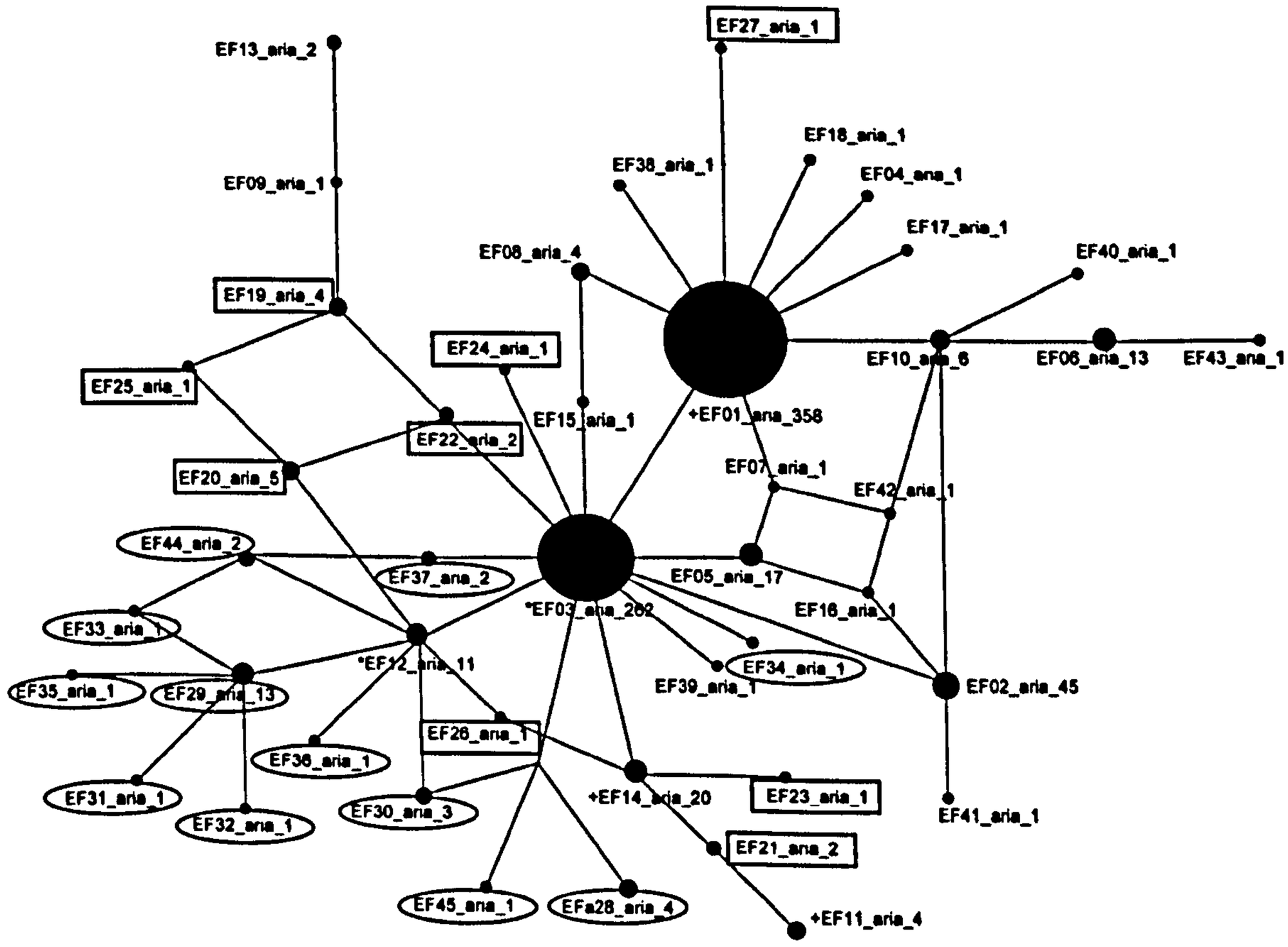


Figure 2.5 Parsimony network (TCS v1.21) showing the genealogical relationships between the 45 EF-1α alleles (length 777 bp) from 403 *P. ariasi*, with a 12 step 95% connection limit. Haplotypes are shown as black filled circles with sizes proportional to their frequency of occurrence, which is given by the number after the allele code (aria_NN). Alleles in boxes and ellipses private to Portugal and Morocco, respectively. *Alleles found in Morocco, Portugal and others; + alleles found in Portugal and others, but absent in Morocco.



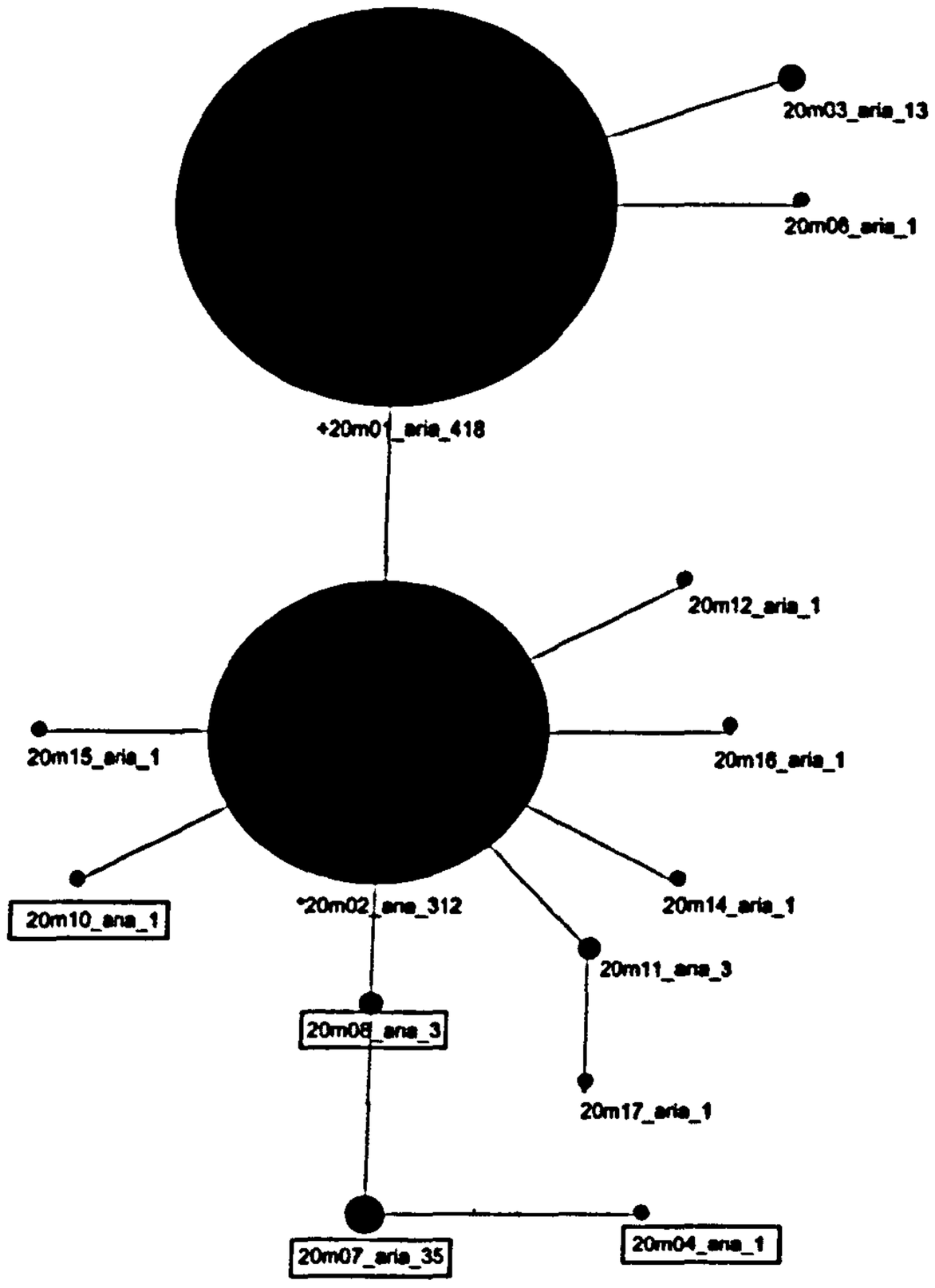


Figure 2.6 (left) Parsimony network (TCS v1.21) showing the genealogical relationships between the 14 AAm20 alleles (length 90 bp) from 396 *P. ariasi*, with a 3 step 95% connection limit. Haplotypes are shown as black filled circles with sizes proportional to their frequency of occurrence, which is given by the number after the allele code (aria_N). Alleles in boxes private to Portugal; * alleles found in Portugal, Morocco and others; + alleles found in Portugal and others, but absent from Morocco.

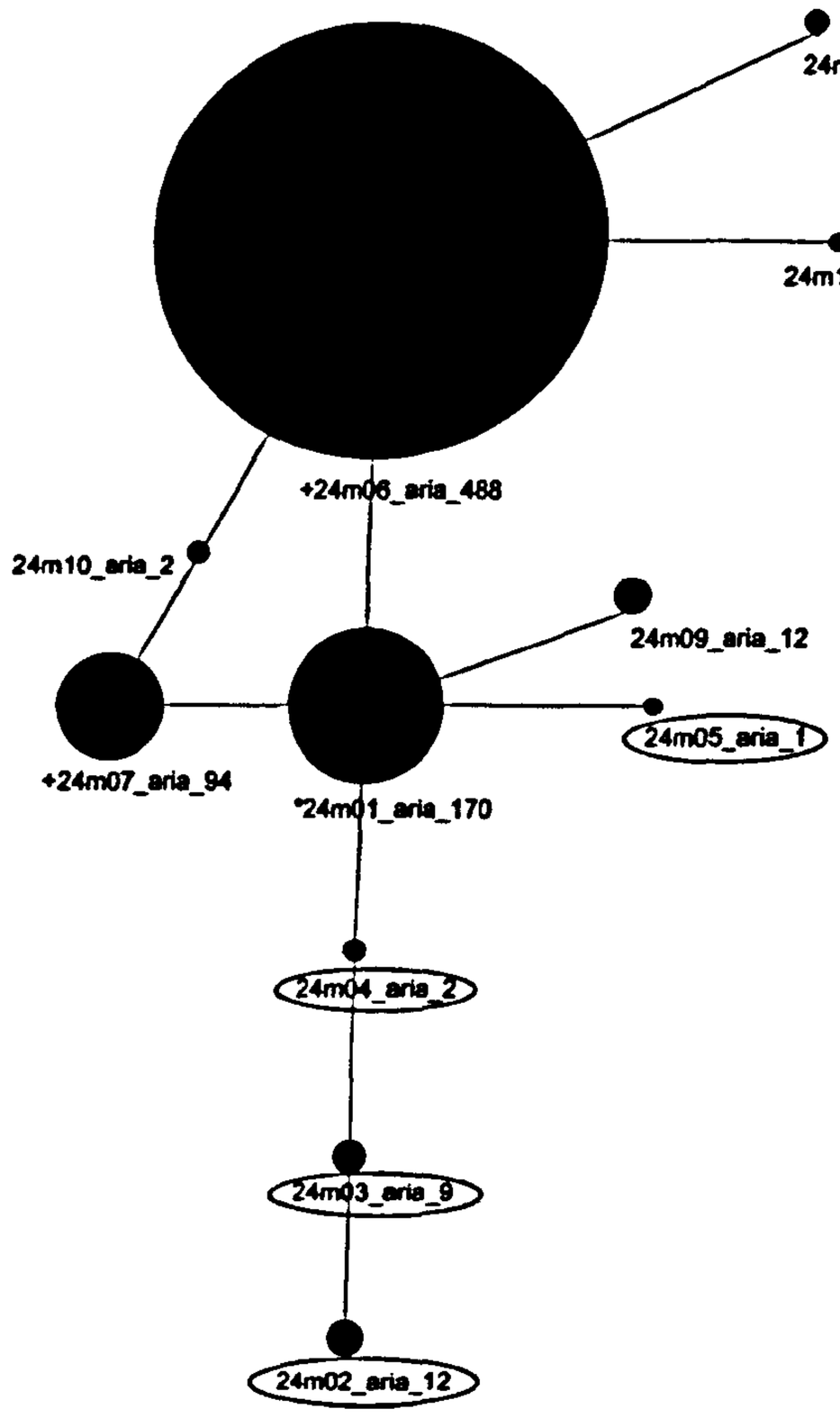


Figure 2.7 (right) Parsimony network (TCS v1.21) showing the genealogical relationships between the 11 AAm24 alleles (length 121 bp) from 398 *P. ariasi*, with a 4 step 95% connection limit. Haplotypes are shown as black filled circles with sizes proportional to their frequency of occurrence, which is given by the number after the allele code (aria_N). Alleles in ellipses private to Morocco; * alleles found in Portugal, Morocco and others; + alleles found in Portugal and others, but absent from Morocco.

showed some geographical structure, with all alleles from Morocco and most from Portugal being associated with just one of the two modal alleles from Europe.

The cyt b network showed most population structure. Haplogroups A-C and F showed 'star-burst' patterns - a central common modal haplotype (darker shades) with rarer haplotypes (lighter shades) derived from it by 1 to 4 mutational steps. This shallow radiation is a signal of recent haplogroup expansion from a small or modest number of founders. Only haplogroup B had sub-haplogroups - two central modes that may reflect two separate histories; CB05 showed a greater expansion pattern than CB04. Furthermore, haplogroup B may have diverged earlier evidenced by 10-14 mutational steps from haplogroups A, E and F, concurring with its basal branch position in the Bayesian phylogeny.

Geographical variation was mapped on the gene networks, and haplogroup phylogeographic structure was observed. Haplogroup F was found in only Morocco where all *P. ariasi* contained it, whereas Haplogroup A was geographically most widespread predominating in all Iberian and French populations (80% of European flies) (Table 2.2; Figure2.4) and except in NW Spain the modal haplotype (CB25) predominated in each population (69% overall). Haplogroup C was uncommon (3.4% of European flies) but geographically widespread in Iberia and France; whilst haplogroup E was rare (0.7% of European flies) and only found in the central Pyrenees and NE Spain. Haplogroup B was less abundant (15.9% of European flies), omnipresent in the French Pyrenees (11.1-48.1%), but absent in Portugal, Morocco and the northern Massif Central, the latter a leading-edge effect where haplogroup A was at fixation and suggesting population sub-division between this region and the French Pyrenees. Haplogroup B was present in low frequencies in NE and NW Spain and the southern Massif Central (4.2-12.5%). The unique demographic history of *P. ariasi* in the French eastern (E) Pyrenees was indicated by the presence of two cyt b sub-haplogroups B and their nearly disjunct distributions. Modal haplotype CB04 and its derived haplotypes were restricted to the French E Pyrenees, east of the Ariège valley to the northeastern edge of the French Pyrenees; where they were omnipresent at low-moderate frequency (0.063-0.318) and constituted 61.4% of haplogroup B. In contrast, modal haplotype CB05 had a contiguous distribution only in geographically bordering central Pyrenees, southern Massif Central and northern Spain. Population PAS at the western Ariège border was represented by both modal haplotypes.

Table 2.2 Frequencies of cyt b haplotypes and haplogroups in 19 spatio-temporal populations of *P. ariasi*. Populations grouped by geographical region. Haplotypes arranged by haplogroup. Numbers of mutational steps from the most geographically widespread haplotype within each haplogroup are given. Number of populations sharing a haplotype is indicated.

Region			Morocco	Portugal	NW Spain	C Pyrenees, France		Eastern Pyrenees, France						NE Spain	Massif Central (MC) and Rhone valley, France					Lot, France			
															MC(S)	MC(S)	MC	MC	Rhone				
Population code			AGH	CHR	CSP	HP1	HP2	PAS	IRL07	TUL	ARQ06	ARQ08	CAT	TRJ	CTU	SPV	ROQ	SAM13	DRAz4	LNP	RME		
N			17	24	24	27	18	52	22	24	38	23	16	23	24	24	15	24	20	24	13		
Haplotype code CBx_aria	Haplo-group code	No. nt steps from CB25_aria																				Total *P. ariasi* with each haplotype	No. of popuations with each haplotype
CB25	A	0		0.583	0.292	0.556	0.833	0.404	0.636	0.625	0.421	0.565	0.438	0.435	0.458	0.417	0.600	0.500	0.800	0.958	1.000	241	Shared 18
CB20	A	1			0.542			0.019						0.043			0.133					17	Shared 4
CB75	A	1													0.167	0.292	0.067		0.050			13	Shared 4
CB79	A	1		0.125								0.043										4	Shared 2
CB18	A	1				0.074				0.042	0.026											4	Shared 3
CB24	A	1						0.019			0.026											2	Shared 2
CB22	A	1											0.063	0.087					0.100			5	Shared 3
CB50	A	1													0.083	0.083						4	Shared 2
CB68	A	1													0.042	0.042						2	Shared 2
CB81	A	1		0.042																		1	Single population
CB82	A	1		0.042																		1	Single population
CB84	A	1		0.042																		1	Single population
CB38	A	1			0.042																	1	Single population
CB100	A	2				0.037																1	Single population
CB101	A	1				0.037																1	Single population
CB103	A	1				0.037																1	Single population
CB99	A	1				0.037																1	Single population
CB106	A	1					0.056															1	Single population
CB01	A	1						0.038														2	Single population
CB28	A	1						0.019														1	Single population
CB31	A	1						0.019														1	Single population
CB45	A	1							0.045													1	Single population
CB47	A	1								0.042												1	Single population
CB61	A	1								0.042												1	Single population
CB35	A	1									0.026											1	Single population
CB55	A	1									0.026											1	Single population
CB63	A	1									0.026											1	Single population
CB40	A	2									0.026											1	Single population
CB21	A	1											0.063									1	Single population
CB10	A	1											0.063									1	Single population
CB19	A	1											0.063									1	Single population
CB15	A	1											0.063									1	Single population
CB23	A	1											0.063									1	Single population
CB86	A	1												0.043								1	Single population
CB89	A	1												0.043								1	Single population
CB90	A	1												0.043								1	Single population
CB91	A	1												0.043								1	Single population
CB92	A	2												0.043								1	Single population
CB87	A	2												0.043								1	Single population
CB41	A	1													0.042							1	Single population
CB64	A	2													0.042							1	Single population
CB66	A	2													0.042							1	Single population
CB67	A	2													0.042							1	Single population
CB69	A	2													0.042							1	Single population
CB107	A	1														0.083						2	Single population
CB108	A	1														0.042						1	Single population
CB94	A	1															0.067					1	Single population
CB73	A	2															0.067					1	Single population
CB74	A	2															0.067					1	Single population
CB12	A	1																0.125				3	Single population
CB11	A	1																0.083				2	Single population
CB14	A	2																0.292				7	Single population
CB76	A	2																	0.050			1	Single population
CB109	A	1																		0.042		1	Single population
Total *P. ariasi* in haplo-group A			0	20	21	21	16	27	15	18	22	14	13	19	23	23	15	24	20	24	13	348	

Table 2.2 Continued.

Region			Morocco	Portugal	NW Spain	C Pyrenees, France		Eastern Pyrenees, France						NE Spain	Massif Central (MC) and Rhone valley, France					Lot, France			
															MC(S)	MC(S)	MC	MC	Rhone				
Population code			AGH	CHR	CSP	HP1	HP2	PAS	IRL07	TUL	ARQ06	ARQ08	CAT	TRJ	CTU	SPV	ROQ	SAM13	DRAz4	LNP	RME		
N			17	24	24	27	18	52	22	24	38	23	16	23	24	24	15	24	20	24	13		
Haplotype code CBx_aria	Haplo-group code	No. nt steps from CB05 (CB04)_aria																				Total *P. ariasi* with each haplotype	No. of popuations with each haplotype
CB05	B	0 (1)			0.125	0.074	0.056	0.231						0.043		0.042						20	Shared 6
CB04	B	1 (0)						0.154	0.273	0.167	0.211	0.174	0.063									31	Shared 6
CB52	B	1 (2)									0.026	0.043										2	Shared 2
CB53	B	1 (2)				0.037																1	Single population
CB98	B	1 (2)				0.037																1	Single population
CB105	B	1 (2)					0.056															1	Single population
CB32	B	1 (2)						0.038														2	Single population
CB06	B	1 (2)						0.019														1	Single population
CB17	B	1 (2)						0.019														1	Single population
CB07	B	2 (1)						0.019														1	Single population
CB44	B	2 (1)							0.045													1	Single population
CB60	B	1 (2)									0.026											1	Single population
CB95	B	1 (2)									0.026											1	Single population
CB36	B	2 (1)									0.026											1	Single population
CB96	B	1 (2)										0.043										1	Single population
CB97	B	2 (1)										0.043										1	Single population
CB03	B	1 (2)											0.063									1	Single population
CB88	B	1 (2)												0.043								1	Single population
Total *P. ariasi* in haplo-group B		No. nt steps from CB26_aria (below)	0	0	3	4	2	25	7	4	12	7	2	2	0	1	0	0	0	0	0	69	
CB26	C	0								0.042	0.053	0.087		0.043								6	Shared 4
CB83	C	1		0.042																		1	Single population
CB85	C	3		0.042																		1	Single population
CB80	C	3		0.042																		2	Single population
CB62	C	1								0.042												1	Single population
CB37	C	1									0.026											1	Single population
CB39	C	3									0.026											1	Single population
CB02	C	1											0.063									1	Single population
CB65	C	1													0.042							1	Single population
Total *P. ariasi* in haplo-group C		No. nt steps from CB104_aria (below)		4	0	0	0	0	0	2	4	2	1	1	1	0	0	0	0	0	0	15	
CB104	E	0				0.037																1	Single population
CB102	E	1				0.037																1	Single population
CB93	E	1												0.043								1	Single population
Total *P. ariasi* in haplo-group E		No. nt steps from CB110_aria (below)	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	
CB110	F	0	0.471																			8	Single population
CB111	F	1	0.059																			1	Single population
CB113	F	1	0.059																			1	Single population
CB114	F	1	0.059																			1	Single population
CB119	F	1	0.059																			1	Single population
CB115	F	2	0.059																			1	Single population
CB116	F	2	0.059																			1	Single population
CB112	F	2	0.059																			1	Single population
CB117	F	3	0.059																			1	Single population
CB118	F	4	0.059																			1	Single population
Total *P. ariasi* in haplo-group F			17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	

Haplotype/allele distribution showed geographical regional grouping (restricted gene flow) within France: cyt b alleles CB75, CB50, CB68 were localized to the Massif Central, E Pyrenees (CB24 and CB04), or Pyrenees only (CB18) (Table 2.2). Distinction between the French Pyrenees and the northern Massif Central with Lot was observed for EF-1α (predominant EF01 and EF03, respectively), locus AAm20 (20m02 with near non-overlapping frequencies of 0.259-0.354 and 0.283-0.614, respectively), and locus AAm24 (24m08 found only in E Pyrenees) (Tables 2.3 to 2.5). The E Pyrenees showed distinction from its bordering regions (C Pyrenees and NE Spain) by the absence of allele EF11. NW Spain was recorded to be distinct (different predominating alleles at all nuclear loci) from NE Spain and France, whose similar frequencies at nuclear loci indicated contemporary gene flow.

2.3.4 No reproductive isolation between *P. ariasi* populations defined by cyt b haplogroups, within populations or overall between locus pairs

A biological species is a group of interbreeding natural populations that are reproductively isolated from other such groups. Two populations of *P. ariasi* in the E Pyrenees (PAS, ARQ - the latter pooled for two collection years, valid as each nuclear locus remained in HWE) contained sufficient flies with cyt b haplogroups A or B to test for reproductive isolation, therefore cryptic speciation. No evidence of haplogroup associated biological speciation was found. Observed genotype frequencies did not differ from those expected in a single randomly-mating population in each location: no significant deviation from HWE was found ($P > 0.05$) (Table 2.6). No linkage disequilibrium between haplotypes or alleles at different loci (cyto-nuclear and nuclear-nuclear) occurred: pairwise comparisons of LD showed no significant difference from the null hypothesis of independent haplotype/allele association between loci ($P > 0.05$) (Table 2.7).

An assessment of all populations independently concluded against biological speciation irrespective of cyt b haplogroup content. HWE was supported at each nuclear locus (Table 2.13), except population of NE Spain at EF-1α whose significantly reduced heterozygosity might have resulted from mixing sub-populations on a short evolutionary time-scale. Four out of 108 Fisher exact probability tests undertaken (per population for each locus pair) statistically supported LD ($P < 0.05$), however these were randomly scattered among populations and locus pairs. Overall, for both the haplogroup and individual population tests, no locus pair showed LD ($P > 0.185$).

Table 2.3 Allele frequencies of EF-1α alleles in 18 spatio-temporal populations of *P. ariasi*. I = inferred based on a defined algorithm (section 2.2.4).

Region	Morocco	Portugal	NW Spain	C Pyrenees, France		Eastern Pyrenees, France						NE Spain	Massif Central (MC) & Rhone, France				Lot, France	
													MC(S)	MC(S)	MC	Rhone		
Allele/ Population code	AGH	CHR	CSP	HP1	HP2	PAS	IRL07	TUL	ARQ06	ARQ08	CAT	TRJ	CTU	SPV	SAM13	DRAz4	LNP	RME
EF03	0.029	0.109	0.429	0.18	0.25	0.146	0.364	0.304	0.429	0.152	0.125	0.152	0.229	0.326	0.729	0.432	0.875	0.885
EF01		0.043	0.214	0.68	0.722	0.74	0.477	0.543	0.464	0.543	0.563	0.565	0.542	0.522	0.146	0.523	0.104	0.115
EF12	0.059	0.13				0.01		0.022				0.022						
EF38	0.029									0.022								
EF11		0.043		0.02								0.022						
EF14		0.283	0.119			0.01	0.023											
EF02			0.071		0.028	0.031	0.091		0.036	0.196	0.188	0.022	0.188	0.065	0.125	0.023	0.021	
EF05			0.048				0.045	0.087	0.036	0.087	0.031	0.022		0.022				
EF06			0.071	0.02		0.021		0.043			0.031	0.043	0.042	0.022				
EF13			0.024			0.01												
EF08				0.04		0.01						0.022						
EF10				0.02		0.01						0.087						
EF29	0.382																	
EF28	0.088																	
EF30	0.088																	
EF37	0.059																	
EF44; I	0.059																	
EF31	0.029																	
EF32	0.029																	
EF33	0.029																	
EF34	0.029																	
EF35	0.029																	
EF36	0.029																	
EF45; I	0.029																	
EF20		0.109																
EF19		0.087																
EF21		0.043																
EF22		0.043																
EF23		0.022																
EF24		0.022																
EF25		0.022																
EF26		0.022																
EF27		0.022																
EF42			0.024															
EF09				0.02														
EF43; I				0.02														
EF15						0.01												
EF39									0.036									
EF40; I											0.031							
EF04											0.031							
EF17												0.022						
EF18												0.022						
EF16														0.022				
EF41														0.022				
EF07																0.023		
N	17	23	21	25	18	48	22	23	14	23	16	23	24	23	24	22	24	13
No. alleles	15	14	8	8	3	10	5	5	5	5	7	11	4	7	3	4	3	2

Table 2.4 Allele frequencies of AAm20 alleles in 18 spatio-temporal populations of *P. ariasi*. I = inferred based on a defined algorithm (section 2.2.4).

Region	Morocco	Portugal	NW Spain	C Pyrenees, France		Eastern Pyrenees, France						NE Spain	Massif Central (MC) and Rhone, France				Lot, France	
													MC(S)	MC(S)	MC	Rhone		
Allele / Population code	AGH	CHR	CSP	HP1	HP2	PAS	IRL07	TUL	ARQ06	ARQ08	CAT	TRJ	CTU	SPV	SAM13	DRAz4	LNP	RME
20m02	1	0.125	0.667	0.259	0.306	0.354	0.227	0.313	0.313	0.326	0.344	0.283	0.417	0.458	0.5	0.614	0.283	0.5
20m01		0.042	0.188	0.704	0.556	0.646	0.727	0.667	0.688	0.652	0.656	0.652	0.583	0.521	0.5	0.386	0.717	0.5
20m03			0.042	0.037	0.139			0.021		0.022		0.043						
20m04		0.021																
20m07		0.729																
20m08		0.063																
20m10		0.021																
20m11			0.063															
20m12			0.021															
20m17; I			0.021															
20m14; I							0.023											
20m15; I							0.023											
20m06												0.022						
20m16; I														0.021				
N	17	24	24	27	18	24	22	24	24	23	16	23	24	24	24	22	23	13
No . Alleles	1	6	6	3	3	2	4	3	2	3	2	4	2	3	2	2	2	2

Table 2.5 Allele frequencies of AAm24 alleles in 18 spatio-temporal populations of *P. ariasi*.

Region	Morocco	Portugal	NW Spain	C Pyrenees, France		Eastern Pyrenees, France						NE Spain	Massif Central (MC) and Rhone, France				Lot, France	
													MC(S)	MC(S)	MC	Rhone		
Allele / Population code	AGH	CHR	CSP	HP1	HP2	PAS	IRL07	TUL	ARQ06	ARQ08	CAT	TRJ	CTU	SPV	SAM13	DRAz4	LNP	RME
24m01	0.294	0.771	0.583	0.167	0.083	0.104	0.159	0.208	0.063	0.196	0.156	0.174	0.104	0.125	0.188	0.348		
24m06		0.167	0.354	0.407	0.722	0.729	0.477	0.646	0.813	0.63	0.625	0.609	0.75	0.729	0.771	0.652	1	1
24m07		0.063	0.042	0.389	0.167	0.083	0.295	0.104	0.104	0.109	0.188	0.174	0.146	0.146	0.042			
24m09			0.021	0.037		0.063	0.045	0.021		0.022	0.031	0.022						
24m10					0.028							0.022						
24m08							0.023	0.021	0.021	0.043								
24m02	0.353																	
24m03	0.265																	
24m04	0.059																	
24m05	0.029																	
24m11						0.021												
N	17	24	24	27	18	24	22	24	24	23	16	23	24	24	24	23	24	13
No . alleles	1	3	4	4	4	5	5	5	4	5	4	5	3	3	3	2	1	1

Table 2.6 Panmixis in two populations of *P. ariasi* indicated by non-significant Hardy-Weinberg (HW) tests ($P > 0.05$) at three nuclear loci (EF-1α, AAm20, AAm24) for sandflies associated with three cyt b haplogroups.

Population	N	No. *P. ariasi* with each cyt b haplogroup				Locus	HW *P*-value
		A	B	C	Other		
PAS	23	12	10	0	1	EF-1α	0.70039
						AAm20	0.39035
						AAm24	0.40498
ARQ	37	22	11	4	0	EF-1α	0.16438
						AAm20	0.65122
						AAm24	0.48627

Table 2.7 No linkage disequilibrium between haplotypes or alleles between pairs of different loci (cyto-nuclear and nuclear-nuclear), according to a model of linkage disequilibrium with a null hypothesis of independent haplotype/allele association between loci (non-significant $P > 0.05$).

Population	Locus 1	Locus 2	*P*-value (± S.E.)
PAS	EF-1α	AAm20	0.913189 ± 0.000214
	EF-1α	AAm24	0.072994 ± 0.00063
	AAm20	AAm24	0.361535 ± 0.000393
	EF-1α	cyt b	0.376315 ± 0.001537
	AAm20	cyt b	0.171733 ± 0.000419
	AAm24	cyt b	0.087596 ± 0.000781
ARQ	EF-1α	AAm20	0.821059 ± 0.000472
	EF-1α	AAm24	0.768402 ± 0.000893
	AAm20	AAm24	0.430054 ± 0.00062
	EF-1α	cyt b	0.488901 ± 0.002063
	AAm20	cyt b	0.764847 ± 0.000883
	AAm24	cyt b	0.311091 ± 0.001693

Table 2.8 Absence of selection at two loci of *P. ariasi* indicated by non-significant McDonald-Kreitman tests (Fisher's exact two-tailed test, significant when $P < 0.05$).

Locus	*P. ariasi* with outgroup species	Ds	Ps	Dn	Pn	Fisher's exact *P*-value	NI
Cyt b	near *P. ariasi*†	14	66	2	18	0.5162	1.909
EF-1α	near *P. ariasi*†	7	21	1	0	0.2759	NA
EF-1α	*P. neglectus*	31	21	1	0	1.0	NA

Legend Ds, Ps, Dn and Pn, correspond to the number synonymous (s) and nonsynonymous (n) substitutions per site that are polymorphic (P) in *P. ariasi*, or fixed (D) between *P. ariasi* and a selected outgroup per locus. NI = Neutrality Index. Strict neutrality has an index of 1.0; NI < 1 indicates positive selection; NI > 1 indicates purifying selection. Test conducted in DNASP (v. 4.90.1). † Likely to be *P. chadlii*.

2.3.5 Neutral evolution of cyt b and the three nuclear loci

Directional or balancing selection was not detected at any of the four loci using two sorts of tests. PAML CODEML and MK investigate polymorphism in codons relative to divergence with other *Phlebotomus* and were appropriate for analysing the relatively long protein-coding fragments for cyt b and EF-1α. Tajima's *D* is an allele frequency-based test and, therefore, additionally appropriate for the anonymous loci AAm20 and AAm24. It is more sensitive than the former analyses for detecting recent selection, but significant results can arise from either selection, demography or recombination.

CODEML using species divergence data concluded against positive directional or diversifying selection ($\omega < 1$) along the *P. ariasi* branch (ω_a) against the background branches (ω_o) for three phylogenies re-estimated from their Bayesian topology's: cyt b Figure 2.2; short EF-1α Figure 2.3a, and long EF-1α Figure 2.3b. The LRT supported significant heterogeneity in purifying selection pressure between branches in cyt b only ($2\Delta l = 2(-4118.35 - (-4120.55)) = 4.4$ at χ^2 df = 1; $0.01 < P < 0.05$), where *P. ariasi* was shown to be under greater purifying selection pressure compared to background branches ($\omega_a = 0.0084$, and $\omega_o = 0.0202$, respectively).

CODEML has low power for detecting intra-specific selection (Anisimova *et al.*, 2002). The more sensitive MK population test for protein-coding regions, showed no significant departures from neutral expectation in the number of polymorphic versus fixed substitutions for both loci of *P. ariasi* ($P = 0.2759$-1.0, two-tailed Fisher's exact test; Table 2.8). The associated NI values indicated the direction of selection of cyt b tended towards purifying selection, signalled by few fixed nonsynonymous substitutions. For EF-1α the NI could not perform well with none to one fixed differences (NI = 0.0). Confidence in the power of MK test was achieved through appropriate outgroup selection where d_S was not approaching saturation (< 0.5): near *P. ariasi* was the suitable outgroup for cyt b (d_S 0.2686-0.2702); for EF-1α both near *P. ariasi* (d_S 0.0721) and *P. neglectus* (d_S 0.4723) were assessed (d_S values quoted according to the ML model of Goldman and Yang (1994); Appendix 2.14)).

Both PAML and MK test for selection at the protein level, which may not be applicable for anonymous nuclear loci AAm20 and AAm24, or have insufficient power when there are few polymorphic nonsynonymous changes. Therefore, allele frequency spectrum population based neutrality tests were implemented to test for recent selection (Table 2.13). Tajima's *D* was significant ($P < 0.05$; none after sequential Bonferroni

correction) and negative only for: cyt b, Morocco, NE Spain, southern Massif Central and Massif Central (ROQ, DRAz4); and EF-1α, NE Spain and central Pyrenees (HP1). Negative *D* values signal directional selection or population expansions by reflecting an excess of low frequency haplotypes/alleles, the latter the more likely interpretation as all but one (NE Spain) showed a corresponding significant and negative Fu's F_S value. It is recognised that recombination tends to reduce the variance of Tajima's *D* leading to conservative estimates (e.g. Ramirez-Soriano *et al.*, 2008). This is unlikely to have affected the results presented here because, considering the longer gene fragments, the estimated minimum number of recombination events (Rm) was low for cyt b, as expected for mitochondrial DNA (0 for 14 populations, and 1-3 for four populations), and for EF-1α (0 for nine populations, 1-2 for nine populations).

2.3.6 Genetic diversity and population structure of *P. ariasi*

The cyt b network showed haplogroups to not correspond directly to current geographical populations, indicating these mtDNA lineages to probably have had distinct and separate demographic histories. Coalescent modelling in MDIV jointly estimates migration rates and divergence times to distinguish the retention of ancestral polymorphisms from ongoing gene flow. The highest likelihood value for *M* in all haplogroup pairwise comparisons was found to be very close to zero, showing little to no evidence of ongoing migrant exchange, which combined with estimated divergence times > 100,000 years before present (conservatively, the end of the previous interglacial), allowed haplogroups to be considered as demographically independent with respect to climate change since the LGM (e.g. Smith and Farrell, 2005). In part, the aim of this chapter was to investigate the historical population structure of *P. ariasi* i.e. the impact of the glacial cycles on species distribution. Therefore, to avoid errors in estimates of demographic parameters through the effects of mixed ancestry in contemporary populations, some tests were assessed using cyt b haplogroups of *P. ariasi* as populations.

2.3.6.1 Demographic history of cyt b haplogroups

Percentage specimen representation of the five assigned cyt b haplogroups observed by Bayesian phylogenetic and clustered by the network reconstruction were A (76.99%), B (15.26%), C (3.32%), E (0.66%), and F (3.76%). Diversity statistics indicated haplogroup A having gone through a prolonged or severe bottle-neck (or

selective sweep): lowest H_d (0.517±0.0334) and π values (0.000947±0.000794) (Table 2.9). Assuming each haplogroup has the same mutation rate, overlapping π values (degree of polymorphism) amongst cyt b haplogroups would indicate relatively contemporaneous divergence times. Haplogroups C and F showed the highest absolute averages of π and relatively high H_d a signal of a large and stable long-term N_e (or an admixed population of historically divided populations), reflected in the network by relatively larger number of mutational steps (1-4, as opposed to 1-2) from the modal haplotype within their haplogroups (Figure 2.4; Table 2.2).

Two mutation rates (μ = 2.3% or 1%) were used to bound confidence intervals for dating. Using generation time of 1 year, more common of *P. ariasi* in colder climes, gene coalescence times (*t*MRCA = 949,771-658,380 or 2,182,175-1,514,275 years ago (y.a.)) and divergence times (t = 545,997-376,757 or 1,168,761-866,541 y.a.) (Table 2.10) between all haplogroup pairs dated to within the Pleistocene epoch (2,588,000-10,000 y.a.). If three generations was credible in the warmest places during the interglacials, estimates would be reduced by two-thirds and gene coalescence time would remain before the last interglacial (Eemian, 125-110 k.y.a.) and both coalescence and divergence not more recently during the Holocene. Dating methods allowed this study to conclude against the opening of the Gibraltar Straits (5.5-4.9 m.y.a.; Pliocene epoch) as the vicariance event causing the differentiation of Moroccan haplogroup F: pairwise comparisons with this haplogroup showed both *t*MRCA and t to date within the Pleistocene (maximum *t*MRCA, 1,983,128 y.a.). To lend support to these dates, *t*MRCAs calculated by MDIV were consistent with the branching order obtained by Bayesian reconstruction and the parsimony network structure: Haplogroup B showed the most ancient divergences from the common ancestor of haplogroups C, A and F (*t*MRCA 2,182,175-1,546,729 y.a.).

Evidence for past sudden demographic expansion of unstructured mtDNA populations (haplogroups or sub-haplogroups) was supported by mismatch distribution of pairwise nucleotide differences among individuals. This confirmed that the star-bursts observed in the cyt b genealogy statistically fitted a model of sudden demographic expansion. This type of expansion was supported for haplogroups A, F (unimodal mismatch distribution) and C (Raggedness index $P > 0.05$, under a null hypothesis of sudden expansion). Haplogroup B gave a unimodal mismatch distribution but the hypothesis of expansion was rejected; Raggedness index $P < 0.05$. Expansion was supported for sub-haplogroup B CB04, which occurred only in E Pyrenees, but not

Table 2.9 Descriptive population statistics for each cyt b haplogroup found in populations of *P. ariasi*.

Haplo-group	No. *P. ariasi* (% of total)	No. pops	S	*h*	H_d (± SD)	π (± SD)
A	348 (76.99%)	18	53	54	0.517±0.0334	0.000947±0.000794
B	69 (15.26%)	11	16	18	0.720±0.0429	0.001391±0.001044
C	15 (3.32%)	7	12	9	0.848±0.0878	0.002891±0.001908
E	3 (0.66%)	2	2	3	1.000	0.001807±0.001856
F	17 (3.76%)	1	13	10	0.794±0.1035	0.002630±0.001756

Legend S = number of segregating sites, *h* = number of haplotypes. All statistics estimates in ARLEQUIN (v3.11). H_d = Haplotype (gene) diversity and π = Nucleotide diversity (Nei, 1987).

Table 2.10 Isolation with migration coalescence model in MDIV: to estimate gene coalescence and divergence times for pairs of cyt b haplogroups found in populations of *P. ariasi*. Confidence intervals estimated with two mutation rates 2.3% and 1%. Generation time = 1 per annum.

Cyt b haplogroup	θ	T	TMRCA	*t*MRCA (y.a.) (μ 2.3%)	*t*MRCA (y a.) (μ 1%)	*t* (y.a.) (μ 2.3%)	*t* (y.a.) (μ 1%)
B + A	5.444	1.644	2.862	949,771	2,182,175	545,997	1,254,492
B + C	2.079	2.976	5.312	672,491	1,546,729	376,757	866,541
B + F	2.507	3.312	5.648	862,230	1,983,128	505,613	1,162,911
A + C	5.787	1.272	2.074	730,989	1,681,276	448,321	1,031,139
A + F	6.479	1.288	2.026	799,321	1,838,439	508,156	1,168,761
C + F	2.408	2.72	4.49	658,380	1,514,275	399,840	917,333

Legend θ = scaled parameter for nucleotide heterozygosity; *T* = scaled gene divergence time; TMRCA = expected time to the most recent common ancestor; *t*MRCA = gene coalescence time; *t* = gene divergence time; y.a. = years ago.

Table 2.11 Mismatch distribution statistics for *P. ariasi* cyt b haplogroups and sub-haplogroups. Sudden demographic expansion detected when significance of Raggedness index $P > 0.05$. Time elapsed since beginning of expansion event (*t*) calculated by $\tau = 2ut$. 95% confidence intervals of τ were estimated around mutation rates 2.3% and 1% at $\alpha = 0.05$. Generation time = 1 per annum; y.a. = years ago.

(Sub-)haplogroup	CB_A	CB_B	CB_C	CB_F	CB_B05	CB_B04
Raggedness index	0.07941	0.11236	0.09959	0.0157	0.1656	0.36789
Raggedness *P*	0.4784	0.0229	0.3379	0.996	0.0292	0.6405
Tau (τ)	0.73 (0.60-0.92)	1.199 (0.86-1.69)	2.826 (0.79-5.01)	2.168 (0.53-4.68)	1.020 (0.36-1.65)	3.000 (0.42-3.00)
t 2.3% (y.a.)	44,453 (36,869-55,780)	NA	172,086 (48,406-305,302)	132,018 (32,350-284,965)	NA	182,682 (25,452-182,682)
t 1% (y.a.)	102,240 (84,800-128,294)	NA	395,798 (111,333-702,195)	303,641 (74,405-655,419)	NA	420,168 (58,539-420,168)

for sub-haplogroup B CB05 (Table 2.11).

Where demographic expansion was supported the time since the beginning of these expansions was approximated given the estimates of parameter τ (Table 2.11) and its 95% confidence intervals. Two divergence rates were used [per nucleotide per generation time of 1: 0.0115 (2.3%) and 0.005 (1%) see Materials and methods]. Estimates since the beginning of cyt b (sub-)haplogroup demographic expansion were dated to within the Pleistocene epoch: predating the LGM (18,000 y.a.) and therefore were not rapid/recent expansions during the Holocene warm interglacial, using either divergence rate: *t* range 44,453-420,168 y.a.; 95% CI range 21,884-702,195 y.a. Large overlapping intervals around expansion dates did not allow for definitive distinctions between (sub-)haplogroup expansions, but it might be hypothesized based on the mean estimates of τ, that haplogroup A expanded most recently (102,240-44,453 y.a.; other haplogroups between 420,168-132,018 y.a.), and represents the most extensive population expansion evidenced by the largest star-burst in the network; contained the greatest number of radiating haplotypes (54).

To understand patterns of historical gene flow and random genetic drift of European *P. ariasi*, dependence between genetic and geographical distance was modelled using flies associated with predominating cyt b haplogroup A to eliminate the effects of multiple haplogroup histories. Pairwise F_{ST} estimates for genetic differentiation had "very great" values (> 0.25; Wright, 1978) signalling high levels of inter-population genetic variance in: cyt b-A, -0.0340 to 0.5642; EF-1α-A, -0.0247 to 0.7300; AAm20-A, -0.0296 to 0.5401; and AAm24-A, -0.0302 to 0.7678.

Genetic differentiation was observed between Portugal, NW Spain and N Massif Central (including Lot and Rhone valleys reflecting their leading-edge effect) compared with all other populations: defined by "great" ($0.15 < F_{ST} < 0.25$) and "very great" (> 0.25) F_{ST} estimates. Conversely populations in the French Pyrenees with NE Spain showed gene flow across this region: genetic differentiation was rarely "great"/significant.

To summarize, $F_{ST} > 0.25$ with $P < 0.05$: cyt b-A [NW Spain *vs.* all populations except NE Spain, E Pyrenees (CAT), Massif Central (CTU, ROQ); Lot *vs.* Massif Central (SPV, SAM13)], EF-1α-A [Portugal *vs.* C Pyrenees (HP2), E. Pyrenees (PAS), N Massif Central (SAM13), Lot; NW Spain *vs.* E Pyrenees (PAS); N Massif Central (SAM13) *vs.* all but NW Spain, E Pyrenees (IRL07), S. Massif Central (SPV); Lot *vs.* all except NW Spain, E Pyrenees (IRL07), N Massif Central (SAM13)], AAm20-A

[Portugal *vs.* all populations; NW Spain *vs.* Central Pyrenees and 6/13 other locations], and AAm24-A [Portugal *vs.* all populations but NW Spain; NW Spain *vs.* 9/14 French populations; plus Lot *vs.* all except Central Pyrenees (HP2), E Pyrenees (PAS, ARQ), S Massif Central, N Massif Central (SAM13)] (Appendix 2.15).

Globally there was no relationship between genetic and geographical distance (IBD) at cyt b-A, assessed by a Mantel test fitting $F_{ST}/(1-F_{ST})$ against distance ($P > 0.05$) (Table 2.12), a likely consequence of higher than expected number of shared haplotypes between Portugal and other populations in Iberia and France (black oval Figure 2.8a; Appendix 2.15a). In contrast, IBD was supported by Mantel tests ($P < 0.05$) for each of the three nuclear loci (Figure 2.8b) using haplogroup A flies, however, for EF-1α only 6.5% of the genetic variation was associated with geographical distance ($R^2 = 0.0648$).

As IBD and population sub-division are not always mutually exclusive, (Mills *et al.*, 2007) association of genetic variation by geographical distance was supported for: cyt b-A within the Massif Central ($R^2 = 0.47$, Mantel test $P = 0.008$), and all three nuclear loci showed positive yet statistically non-significant association ($P > 0.05$), the latter a similar trend in the E Pyrenees (Table 2.12). Yet for cyt b-A, to merge these regions gave no support for IBD, linear regression line $R^2 = 0.0000$ (yellow symbols and black line in Figure 2.8a), and IBD was also not supported ($R^2 = 0.0002$, $P = 0.211$) between the E Pyrenees and the southern Massif Central (Table 2.12). In the latter within and between regional geographical distances were comparable, however, pairwise genetic distance was relatively higher than expected by geographical distance alone - indicative of a step-change/genetic discontinuity between these two regions. In contrast to the Massif Central gene flow, not IBD, was demonstrated along the Pyrenees at cyt b-A ($R^2 = 0.1157$; $P > 0.05$), nuclear gene analyses being concordant.

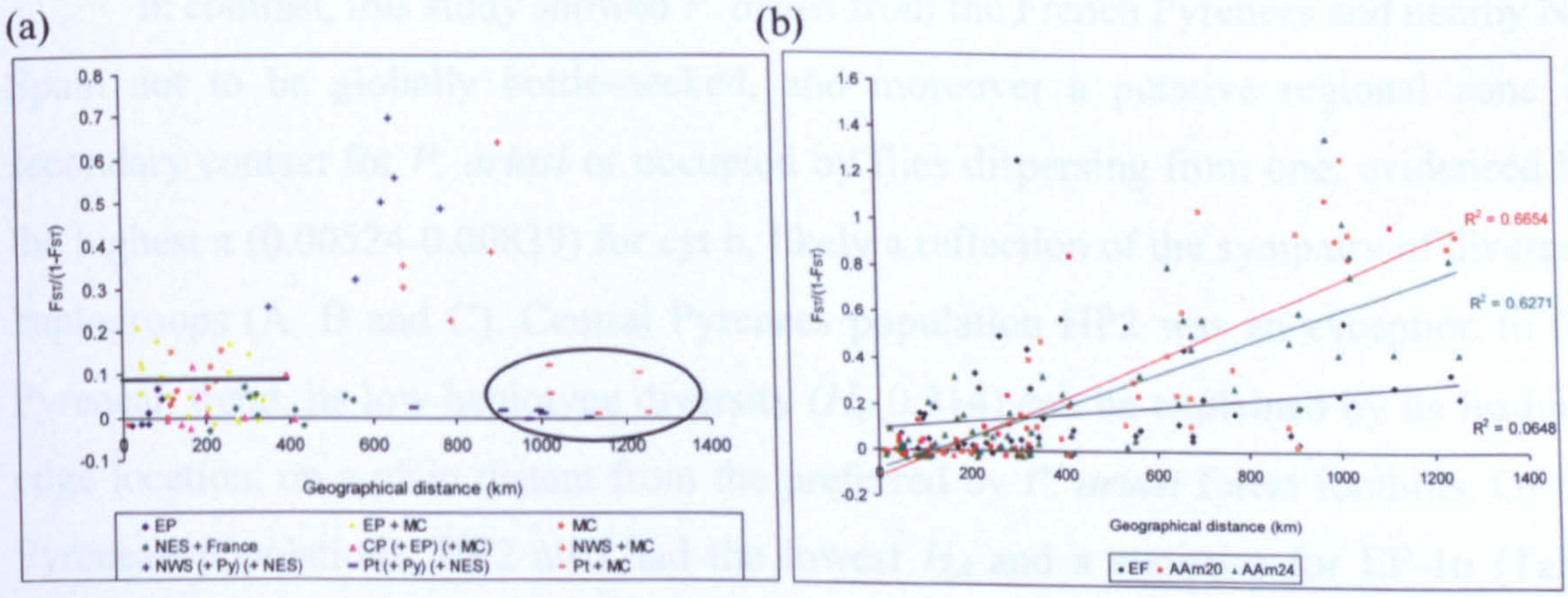
Demographic population structure analysis of genetic variance was implemented using hierarchical AMOVA for the more polymorphic cyt b. Support was shown for Europe sub-divided from Morocco with further sub-structure between populations within regions (percentage variation among regions = 56.0%, $P < 0.05$ and within regions = 6.57%, $P < 0.001$). A Mantel test supported the correlation in fitting genetic distance ($\Phi_{ST}/(1-\Phi_{ST})$ to correlate with AMOVA result, F_{ST} substituted for Φ_{ST}) to geographical distance ($P = 0.006$) between pairwise comparisons of flies from Europe and Morocco. This result was confirmed by dbRDA analysis, where a marginal test showed a significant relationship between genetic distance and geographical distance or

Table 2.12 Testing the association between genetic and geographical distance between *P. ariasi* populations (flies associated with cyt b haplogroup A only) is according to predictions of IBD: fitting estimates of $F_{ST}/(1-F_{ST})$ to geographical distance (km). Significance permuted using a Mantel test; * $P < 0.05$, ** $P < 0.01$ (GENEPOP v4.0).

	Fitting $F_{ST}/(1-F_{ST})$ to distance			Fitting $F_{ST}/(1-F_{ST})$ to ln distance		
Population region; locus	a	b	*P* (c>o)	a	b	*P* (c>o)
Iberia and France; cyt b	0.0513799	0.00013665	0.146	-0.18486	0.05216799	0.146
Iberia and France; AAm20	-0.1113924	0.00085378	0**	-1.04137	0.2289904	0**
Iberia and France; AAm24	-0.0756238	0.00068608	0.004*	-0.82515	0.18440977	0.004*
Iberia and France; EF-1α	0.0943249	0.00016258	0.035*	-0.21804	0.06721853	0.035*
MC; Cyt b	-0.0205183	0.00058778	0.008**	-0.16966	0.04932542	0.008**
MC; AAm20	-0.0066951	0.00015263	0.071	-0.03658	0.01155986	0.071
MC; AAm24	-0.0239628	0.00027105	0.071	-0.08353	0.02186348	0.071
MC; EF1α	0.0619743	0.00043	0.86	-0.12811	0.05431634	0.86
EP; cyt b	-0.0607589	0.00118921	0.098	-0.21366	0.05613781	0.092
EP; AAm20	-0.0113186	0.00016905	0.547	-0.02703	0.00649227	0.535
EP; AAm24	0.110839	-0.0015125	0.921	0.355786	-0.083871	0.916
EP + SMC; Cyt b	0.0392774	0.000295	0.211	-0.06147	0.0295631	0.211
EP + SMC; AAm20	-0.0026529	0.00020919	0.644	-0.04186	0.01304208	0.644
EP + SMC; AAm24	0.0434225	-0.0006295	0.984	0.122729	-0.02974237	0.984
EP + SMC; EF1α	0.008633	0.00058673	0.356	-0.10714	0.0381856	0.356

Legend EP = Eastern Pyrenees; MC = Massif Central (including Lot and Rhone valleys); SMC = southern Massif Central.

Figure 2.8 Plots of genetic against geographical distance to test for isolation-by-distance: (a) locus cyt b haplogroup A; (b) 3 nuclear loci combined (haplogroup A). Extent of correlation given as R^2 values.



Legend (a) ringed data = pairwise estimates with Portugal. Black regression line for populations of Massif Central with E Pyrenees. CP = central Pyrenees; EP = eastern Pyrenees; MC = Massif Central; NES = northeast Spain; NWS = northwest Spain; Py = Pyrenees; SMC = southern Massif Central; Pt = Portugal.

geographical region (54.0% variation explained, P = 0.001 and 65.1%, P = 0.001, respectively); the latter categorizing pairwise comparisons between Morocco/Europe separate from within Europe. When geographical distance was taken into account as a covariate in the multiple regression analysis, geographical region remained correlated to genetic distance (11.7%, P = 0.001).

2.3.6.2 Population genetic structure of geographical regions

Diversity statistics demonstrated that northern Massif Central populations were leading-edge (Table 2.13). The fixation of cyt haplogroup A in the north caused significantly (non-overlapping) lower H_d (0-0363) for the Lot (LNP and RME) and Rhone valleys (DRAz4) compared to all other populations, and lower π (0.00011-0.00248) in the Massif Central, Lot and Rhone valleys, as compared to all other populations (H_d 0.541-0.825 and π 0.00263-0.00839). EF-1α was the only nuclear gene to show geographical variation in diversity, and this concurred with cyt b: highest in the south to lowest in the north [Morocco, Portugal and NW Spain (H_d 0.761-0.881, π 0.00174-0.0271); NE Spain, French Pyrenees and southern Massif Central (H_d 0.427-0.659, π 0.00060-0.00138); and, northern Massif Central, Lot and Rhone valleys (H_d 0.212-0.552, π 0.00027-0.00077)].

In contrast, this study showed *P. ariasi* from the French Pyrenees and nearby NE Spain not to be globally bottle-necked, and moreover a putative regional zone of secondary contact for *P. ariasi* or occupied by flies dispersing from one; evidenced by the highest π (0.00524-0.00839) for cyt b, likely a reflection of the sympatry of diverged haplogroups (A, B and C). Central Pyrenees population HP2 was an exception to the Pyrenean trend, its low haplotype diversity (H_d 0.314) can be explained by its leading-edge location, on a plain distant from the preferred by *P. ariasi* forest foothills. Of the Pyrenean populations, HP2 also had the lowest H_d and π statistics for EF-1α (Table 2.13).

Considering patterns in mtDNA haplotype distribution, in concert with knowledge of environmental determinants controlling *P. ariasi* distribution and abundance (i.e. altitude and habitat preference), led to the implementation of *post-hoc* AMOVA analysis to test support for geographical regional sub-divisions. Table 2.14 details the sub-divisions tested, namely to seek support that the E Pyrenees and upland Massif Central are separated by an unsuitable low land corridor and the uniqueness of the E Pyrenees from its bordering regions. For cyt b although within population

Table 2.13 Descriptive and neutrality statistics for cyt b and three nuclear loci characterized by population of *P. ariasi*.

Population region		Morocco	Portugal	NW Spain	C Pyrenees, France		Eastern Pyrenees, France					
Locus	Population	AGH	CHR	CSP	HP1	HP2	PAS	IRL07	TUL	ARQ06	ARQ08	CAT
Cyt b	N	17	24	24	27	18	52	22	24	38	23	16
	S	13	14	13	23	13	18	14	17	26	17	21
	h	10	8	4	11	4	12	4	7	15	7	10
	H_d (± SD)	0.794 ± 0.104	0.656 ± 0.105	0.630 ± 0.074	0.695 ± 0.098	0.314 ± 0.138	0.769 ± 0.042	0.541 ± 0.094	0.598 ± 0.107	0.788 ± 0.058	0.664 ± 0.099	0.825 ± 0.098
	π (± SD)	0.00263 ± 0.00176	0.00367 ± 0.00225	0.00422 ± 0.00253	0.00594 ± 0.00337	0.00342 ± 0.00216	0.00839 ± 0.00449	0.00764 ± 0.00425	0.00593 ± 0.00338	0.00867 ± 0.00467	0.00808 ± 0.00446	0.00567 ± 0.00333
	Tajima's *D*	-1.87662*	-0.97078	-0.36766	-0.958	-1.24021	1.73183	1.67199	-0.13829	0.1164	1.06438	-1.36893
	Fu F_S	-5.557*	-0.738	3.784	-1.087	2.332	1.642	6.731	1.700	-1.056	2.830	-2.432
	Rm	1	0	0	2	0	1	0	0	3	0	0
EF-1α	N	17	23	21	25	18	48	22	23	14	23	16
	S	10	9	7	9	2	7	4	4	4	4	6
	h	14	14	8	8	3	10	5	5	5	5	7
	H_O	0.824	0.783	0.571	0.600	0.333	0.458	0.500	0.478	0.429	0.522	0.688
	H-W *P*	0.452	0.061	0.097	0.377	0.649	0.700	0.249	0.187	0.226	0.320	0.833
	H_d (± SD)	0.840 ± 0.056	0.881 ± 0.030	0.761 ± 0.050	0.512 ± 0.076	0.427 ± 0.076	0.434 ± 0.059	0.644 ± 0.045	0.615 ± 0.054	0.619 ± 0.054	0.649 ± 0.060	0.651 ± 0.080
	π (± SD)	0.00271 ± 0.00172	0.00262 ± 0.00166	0.00174 ± 0.00122	0.00099 ± 0.00082	0.00060 ± 0.00060	0.00082 ± 0.00071	0.00105 ± 0.00085	0.00112 ± 0.00089	0.00094 ± 0.00080	0.00125 ± 0.00096	0.00138 ± 0.00104
	Tajima's *D*	-0.4273	-0.01808	-0.46112	-1.72213*	-0.05659	-1.26018	-0.2681	-0.10794	-0.75785	0.15706	-0.79343
	Fu F_S	-7.269*	-6.086*	-2.093	-4.128*	0.047	-6.771*	-0.791	-0.579	-1.534	-0.272	-2.397
	Rm	2	2	2	0	0	2	0	1	0	0	1
AAm20	N	17	24	24	27	18	24	22	24	24	23	16
	S	0	5	5	2	2	1	3	2	1	2	1
	h	1	6	6	3	3	2	4	3	2	3	2
	H_O	-	0.458	0.625	0.481	0.778	0.375	0.545	0.458	0.542	0.522	0.563
	H-W *P*	-	0.560	0.770	0.649	0.383	0.390	0.611	0.429	0.360	1	0.592
	H_d (± SD)	-	0.456 ± 0.084	0.525 ± 0.075	0.444 ± 0.061	0.595 ± 0.056	0.467 ± 0.043	0.428 ± 0.074	0.467 ± 0.054	0.439 ± 0.052	0.478 ± 0.052	0.466 ± 0.056
	π (± SD)	-	0.01420 ± 0.01008	0.00757 ± 0.00652	0.00516 ± 0.00509	0.00758 ± 0.00658	0.00519 ± 0.00513	0.00552 ± 0.00534	0.00534 ± 0.00522	0.00488 ± 0.00493	0.00548 ± 0.00531	0.00517 ± 0.00517
	Tajima's *D*	NA	0.32919	-0.96528	0.10024	0.80671	1.45766	-0.59536	0.11932	1.28670	0.15082	1.33679
	Fu F_S	NA	-0.344	-2.271	0.264	0 841	1.770	-0.935	0.270	1.640	0.297	1.537
	Rm	NA	3	0	0	0	0	0	0	0	0	0
AAm24	*N*	17	24	24	27	18	24	22	24	24	23	16
	S	4	2	3	3	2	4	4	4	3	4	3
	h	5	2	4	4	4	5	5	5	4	5	4
	H_O	0.765	0.333	0.375	0.630	0.389	0.417	0.591	0.458	0.375	0.565	0.563
	H-W *P*	1	0.359	0.203	0.798	0.491	0.405	0.138	0.113	1	0.315	0.736
	H_d (± SD)	0.736 ± 0.034	0.382 ± 0.078	0.543 ± 0.047	0.666 ± 0.031	0.456 ± 0.090	0.456 ± 0.084	0.672 ± 0.046	0.539 ± 0.070	0.332 ± 0.083	0.562 ± 0.071	0.567 ± 0.082
	π (± SD)	0.01212 ± 0.00838	0.00333 ± 0.00349	0.00488 ± 0.00442	0.00867 ± 0.00649	0.00585 ± 0.00501	0.00579 ± 0.00494	0.00886 ± 0.00662	0.00602 ± 0.00507	0.00426 ± 0.00406	0.00641 ± 0.00529	0.00711 ± 0.00573
	Tajima's *D*	1.24144	-0.18909	-0.26401	1.20467	0.90902	-0.51045	0.38681	-0.44064	-0.49379	-0.34103	0.3598
	Fu F_S	0.554	-0.083	-0.473	1.079	-0.266	-1.115	-0.047	-1.012	-0.787	-0.879	0.099
	Rm	0	0	0	0	1	0	0	0	0	0	0

Table 2.13 Continued.

Population region		NE Spain	Southern M.C., France		M.C., France	Rhone, France	LOT, France		M.C., France
Locus	Population	TRJ	CTU	SPV	SAM13	DRAz4	LNP	RME	ROQ
Cyt b	N	23	24	24	24	20	24	13	15
	S	28	15	16	4	4	1	0	7
	h	13	10	7	4	4	2	1	6
	H_d (± SD)	0.818 ± 0.082	0.775 ± 0.080	0.754 ± 0.064	0.670 ± 0.069	0.363 ± 0.131	0.083 ± 0.075	0	0.648 ± 0.134
	π (± SD)	0.00524 ± 0.00304	0.00232 ± 0.00156	0.00248 ± 0.00165	0.00169 ± 0.00124	0.00066 ± 0.00066	0.00011 ± 0.00024	- -	0.00142 ± 0.00112
	Tajima's *D*	-1.85639*	-2.02917*	-2.03716	0.45663	-1.63814*	-1.15933	NA	-1.84878*
	Fu F_S	-4.216*	-4.662*	-1.112	0.733	-1.613*	-1.028	NA	-2.666*
	Rm	0	0	0	0	0	0	NA	0
EF-1α	N	23	24	23	24	22	24	23	
	S	9	3	4	2	3	2	1	
	h	11	4	6	3	4	3	2	
	H_O	0.348	0.542	0.696	0.375	0.455	0.250	0.231	
	H-W *P*	<0.001	0.144	1	0.255	0.393	1	1	
	H_d (± SD)	0.659 ± 0.072	0.630 ± 0.052	0.628 ± 0.051	0.441 ± 0.077	0.552 ± 0.035	0.228 ± 0.075	0.212 ± 0.097	
	π (± SD)	0.00128 ± 0.00098	0.00121 ± 0.00094	0.00113 ± 0.00090	0.00061 ± 0.00060	0.00077 ± 0.00070	0.00030 ± 0.00039	0.00027 ± 0.00038	
	Tajima's *D*	-1.45888	0.81104	-0.08585	0.10868	-0.28204	-0.87325	-0.31053	
	Fu F_S	-7.098*	0.67	-1.529	0.258	-0.505	-1.118	0.162	
	Rm	2	1	1	0	0	0	0	
AAm20	N	23	24	24	24	22	23	13	
	S	3	1	2	1	1	1	1	
	h	4	2	3	2	2	2	2	
	H_O	0.565	0.500	0.458	0.583	0.591	0.478	0.385	
	H-W *P*	0.774	1	0.677	0.684	0.386	0.628	0.577	
	H_d (± SD)	0.503 ± 0.062	0.496 ± 0.028	0.529 ± 0.026	0.511 ± 0.015	0.485 ± 0.037	0.414 ± 0.059	0.520 ± 0.028	
	π (± SD)	0.00603 ± 0.00564	0.00552 ± 0.00533	0.00613 ± 0.00569	0.00567 ± 0.00542	0.00539 ± 0.00526	0.00461 ± 0.00477	0.00578 ± 0.00558	
	Tajima's *D*	-0.43089	1.63398	0.40291	1.71946	1.54396	1.12856	1.61005	
	Fu F_S	-0.698	1.901	0.561	1.963	1.802	1.504	1.630	
	Rm	0	0	0	0	0	0	0	
AAm24	*N*	23	24	24	24	23	24	13	
	S	3	2	2	2	1	0	0	
	h	5	3	3	3	2	1	1	
	H_O	0.652	0.375	0.417	0.292	0.435	-	-	
	H-W *P*	0.723	0.489	0.813	0.053	1	-	-	
	H_d (± SD)	0.581 ± 0.067	0.414 ± 0.078	0.441 ± 0.077	0.377 ± 0.075	0.464 ± 0.045	-	-	
	π (± SD)	0.00695 ± 0.00559	0.00527 ± 0.00464	0.00544 ± 0.00474	0.00366 ± 0.00369	0.00383 ± 0.00380	-	-	
	Tajima's *D*	0.49057	0.74676	0.82829	-0.03311	1.426	NA	NA	
	Fu F_S	-0.661	0.884	0.957	0.101	1.731	NA	NA	
	Rm	1	0	0	0	0	NA	NA	

Legend N = Sample size; S = number of segregating sites; *h* = number of haplotypes; H_o = observed heterozygosity; (H-W *P*) = Hardy-Weinberg *P*-value; H_d = Haplotype (gene) diversity (Nei, 1987) (in diploid genomes is equivalent to expected heterozygosity); Nucleotide diversity (π) = average number of nucleotide differences per site between two sequences (Nei, 1987). #Significant *P*-values for Tajima's *D* (1989), Fu F_S (1997) tests when $P < 0.05$, after sequential Bonferroni correction in bold. Rm = number of recombination events as revealed by the four gamete model (Hudson and Kaplan, 1985).

variance was greatest, significant support for among region variance ($P < 0.05$) sub-divided the E Pyrenees, Massif Central (including Lot and Rhone valleys) and C Pyrenees/NE Spain (sub-divisions 1 to 4 Table 2.14). The validity of these regional grouping was supported by the homogeneity of within-region variances ($P > 0.05$) except for E Pyrenees from Massif Central (sub-division 2); but homogeneity was attained here by analyzing the south and north Massif Central independently (sub-divisions 5 and 6). Testing for population sub-division at the most polymorphic nuclear locus EF-1α showed lower but consistent sub-division resolution than cyt b. Among regional variation was supported ($P < 0.05$) between all three regions, the E Pyrenees from N Massif Central, and Massif Central from C Pyrenees/NE Spain, however, all comparisons tested supported ($P < 0.05$) genetic heterogeneity both within regions and populations.

Mismatch distribution was used to investigate the occurrence of sudden demographic expansion of cyt b haplogroups in those geographical regions supported by AMOVA: only two haplogroups (A and B) were used as these had sufficient numbers for these tests. For haplogroup A, the C Pyrenees plus NE Spain, E Pyrenees and the N Massif Central showed unimodal mismatch distributions (Figure 2.9) and significant signals of sudden expansion (Raggedness index $P > 0.05$) (Table 2.15). Using estimates of τ the expansions began mainly during the last glacial period (110,000-12,000 y.a.), except for in the N Massif Central, which dated to the Holocene using a 2.3% mutation rate (0-12,849 y.a.). Haplogroup B was only investigated within the E Pyrenees, where sub-haplogroup B05 was not expanding (Raggedness index $P < 0.05$), whereas sub-haplogroup B04 (Table 2.11, labelled "CB_B04") was and estimates of τ dated this event over three-fold earlier preceding the last glacial (420,168-182,682 y.a.) than any haplogroup A regional expansion (123,389-0 y.a.), however, confidence intervals between haplogroups overlapped.

Analysing all *P. ariasi*, following the cyt b result (above), globally IBD was also supported at all nuclear loci ($P < 0.05$ Table 2.16 and Figures 2.10 a, b, c), but for EF-1α and AAm24 less than 32% of the genetic variation was associated with geographical distance ($R^2 = 0.1463$ and 0.3213, respectively). Inter-regional comparisons did not support the overall results: EF-1α showed a considerable amount of variance around the regression line between pairs of populations between the Pyrenees and NE Spain with the Massif Central including Lot (SAM13, LNP and RME; data points $F_{ST}/(1-F_{ST}) > 0.25$ and < 400 km; Figure 2.10a); and for locus AAm24 statistical outliers (defined by

a z-test) included pairwise comparisons between the two Lot populations (LNP and RME) with outgroup CHR (Figure 2.10b). Locus AAm20 was monomorphic in Morocco, however, non-shared haplotypes still allowed this locus to follow IBD predictions (Figure 2.10c).

IBD was also used as an exploatory tools to infer alternative dispersal pathways circum-Pyrenees, but there was no greater statistical support for migrations through the western *vs*. eastern coastal foothills. Failure of this result is likely to be of consequence of scare sampling south of the Pyrenees. Observation of significance was the same for both models of dispersal for all results: one-dimensional or two-dimensional habitats estimated by $F_{ST}/(1-F_{ST})$ against geographical distance or log distance, respectively.

Tests were conducted to ascertain whether the effects of IBD were sufficient to explain AMOVA supported geographical regionality of the E Pyrenees and N Massif Central. At cyt b, marginal tests (DISTLM) also supported the correlation between genetic distance ($\Phi_{ST}/(1-\Phi_{ST})$ and geographical distance (38% variation explained P = 0.001) or geographical region (52%, P = 0.001), when using data points categorised into two discrete classes either within the E Pyrenees/N Massif Central *vs*. between region comparisons. Eliminating geographical distance by taking it as a covariate, a dbRDA analysis supported the AMOVA obtained sub-structure; categorization into within or between regional comparisons remained significantly (16%, P = 0.001) correlated to genetic distance. A result which supported the Carcassonne corridor as a habitat barrier between these two regions. At EF-1α the same comparison showed significant correlation between genetic distance and geographical distance (16%, P = 0.008), but not for categorising regional pairwise comparisons (< 1%, P = 1). This result might be explained by the confounding effects of bottle-necked LOT populations on within N Massif Central data. Further sampling would be needed to understand fully the observed regional effects shown in the AMOVAs.

Table 2.14 Testing for geographical regional population sub-structure by hierarchical AMOVA using 7 *a priori* sub-divisions. F Indices, percentage variation and *P*-values given for cyt b, *P*-values only for EF-1α. Calculations used *P. ariasi* associated with all cyt b haplogroups. Significant *P* values after 16,000 permutations: * < 0.05, ** < 0.01 *** < 0.001 (ARLEQUIN v3.11).

Sub-division tested	Cyt b				EF-1α
	df	F Indices	% variation	*P* - value	*P* - value
1. E Pyrenees vs. *Massif Central* vs. *C Pyrenees & NE Spain*					
Among regions	2	0.18494	18.49	<0.001***	*
Among pop within regions	13	0.01829	1.49	0.07218 ± 0.0026	***
Within pop	371	0.19984	80.02	<0.001***	***
2. E Pyrenees vs. *Massif Central*					
Among regions	1	0.25422	25.42	<0.001***	0.076
Among pop within regions	11	0.02525	1.88	0.0497 ± 0.0023*	***
Within pop	306	0.27305	72.69	<0.001***	***
3. E Pyrenees vs. *C Pyrenees & NE Spain*					
Among regions	1	0.08231	8.23	<0.05*	0.170
Among pop within regions	7	0.01963	1.80	0.11673± 0.0034	***
Within pop	234	0.10033	89.97	<0.05*	***
4. Massif Central vs. *C Pyrenees & NE Spain*					
Among regions	1	0.09306	9.01	<0.001***	*
Among pop within regions	8	0.00390	0.38	0.22307± 0.0039	***
Within pop	202	0.93562	90.61	<0.05*	***
5. E Pyrenees vs. *N Massif Central*					
Among regions	1	0.24455	24.46	< 0.01**	*
Among pop within regions	9	0.02714	2.05	0.05842 ± 0.0022	***
Within pop	260	0.26505	73.49	< 0.001***	***
6. E Pyrenees vs. *S Massif Central*					
Among regions	1	0.18052	18.05	< 0.05*	0.676
Among pop within regions	6	0.02829	2.32	0.09099 ± 0.0028	***
Within pop	215	0.20370	79.63	<0.001***	***

Legend Populations in geographical regions: eastern (E) Pyrenees, ARQ, CAT, IRL07, PAS, TUL; northern (N) Massif Central, LNP, RME, SAM13, DRAz4, ROQ; southern (S) Massif Central, CTU, SPV; central (C) Pyrenees, HP1, HP2; northeast (NE) Spain, TRJ. *P*-value represents the significance of the variance components and Φ_{ST} statistics tested under a permutation approach, whose null is panmixia at the different levels of hierarchy.

Table 2.15 Mismatch distribution statistics for *P. ariasi* cyt b haplogroups A (CB_A) and B (CB_BN) by geographical region supported by AMOVA. Sudden demographic expansion detected when significance of Raggedness index $P > 0.05$. Time elapsed since beginning of expansion event (t) calculated by $\tau = 2ut$. 95% confidence intervals of τ were estimated around mutation rates 2.3% and 1% at $\alpha = 0.05$. Generation time = 1 per annum; y.a. = years ago.

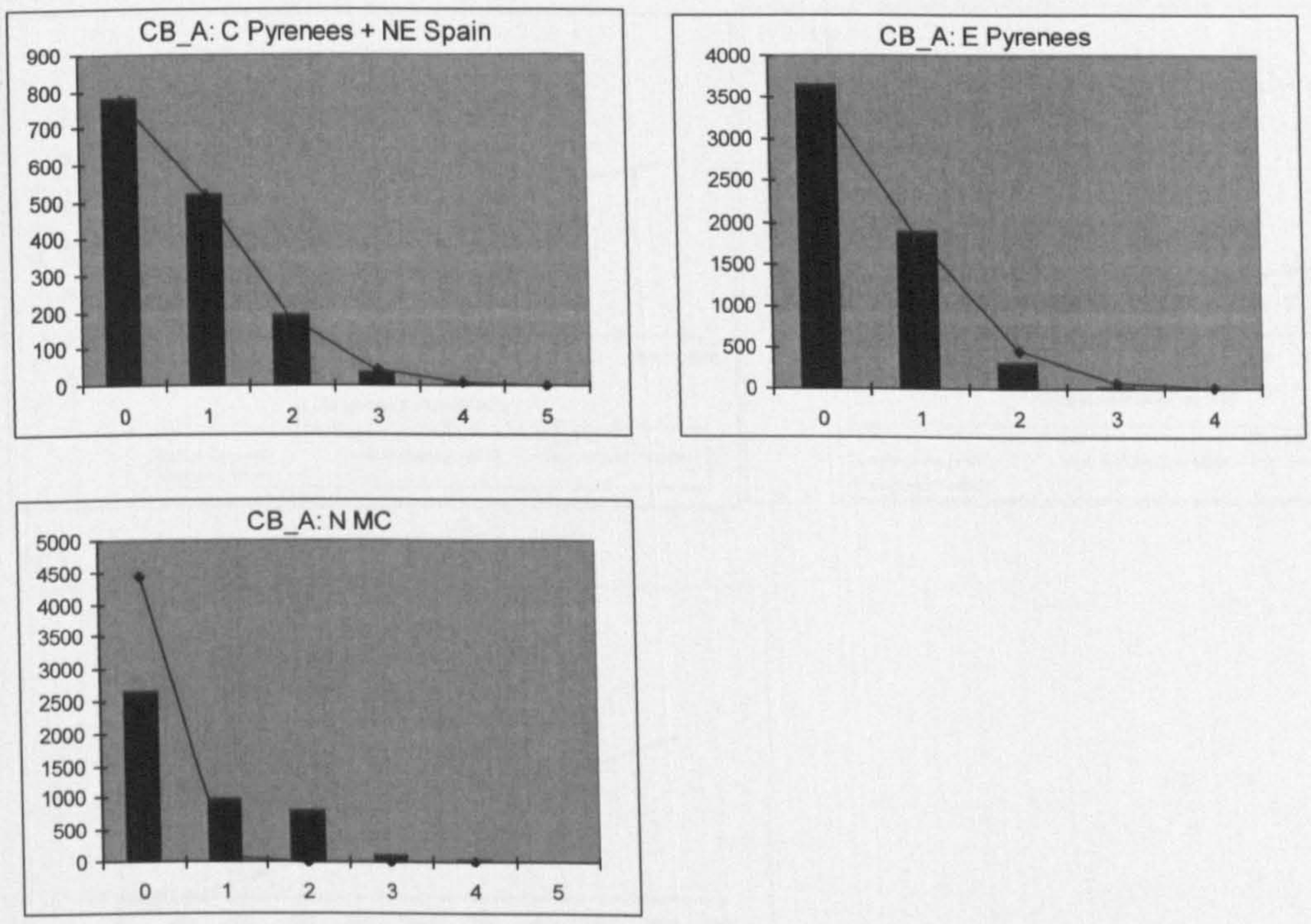
Geographical region	CB_A NE Spain + C Pyrenees	CB_A E Pyrenees	CB_A N MC	CB_B05 E Pyrenees
Raggedness index	0.08447	0.16437	0.16080	0.19644
Raggedness *P*	0.7717	0.4193	0.9995	0.0474
Tau (τ)	0.719 (0.00-1.43)	0.482 (0.26-0.76)	0.00 (0.00-0.211)	1.127 (0.52-2.02)
t 2.3% (y.a.)	43,783 (0-87,078)	29,351 (15,832-46,279)	0.00 (0-12,849)	NA
t 1% (y.a.)	100,700 (0-200,280)	67,507 (36,415-106,443)	0.00 (0-29,551)	NA

Table 2.16 Testing the association between genetic and geographical distance between *P. ariasi* populations is according to predictions of IBD: fitting estimates of $F_{ST}/(1- F_{ST})$ to geographical distance (km). Significance permuted using a Mantel test; * $P < 0.05$, ** $P < 0.01$ (GENEPOP v4.0).

	Fitting $F_{ST}/(1-F_{ST})$ to distance			Fitting $F_{ST}/(1-F_{ST})$ to ln distance		
Population region; locus	a	b	*P* (c>o)	a	b	*P* (c>o)
All populations; Cyt b	0.059483	0.0002576	0.009**	-0.28448	0.08250372	0.015*
All populations; AAm20	-0.1172216	0.00103504	0**	-1.32029	0.30024913	0**
All populations; AAm24	0.0490214	0.00057115	0**	-0.75881	0.19218457	0.001**
All populations; EF1α	0.1795947	0.00018567	0.003**	-0.1988	0.08365327	0.003**
EP + NMC; Cyt b	0.0353954	0.00046752	0.008**	-0.10283	0.04502734	0.005**
EP + NMC; EF-1α	0.0705571	0.00134991	0.03*	-0.24956	0.11510768	0.02*

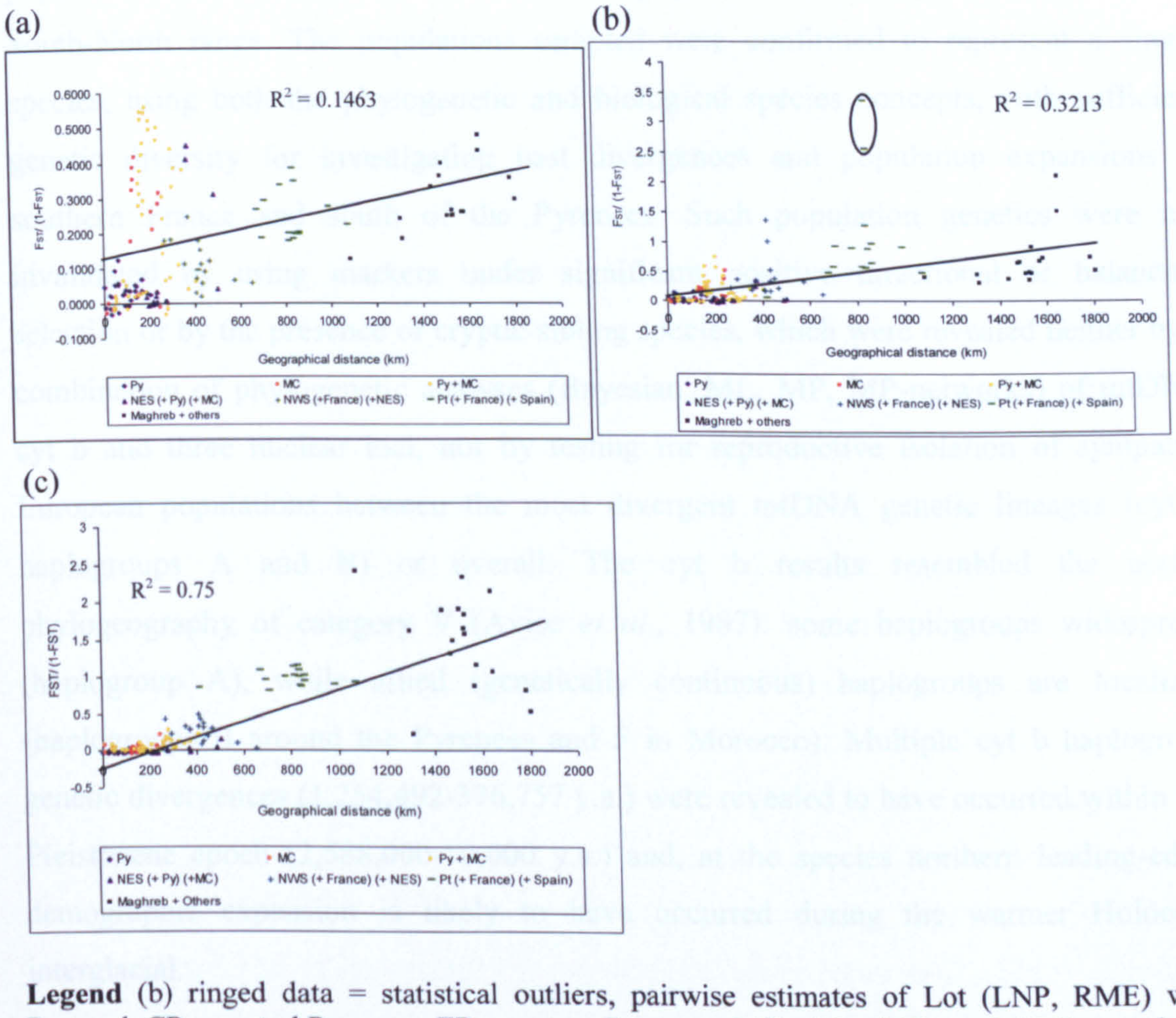
Legend EP = Eastern Pyrenees; NMC = northern Massif Central (including Lot and Rhone valleys)

Figure 2.9 Mismatch distributions for cyt b (sub-)haplogroups A and B by geographical region. Bars represent observed number of nucleotide differences between pairs of individuals; curves correspond to the mismatch distribution fitted to the data under an expected model of sudden demographic expansion (ARLEQUIN v3.11).



Legend Geographical regions: central (C) Pyrenees; northeast (NE) Spain; eastern (E) Pyrenees; north (N) Massif Central (MC).

Figure 2.10 Plots of genetic against geographical distance between populations of *P. ariasi*. (a) EF-1α; (b) AAm24; (c) AAm20. Extent of correlation given as R^2 values.



Legend (b) ringed data = statistical outliers, pairwise estimates of Lot (LNP, RME) with Portugal. CP = central Pyrenees; EP = eastern Pyrenees; MC = Massif Central; NES = northeast Spain; NWS = northwest Spain; Py = Pyrenees; SMC = southern Massif Central; Pt = Portugal.

2.4 Discussion

This chapter investigated for the first time the temporal and geographic population structure of *P. ariasi*, for which specimens were sampled from across its South-North range. The populations sampled were confirmed to represent a single species, using both the phylogenetic and biological species concepts, with sufficient genetic diversity for investigating past divergences and population expansions in southern France and south of the Pyrenees. Such population genetics were not invalidated by using markers under significant positive directional or balancing selection or by the presence of cryptic sibling species, which were revealed neither by a combination of phylogenetic analyses (Bayesian, ML, MP, MP-networks) of mtDNA cyt b and three nuclear loci, nor by testing for reproductive isolation of sympatric European populations between the most divergent mtDNA genetic lineages (cyt b haplogroups A and B) or overall. The cyt b results resembled the nested phylogeography of category V (Avise *et al.*, 1987): some haplogroups widespread (haplogroup A), while allied (genetically continuous) haplogroups are localized (haplogroups B around the Pyrenees and F in Morocco). Multiple cyt b haplogroup genetic divergences (1,254,492-376,757 y.a.) were revealed to have occurred within the Pleistocene epoch (2,588,000-10,000 y.a.) and, at the species northern leading-edge, demographic expansion is likely to have occurred during the warmer Holocene interglacial.

2.4.1 Locus neutrality and clock validity

Confidence that the population structure investigated described polymorphisms at four selectively neutral loci was obtained by phylogenetic and population genetic tests (PAML CODEML, MK test and Tajima's *D* statistic), which sought signals of long-term and recent selection. As expected for conserved protein-coding loci, evolution was driven by purifying selection pressures (Zhao *et al.*, 2003). However, this would not confound population inferences, but rather be of detriment in respect to the level of genetic variation accumulated.

Dates were estimated assuming a 'strict' molecular clock for each locus, which Drummond *et al.* (2006) argued to be biologically unrealistic and less appropriate than a 'relaxed' clock. Moreover, others have argued that it is invalid to extrapolate mutation rates across different evolutionary time-scales (Ho *et al.*, 2005; Penny, 2005). However, Weir and Schluter (2008) supported the use of a 2.1% mutation rate for cyt b, showing

its approximate maintenance over a 12 m.y. interval across numerous avian taxonomic orders. As *Phlebotomus* sandflies lack both outgroup and ingroup fossil records no rate curve could be estimated and uses of biogeographical calibration points are accompanied by their own confidence error (Esseghir *et al.*, 1997). Although *Phlebotomus* clock calibration was not possible, it was still valuable to estimate divergence times. I compensated for the lack of an accurate clock by the utilization of two standard mtDNA mutation rates (2.3% and 1%) and two sandfly generation times (1 or 3 p.a.). It follows that such a method for dating is approximate, and thus conclusions are discussed in the context of broad time-scales.

2.4.2 Quaternary genetic divergences and population expansions

Dating of coalescence and divergence events of *P. ariasi* by MDIV were consistent with the branching order of the Bayesian phylogeny (Figure 2.2) as supported by the parsimony network (Figure 2.4): haplogroup B branched first some 1.2 m.y.a-377 k.y.a., and the other branch (macrohaplogroup A) showed poor lineage sorting over a similar period. Coalescence and divergences were dated to within the Pleistocene: 2.2 m.y.a.-660 k.y.a. and 1.2 m.y.a-380 k.y.a. for 1% and 2.3% mutation rate, respectively, at 1 generation p.a. Dates are one third less for 3 generations p.a. and thus remain within the Pleistocene epoch (2,588,000-10,000 y.a.; Gibbard and van Kolfschoten, 2004) when speciation has been recorded for other organisms. Avise *et al.* (1998) calculated speciation duration within vertebrates to require at least two million years on average, and Ribera and Vogler (2004) showed most phylogenetic species of endemic Iberian water beetles to have diverged within the Pleistocene less than 1 m.y.a., with a few species pairs corresponding to as little as ~80,000 y.a. Both studies were based on a standard 2% mtDNA clock, comparable to the lower limit used in this study of 2.3% (Brower, 1994). Bayesian and ML cyt b gene trees indicated the presence of cryptic speciation in *P. ariasi* through the support of two primary monophyletic groups, haplogroups B and macrohaplogroup A. However, contrary to these results a thorough investigation rejected cryptic speciation of *P. ariasi*: nuclear gene phylogenies showed no obvious lineage sorting, although whether these markers are sufficiently polymorphic to resolve cryptic species or species complexes is questionable (Esseghir *et al.*, 2000; Parvizi and Assmar, 2007); single parsimony networks for each locus were reconstructed; and no evidence of reproductive isolation/biological speciation in *P. ariasi* was supported. In fact, biological speciation was not tested among water beetles,

and so some might only be 'phylogeographic species' with morphology fixed regionally by limited dispersal. In summary, *P. ariasi* supports the argument that although intra-specific divergence was initiated during the Pleistocene, which drove changes in population structure through species tracking favourable habitats (Coope, 1994; Hewitt, 1996), it was not a time of significant evolutionary divergence by adaptation (Knowles and Richards, 2005).

P. ariasi offers an intra-specific population history consistent with cold intolerant western Palaearctic species, where results revealed support for multiple refugia, multiple expansion events, and a zone of post-glacial secondary contact (here, north of the Pyrenees), in response to the Pleistocene's cyclical climate changes. The Pleistocene is reported to have generated significantly higher numbers of intra-specific haplogroups (some in Europe) than either the Pliocene or Miocene (Avise *et al.*, 1998). *P. ariasi* supports this, evidenced by five mtDNA divergence events (haplogroups) where the data indicated their restriction in multiple allopatric refugia - a result consistent with the refugia-within-refugia scenario (Gómez and Lunt, 2006).

Five cyt b haplogroups were hypothesized based on Bayesian support for the grouping of haplogroups A-C and E, and non-overlapping lower within-haplogroup differentiation versus higher between-haplogroup divergence. Avise *et al.* (1987) recognized haplotype grouping (a haplogroup) when the number of mutational steps between groups is greater than the maximum differentiation within a group. A similar approach of pairwise sequence difference within and between haplogroups was used by Naderi *et al.* (2007) when discussing their standard criteria for defining goat mtDNA haplogroups. However, Naderi noted that this threshold may be inadequate, because some haplotypes may lie at the boundary between within- and between-haplogroup pairwise differences, as observed in *P. ariasi.*

It is reported that intra-specific phylogenetic clades can form in a continuously distributed and spatially structured species, and not only as a consequence of geographical barriers to dispersal, cryptic species boundaries or recent contacts between historically allopatric populations (Irwin, 2002). However, at least two of the haplogroups reported in this chapter could reasonably be associated with biogeographical boundaries that isolate Iberia – the Gibraltar Straits in the south (haplogroup F), the Pyrenees in the north (haplogroup B) – and known refugia of similarly distributed species (Esseghir *et al.*, 2000; Lumaret *et al.*, 2002). However, as is often cited, caution should be taken when inferring a species' history based on a single

genealogy, where ideally independent estimates of the species tree should be considered in combination within a coalescent statistical framework (Ballard and Whitlock, 2004).

All cyt b coalescence and divergences were long after the final opening of the Gibraltar Straits, 5.5-4.9 m.y.a. (Steininger and Rogl, 1984), and so any vicariant evolution was caused by other barriers, probably related to the climate oscillations that heightened in the early Pleistocene (Hewitt, 2004a). Moreover, the lack of discontinuous genetic variation between haplogroup F (Morocco) from European haplogroups (e.g. phylogenetic polytomy and reticulate network loops within macrohaplogroup A) argues against the Gibraltar Straits as a long-term zoogeographical barrier to gene flow (category I; Avise *et al.*, 1987). Only more widespread sampling might indicate the origins of the Pleistocene "Eve" of *P. ariasi* and if haplogroup F is restricted to the Atlas region where it was found. However, the current phylogeographic distribution of European/Moroccan haplogroups is only possible if at least one made an intercontinental crossing. Dispersal across the Gibraltar Straits has been identified in flying insects (Schmitt *et al.*, 2005; Lozier and Mills, 2009). However, *P. ariasi* is not a strong flier, with wind speeds > 1.5 m/sec and > 4-5 m/sec inhibiting and stopping flight (Killick-Kendrick *et al.*, 1984). It is therefore more likely that movement of *P. ariasi* across the Gibraltar Straits occurred during times when Pleistocene climates caused short periods of (incomplete) drying, resulting in vegetated islands that permitted stepping-stone dispersal (Flemming *et al.*, 2003; Cosson *et al.*, 2005; Carranza *et al.*, 2006).

Cyt b haplogroups did not correspond directly to geographical populations. Therefore, the approach of O'Loughlin *et al.* (2007) was followed by constructing separate mismatch distributions for unstructured (sub-)haplogroups, in order to infer whether there had been expansions experienced by each lineage rather than assessing the structure of mixed-ancestry populations. The modelling of sudden (rapid) demographic expansion only, was supported for four cyt b (sub-)haplogroups using the predictions of mismatch distribution (haplogroups A, C, F and sub-haplogroup CB_B04 of haplogroup B; Table 2.11). Despite the large confidence intervals for dating, expansions occurred much later than their divergences, and are likely to have occurred in the Pleistocene. The oldest population expansion estimate of *P. ariasi* (CB_B04 ca. 420 k.y.a.) related its post-glacial re-colonization in Iberia/France to MIS 12 (some 433 k.y.a.), one of the two coldest Pleistocene glacials as evidenced by the highest $\delta^{18}O$ benthic stack (Lisiecki and Raymo, 2005). Times of haplogroup expansions and their

current distributions suggest a later geographical replacement of haplogroups B and C by A throughout northern Iberia and France. Haplogroup A predominates to the exclusion of all other haplogroups in the uplands of the Massif Central, where the species is most unlikely to have survived at the LGM, and this suggests a post-glacial spatial expansion northwards in the Holocene. Immigration into France was more likely from NE Spain, based on its broad littoral region that has often been warmer than NW Spain (Delmas *et al.*, 2008), regional similarities at all loci, and low F_{ST} differentiations that grouped NE Spain and southern France apart from those of NW Spain. A similar distribution pattern was observed in the phylogeography of *Quercus ilex* chlorotypes (Lumaret *et al.*, 2002), a floral indicator species for *P. ariasi* (Rioux *et al.*, 1984).

Although mtDNA detected significant changes to population structure of *P. ariasi* during the Pleistocene, no discrete lineages were resolved at any nuclear loci. The unresolved network pattern observed could be explained if the isolation event(s) that promoted divergence of fast-evolving mtDNA cyt b were not of sufficient duration to cause discrete lineage sorting in slower evolving nuclear markers. Then, secondary contact during post-glacial re-colonization would cause genetic homogenization. Indeed, IBD was supported for nuclear markers. It was often shallow, with the exception of comparisons with leading-edge populations, which is consistent with the paradigm of northern purity for temperate species (Hewitt, 2004a).

2.4.3 Refugial populations north of the Pyrenees during the late glacial period

Cyt b phylogeographic structure was observed where haplogroup B was found to be omnipresent in northern Spain and Pyrenean France, but absent in Morocco, Portugal and the northern Massif Central. Sub-haplogroup B CB_B04 was limited to the E Pyrenees only, and its expansion date (420-183 k.y.a.) was early enough for it to have reached the northern slopes of the Pyrenees before the last glacial period, 110-12 k.y.a. (Gibbard and van Kolfschoten, 2004) and to have been an endemic ever since. Alternatively, the current endemicity of this sub-haplogroup in the E Pyrenees might be explained by its arrival in a phalanx-type mass immigration of all haplogroups (A-C) at the start of the current interglacial, probably from NE Spain as previously reasoned. However, this is unlikely because of the absence of sub-haplogroup B CB_B04 in N Spain and its older and disparate expansion time. This study suggests CB_B04 is more likely to be a marker for a population that survived north of the Pyrenees during one or more glacial periods, before its refuge was invaded more recently and rapidly by

abundant interglacial dispersers with haplogroup A. Similarly, it is also possible that sub-haplogroup CB_B05 and haplogroup C are markers for northern refugial populations, but the sampling in this thesis did not allow for their refugia to be inferred.

As observed for sub-haplogroup B CB_B04, endemic or relict populations have not always been the source of major post-glacial re-colonizations (Bilton *et al.*, 1998; Petit *et al.*, 2003; Segarra-Moragues *et al.*, 2007). Endemic/relict populations are often much older than any other population in their range, because they have often persisted longer in isolation (Hampe and Petit, 2005). However, caution must be exercised when inferring a refuge in southern France, because of the sparse sampling of *P. ariasi* in Iberia. The ability of *P. ariasi* to survive *in situ* north of the Pyrenees during glacial periods can only be inferred from its current bioclimate envelope. Its hibernating larvae can survive for weeks at 2-7°C (Ready and Croset, 1980), and the oaks characteristic of its favoured humid and sub-humid Mediterranean bioclimates (Rioux *et al.*, 1984) left pollen traces of their survival in southern France during the last glacial (Beaudouin *et al.*, 2007). However, these oaks flourish in colder climes (Deciduous white oak, *Q. pubescens*) and drier climes (Evergreen, *Q. ilex*) (Jalut *et al.*, 2009) than *P. ariasi*, and there are doubts about the interpretation of the pollen record (Calvet, 2004). The snow-line on the northern face of the E Pyrenees is now much higher (2,700-2,800 m.a.s.l.) than it was at the last glacial maximum, 1,400-1,500 m.a.s.l. (Calvet, 2004), when the upper bound of *P. ariasi* abundance could have dropped from the current ca. 1,500 m.a.s.l. (Rioux and Golvan, 1969) to near sea-level, as it tracked suitable habitats driven by the oscillating climates. Also the Pyrenees is a known region of endemics/relicts of temperate species, a supported refugium within the Atlantic-Mediterranean differentiation centre (e.g. Deffontaine *et al.*, 2009; Gómez and Lunt 2006 and references within). The current study suggests this region was a refugium for a sub-tropical species. Moreover, the current range of haplogroup B CB_B04 could be limited by the local environment, because the Mediterranean climate does not extend far to the west of the river Aude (Calvet, 2004).

2.4.4 Recent post-glacial re-colonizations not blocked by refugial populations north of the Pyrenees

The route or routes of dispersal into France from Iberia could not be ascertained through IBD analysis, as sampling in Iberia was insufficient. Hewitt (1996; 1999) highlighted the effects of different modes of dispersal on the genetic diversity of re-

colonizing populations. This study indicates that the dispersal mode of *P. ariasi* has often been phalanx - along a broad dispersal front, typified by IBD and the mixing of cyt b haplogroups. Long-range pioneer dispersal (leptokurtic) often produces small fragmented pockets of genetically homogeneous populations with high inter-population differentiation (Ibrahim *et al.*, 1996). This was not typical of *P. ariasi*, which showed large geographical regions of homogeneity e.g. no significant genetic differentiation (F_{ST}) and shallow IBD or complete gene flow for most loci across the French Pyrenean slopes or within the Massif Central. Long-range pioneer dispersal would not be expected of *P. ariasi* because of a flight range limited to 0.1-2 km (Killick-Kendrick *et al.*, 1984), and long-distance gene flow has been observed in other sandflies in Europe e.g. populations of *P. perniciosus* sampled over an area of 500 km are genetically homogeneous (Aransay *et al.*, 2003). Cyt b haplogroups were sympatric in the northern Pyrenees, a zone of secondary contact. Phalanx-like dispersal of *P. ariasi* is least likely to have been blocked by small refugial or relict leading-edge populations that had survived glacial periods, and this fits with finding cyt b haplogroup A predominating over haplogroups B and C in the E Pyrenees. The re-colonization of *P. ariasi* could have kept pace with that of its associated woodland, which spread at a rate of 50+ m.p.a. (Hewitt, 1999), and for holm oaks produced an Iberia-Italy hybrid zone in the Rhone valley (Lumaret *et al.*, 2002).

2.4.5 Monopolization currently blocking the northward spread of Pyrenean sandflies and potentially of leishmaniasis

This study suggests that cyt b haplogroup A is a marker for the most recent (128-36.9 k.y.a.) and dominant expansion of *P. ariasi* in Europe, and this probably originated south of the Pyrenees because of the high frequency of this haplogroup and its modal haplotype (CB25) in northern Portugal and Spain. It predominates north of the Pyrenees, to the exclusion of all other haplogroups in bottle-necked, leading-edge populations (low H_d and π diversity of cyt b and EF-1α) in the Massif Central and the nearby Lot and Rhone valleys. There was a step change in F_{ST} values between the E Pyrenees and the Massif Central for cyt b and EF-1α, with the rarity of haplogroups B and C in the Massif Central being detected by AMOVA. The step-change observed between the Pyrenees and Massif Central, which was beyond the affects of IBD as shown in a dbRDA and is likely to be explained by the absence of a forest structure suitable for *P. ariasi* dispersal in the lowland corridor between them.

It could be hypothesized that the stepping stone dispersal across this corridor is being further blocked by the "monopolization" (Loeuille and Leibold, 2008) of the Massif Central by sandfly populations characterized by cyt b haplogroup A. Leading colonizers are believed to establish the allelic content of a 'population', where they can act as a barrier to dispersal for later colonizers (Nichols and Hewitt, 1994). If flies of the Massif Central were found to be relatively poor dispersers or vectors, this would hinder the spread of zoonotic leishmaniasis to northern France. Actually, leishmaniasis foci in the Massif Central are distinctive, characterized by low diversity of regional *L. infantum* strains, high disease prevalence in domestic dogs (the reservoir), frequent cutaneous lesions but low prevalence of symptomatic visceral leishmaniasis in humans, and a preponderance of *P. ariasi* (Pratlong *et al.*, 2004). Re-forestation of the lowland corridor between the two southern uplands might increase gene flow and alter the population structure characteristics of *P. ariasi*. The population differentiation of *P. ariasi* is unlikely to match that of *P. perniciosus*, because this alternative sympatric vector peaks at lower altitudes, in hotter and drier bioclimates (Rioux *et al.*, 1984), and has two cyt b lineages (Iberia, Italy-N Africa) mixing in France (Perrotey *et al.*, 2005).

CHAPTER 3

No contemporary arms race involving the sandfly salivary peptide apyrase: implications for vaccination against Mediterranean zoonotic leishmaniasis

3.1 Introduction

The natural mode of *Leishmania* promastigote transmission is by regurgitation of the parasite (Schlein *et al.*, 1992) in the saliva of infected adult female sandflies (Diptera: Psychodidae) into host haemorrhagic feeding pools (Shortt and Swaminath, 1928; Ribeiro, 1987a; 1995). To counteract their host's protective haemostatic, inflammatory and immunomodulatory responses to capillary laceration, the female sandfly secretes a suite of potent pharmacological substances into her saliva (Ribeiro and Francischetti, 2003). In experimental models, these salivary peptides have been shown to change the course of *Leishmania* infection, either having a protective or exacerbating effect (e.g. Oliveira *et al.*, 2008), and so might be used for vaccination against leishmaniasis (Valenzuela *et al.*, 2001a; Palatnik-de-Sousa, 2008). The suitability of a candidate salivary peptide should be based not only on the knowledge of its effect on *Leishmania*, but also on the degree and nature of the evolutionary processes driving its natural genetic polymorphism. The aim of this chapter is to investigate the systematics and population genetics of the salivary peptide apyrase of *P. ariasi* and other *Phlebotomus*, which in the former produces a delayed-type hypersensitivity (DTH) in a mouse model (Oliveira *et al.*, 2006), a cellular immunity consistent with protection against *Leishmania* (Kamhawi *et al.*, 2000).

The enzyme apyrase has the most abundant transcript in a salivary gland cDNA library of *P. ariasi* (Oliveira *et al.*, 2006). The apyrases are ubiquitous in vertebrates, plants and non-haematophagous arthropods, with a role in nucleotide catabolism (Sarkis *et al.*, 1986). However, as a salivary peptide of haematophagous arthropods the apyrases (E.C. 3.6.1.5) are adapted to be secreted and function as a potent anti-platelet factor by hydrolysing di- and tri-phosphates, e.g. ADP and ATP, the central activators in host haemostasis that are released by both injured cells and during platelet aggregation (Ribeiro *et al.*, 1987a; Marcus and Safier, 1993; Valenzuela *et al.*, 1996; 1998). Highly active in the salivary glands of haematophagous arthropods that have evolved to blood feed independently (Grimaldi and Engel, 2005), the apyrase enzymes have been acquired by convergent evolution (Sarkis *et al.*, 1986). Consequently the apyrases are

classified into three independent protein families (reviewed Champagne and Valenzuela, 1996; Hamasaki *et al.*, 2009), the apyrase of sandflies being a member of the *Cimex* family, uniquely dependent on Ca^{2+} alone (Valenzuela *et al.*, 2001b), and with homologues in human and mouse (Valenzuela *et al.*, 1998).

Experimental models have demonstrated the roles that salivary peptides of sandflies play in changing *Leishmania* pathogenicity. In rodent and dog models, immunization with sandfly saliva (or homogenate) or distinct peptides protects against *Leishmania* infections through a T_h1-associated cytokine interferon-γ (IFN-γ) (DTH) cell-mediated immunity. Conversely, some peptides exacerbate parasite load and subsequently the course of infection and this is correlated with a humoral T_h2-associated cytokine interleukin-4 response (Belkaid *et al.*, 1998; MBow *et al.*, 1998; Kamhawi *et al.*, 2000; Oliveira *et al.*, 2008; Collin *et al.*, 2009). Immunization of rodent models by a DNA plasmid expressing apyrase of *P. ariasi* was mentioned above. However, the role of the apyrase of *P. ariasi* or any other *Phlebotomus* in such protection has not been investigated, perhaps as a consequence of research often focusing on peptides associated with a host antibody and T_h1 response (Collin *et al.*, 2009), the former apyrase of *P. ariasi* has not been recorded to induce (Oliveira *et al.*, 2006). How anti-salivary immunity works is not fully understood, so an antibody response may not be integral for protection: immunity is suggested to occur through creating an unsuitable environment for *Leishmania* development or acceleration of specific anti-*Leishmania* immunity or a combination of both (Collin *et al.*, 2009).

Actually, sandfly salivary peptides are receiving attention as a component of anti-*Leishmania* vaccines, because they control pathogenicity and are a permanent feature in the natural transmission cycle. Anti-*Leishmania* vaccines have been experimentally developed using species-specific salivary peptides of two sandflies; New World *Lutzomyia longipalpis* and Old World *P. papatasi*. Anti-*L. major* peptides targeted include anti-MAX (Morris *et al.*, 2001), and an SP15-DNA vaccine (Valenzuela *et al.*, 2001a; Oliveira *et al.*, 2008). In visceral leishmaniasis (VL) models, immunisation with salivary peptides of *L. longipalpis* in hamster (anti-LIM19) and dogs (anti-LJL143 and -LIM17 in the natural reservoir of VL) conferred protection and parasite killing by a T_h1 with DTH response (Gomes *et al.*, 2008; Collin *et al.*, 2009). To date no studies have investigated the effects of an anti-apyrase vaccine, or any salivary peptide in vectors of Mediterranean ZVL.

Apyrase is potentially a broad spectrum vaccine, having been recorded across four genera/subgenera: *Lutzomyia*, *Phlebotomus*, *Euphlebotomus* and *Larroussius* (Anderson *et al.*, 2006). However, like many salivary peptides its natural polymorphism among sandfly species and their geographical populations has not been well studied. The activity levels of salivary peptides can vary geographically (Warburg *et al.*, 1994), can have high intra-specific amino acid divergence (Lanzaro *et al.*, 1999) associated with differences in antigenicity (Milleron *et al.*, 2004a), and can be weakly cross-reactive (Volf and Rohoušová, 2001).

Two factors inform our understanding of the molecular evolution of salivary peptides, firstly their level of sequence evolution and secondly what processes drive their rate of change. An evolutionary arms race (presumed genetic) (Dawkins and Krebs, 1979) has been postulated between parasitic *Leishmania* species or strains and their mammalian hosts, or natural vectors (Ribeiro, 1987b; Handman, 1999; Beverley and Dobson, 2004). Although not proven, arms race scenarios in the sandfly peptide-host-*Leishmania* triad might be expected based on other insect-borne diseases, including tsetse fly-borne sleeping sickness caused by antigen-switching *Trypanosoma brucei* (Young *et al.*, 2008) and anopheline mosquito-borne malaria caused by *Plasmodium* species with highly polymorphic surface antigens (Tetteh *et al.*, 2009).

Immunity genes evolving under a classical arms race model (e.g. Endo *et al.*, 1996; Jiggins and Kim, 2007) may show positive directional selection driving to fixation a succession of adaptive alleles characterized by elevated inter-specific nonsynonymous amino acid substitutions but a lack of intra-specific polymorphism (Hurst and Smith, 1999; Ford, 2002; Olson, 2002). Alternatively, an arms race driven by balancing selection (Spurgin and Richardson, 2010) will favour polymorphism with multiple adaptive alleles being maintained within a species' populations and giving rise to high heterozygosity and many ancient alleles (e.g. Gilbert *et al.*, 1998; Garrigan and Hedrick, 2003). There is some evidence that sandfly salivary peptides are subject to selection. The presence of multiple maxadilan peptide alleles maintaining vector antigen polymorphisms of *L. longipalpis* was hypothesized to be driven by [balancing] selection as a strategy against host immune system response, but no statistical support for selection was presented (Milleron *et al.*, 2004b). In contrast, adaptive evolution was rejected for the salivary peptide SP15 of *P. papatasi* (Elnaiem *et al.*, 2005), which may result from an incomplete analyses.

My literature review for this thesis found no studies that had characterized the genetic variation in the salivary peptides of a European sandfly species or a vector of *L. infantum*. Furthermore, no study to date has conducted an investigation at both the subgenus and population levels, in order to assess the contribution of neutral versus selective processes operating on a phlebotomine salivary peptide. The aim of this chapter was to investigate the genetic evolution of the salivary peptide apyrase, a peptide that could putatively modify parasite transmission. As the interaction between sandfly peptide-host-*Leishmania* may be subject to an evolutionary arms race through the adaptive pressures of cyclic antagonistic positive directional, or balancing selection, these types of selection were tested for. If apyrase is a potential salivary peptide based candidate for an anti-*Leishmania* vaccine, this study will contribute to the understanding of its molecular evolution both across *Phlebotomus* and within *P. ariasi*.

This chapter's aims were:

1. To design primers to target a fragment of apyrase that contained sites likely to be evolving under selection i.e. calcium and nucleotide binding sites (Dai *et al.*, 2004), ADPase sites (Yang and Kirley, 2004), and putative MHC class II T cell epitopes (Kato *et al.*, 2006).
2. To characterize this fragment of apyrase in nine *Phlebotomus* species, in addition to 20 natural populations of *P. ariasi*.
3. To use selection tests, mainly based either on within-gene heterogeneity of nonsynonymous versus synonymous substitution rates or the allele frequency spectrum, to investigate (a) evidence of long-term evolutionary selection in apyrase of distantly related *Phlebotomus* taxa, and (b) evidence of selection in populations of *P. ariasi* potentially exposed to varying selection pressures associated with their differing ecological niches.
4. To investigate the contribution of demography to the genetic variation of apyrase, by additionally characterizing markers cyt b and EF-1α, loci that have shown no evidence to be subject to positive directional or balancing selection in *P. ariasi*. This was an objective because some statistical tests of selective neutrality assume demographic equilibrium, which is often violated in natural populations.

3.2 Materials and methods

3.2.1 Specimen sampling and preparation

The apyrase alignment of *Phlebotomus* species included: (i) GenBank accessions *P. (Phlebotomus) papatasi* (AF261768); *P. (Phlebotomus) duboscqi* (DQ834331, DQ834335); *P. (Euphlebotomus) argentipes* (DQ136150); *P. (Adlerius) arabicus* (EZ000631, EZ000632, EZ000633); *P. (Larroussius) perniciosus* (DQ192490, DQ192491); *P. (Larroussius) ariasi* (AY845193); (ii) novel apyrase sequences characterized in this thesis from flies or DNA provided by collaborators (see specimen donation acknowledgements, page 22), namely *P. (Adlerius) brevis* (3 flies), *P. (Adlerius) halepensis* (6 flies), and, from the subgenus *Larroussius*, *P. major* (3 flies), *P. neglectus* (7 flies), *P. kandelakii* (4 flies), *P. perfiliewi* (4 flies) and *P. tobbi* (4 flies); and, (iii) novel apyrase sequences characterized from *P. perniciosus* (6 flies), *P. ariasi* (471 flies) and *P. (Transphlebotomus) mascittii* (2 flies) from the collections made for this thesis.

Investigation of population genetic variation of apyrase was evaluated using natural populations of *P. ariasi* collected from 20 locations in one or two summers of 2005-2008. As well as the populations described in Chapter 2, two further populations were characterized: code MLQ (Le Bousquet, Aude, France, 43.0179 N, 1.8424 E) from a high altitude location (1114 m.a.s.l.) in the northeast Pyrenees; and code PLB (Limbrassac, Aude, France, 42.7459 N, 2.1672 E) from a low altitude population (385 m.a.s.l.) in the northeast Pyrenees. Table 3.1 summarizes the population characteristics in relation to their varying natural environments: altitude as a proxy for temperature; forest fragmentation as a proxy for density/genetic bottle-necking; and, associations with hosts.

Specimens were caught and stored as detailed in Chapter 2 section 2.2.1.

3.2.2 Molecular characterization

DNA extraction of sandflies was carried out according to Chapter 2.

3.2.2.1 Polymerase Chain Reaction (PCR) amplification, purification and direct sequencing

Loci cyt b and EF-1α were characterized as described in Chapter 2.

Table 3.1 Population characteristics of *P. ariasi* molecularly characterized in relation to environment and hosts.

Population Code	N	Location	Altitude (m a s l)	Forest fragmentation	General habitat: host(s)
AGH	17	Western Atlas, Morocco	1250	High	Outside road-side wall
CHR	24	Northern Portugal	499	High	Inside rural dwelling: [poultry near]
CSP	24	Northwest, Spain	585	-	Outside rural dwelling: Canidae, Felidae, poultry
HP1	27	Central Pyrenees, France	607 - 789	Intermediate	Outside road-side wall
HP2	19	Central Pyrenees, France	307 - 429	High	Outside road-side wall
PAS	54	Eastern Pyrenees, France	647	Low	Outside rural garden: Canidae
PLB	24	Eastern Pyrenees, France	385	High	Inside farm: Canidae, Leporidae, poultry, Bovinae
MLQ	35	Eastern Pyrenees, France	1114	Intermediate	Inside farm: Bovinae
TUL	24	Eastern Pyrenees, France	467	Low	Inside rural dwelling: Canidae, poultry
IRL07	22	Eastern Pyrenees, France	421	Low	Inside rural dwelling: Canidae
ARQ06	44	Eastern Pyrenees, France	382	Low	Inside/outside farm: Canidae, Bovinae, Equidae
ARQ08	23	Eastern Pyrenees, France	382	Low	Inside/outside farm: Canidae, Bovinae, Equidae
CAT	16	Eastern Pyrenees, France	600	-	Outside farm: Bovinae
TRJ	23	Northeast, Spain	332	High	Inside farm: Leporidae [Canidae near]
CTU	24	Southern Massif Central, France	358	Low	Inside rural garden: Poultry [Leporidae near]
SPV	24	Southern Massif Central, France	351	Low	Inside rural dwelling: Canidae, Leproidae, poultry
SAM13	24	Massif Central, France	510	-	Garden wall
DRAz4	23	Massif Central, France	100	Low	-
LNP	24	Lot, France	316	High	Inside farm: poultry [Bovinae and Leporidae near]
RME	13	Lot, France	270	Intermediate	Inside/outside farm: Canidae, Bovinae, Leporidae, poultry

Table 3.2 Novel primers and PCR conditions for the amplification and direct sequencing of the apyrase gene fragment of *P. ariasi* and * other *Phlebotomus* species. (Tm = one-/or two-step annealing temperature. † Starting nucleotide in GenBank accession AY845193).

Primer APY	† 5' nt	Primer sequence 5' - 3'	Primer pair APY	Tm (°C)	$MgCl_2$ (mM)
-1F	110	CAACMAGATTCATCCCT TTYGC	-1F/ -3R*	52/56	1.5
-3R	652	CCAATTTACRGCCTCATGCCA	-1FC/ -3R	63	0.7
-1FC	199	TATGGCGAATTGAAGGACAAC	-1FT/ -3R	60	0.5
-1FT	198	ATATGGCGAATTGAAGGACAAT	-1FC/ -3RCG	63	0.7
-3RC	582	TTGACTCTTCCAATTGATG	-1FC/ -3RT2	63	0.7
-3RT2	582	GCTGTATTGACTCTTCCAATTGATA	-1FT/ -3RCG	60	0.5
-3RCG	564	GTTGGTGACTTCGCCTTCC	-1FT/ -3RT2	60	0.5
-47G	136	ATCTCCGACTTGGACAAGAAG	-47G/ -3R	64/62	1.0
-161G	248	GTCAAAATCTTCACTACTTCACG	-161G/ -3R	64/62	1.0
-474C	583	TGTATTGACTCTTCCTATTGAC	-1F/ -474C	64/62	1.0

Locus apyrase

At the start of this thesis, a 1161 bp DNA sequence of *P. ariasi* (GenBank AY845193) had been isolated from a cDNA library constructed from salivary gland mRNA (Oliveira *et al.*, 2006). No primers had been published for PCR amplification and direct sequencing of this apyrase. Four of the six least divergent sandfly apyrases in GenBank (2006) were aligned as ca. 1100 bp DNA sequences (data not shown) and as 336 deduced amino acids (Figure 3.1): *P. argentipes* SP03 (GenBank accession DQ136150), *P. ariasi* SP01 (AY845193), *P. perniciosus* SP01 (DQ192490) and *P. perniciosus* SP01B (DQ192491).

A 'conserved' primer pair APY-1F with APY-3R was designed to target a fragment of apyrase 563 bp in length (including primers) from *P. ariasi* and *P. perniciosus*, namely amino acids 26-213 (Anderson *et al.*, 2006) (Figure 3.2). The forward primer (APY-1F) annealed to the first conserved nucleotide section (22 bases) at the 5' of the gene, with 100% nucleotide similarity between the two species. The 3' terminus of the reverse primer (APY-3R) was positioned downstream of an insertion – codons 202 to 203 in *P. argentipes* and *P. perniciosus* but absent in *P. ariasi* – and included four well conserved codons (WHEA) in Old and New World sandflies and in *Cimex* (Anderson *et al.*, 2006). The insertion was included as it might be potentially informative for investigating apyrase evolution.

According to mutagenesis studies of homologues to *Phlebotomus* apyrases, the human calcium activated nucleotidases (CANs), the targeted fragment captured most sites essential to apyrases' anti-haemostatic function as a platelet aggregation antagonist (Figure 3.1): 9 out of 13 and 3 out of 6 nucleotide and calcium binding sites, respectively, (Dai *et al.*, 2004); 3 out of 4 residues essential to nucleotidase activity, a single site known for ADPase nucleotidase function [of invertebrates] and 5 out of 5 point mutations converting human apyrase to a potent anti-platelet aggregation agent (Yang and Kirely, 2004). It also contained 4 out of 6 MHC T-cell epitopes inferred for *P. dubosqci* (Kato *et al.*, 2006).

Nucleotide sequences were aligned and edited as described in Chapter 2 section 2.2.3. Direct sequencing from the PCR product of conserved primer sequences showed genotypes with more than one ambiguous site. Therefore, allele specific primers were designed and associated PASA parameters optimized to preferentially amplify one allele over another, allowing unambiguous scoring of apyrase genotypes. PCR thermocycling parameters for all apyrase primers used a 'hot start' at 85°C, and an initial single 3 min

Figure 3.1 Amino acid alignment of GenBank apyrases from three *Phlebotomus* sandflies: *Phlebotomus argentipes* (DQ136150); *Phlebotomus perniciosus* (DQ192490), (DQ192491); *Phlebotomus ariasi* (AY845193). Conserved and similar amino acids shaded in black and grey, respectively. **x** = conserved forward (APY-1F) and reverse (APY-3R) primers. Functional sites have been reported for: nucleotide binding (+) and calcium binding (*) in the human homologues (Dai *et al.*, 2004); after *in vitro* mutagenesis of the human homologue, o are essential to APDase activity, ▲ single residue mutation from Glu to Tyr with high associated ADPase nucleotidase activity; and ^ (carets under sequence alignment) point mutations that convert the wild-type human CAN into 100-fold more potent ADPase which abolishes platelet aggregation (Yang and Kirely, 2004). [] Brackets enclose putative MHC epitopes in the sandfly *P. dubosqci* (Kato *et al.*, 2006).

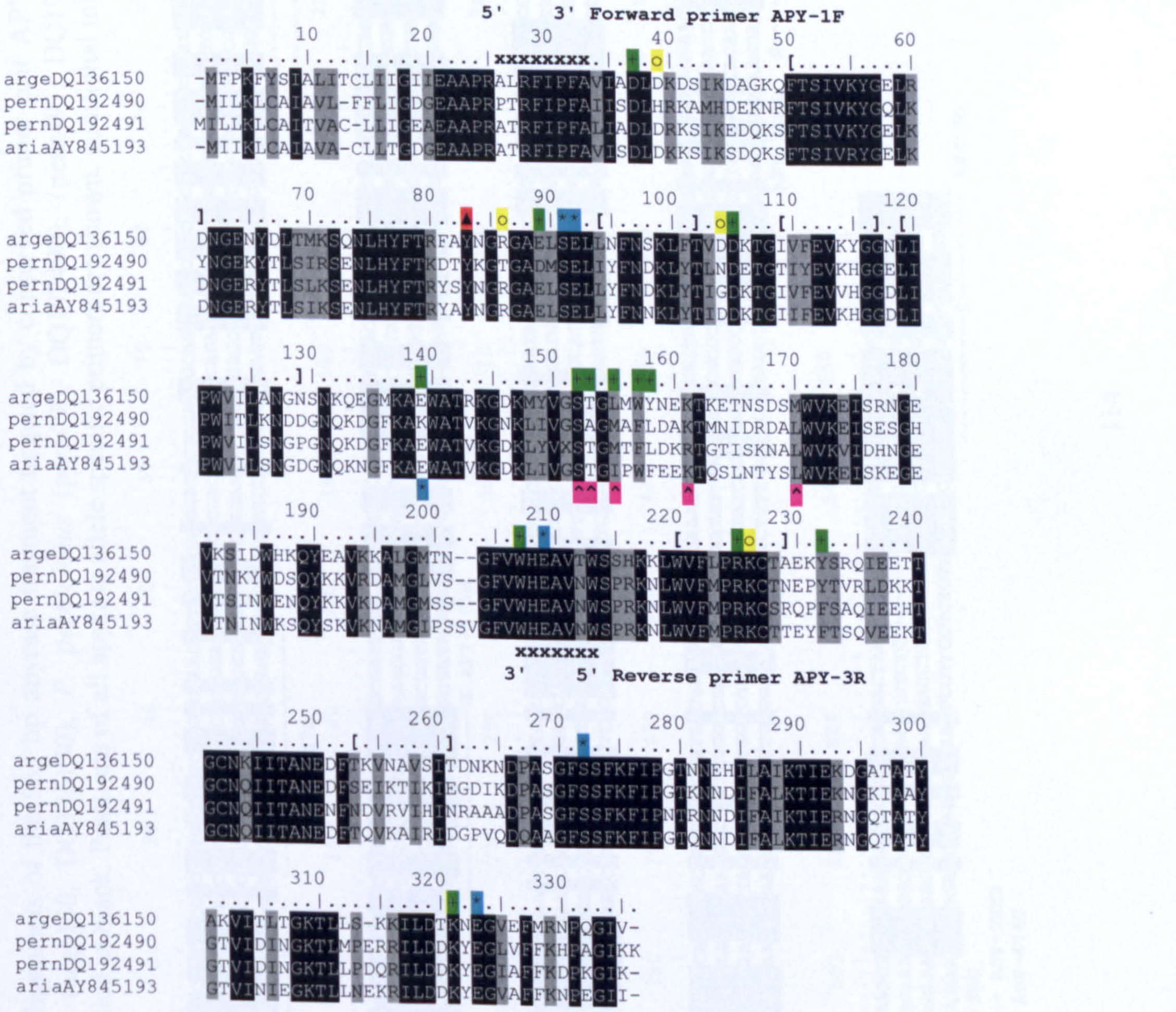
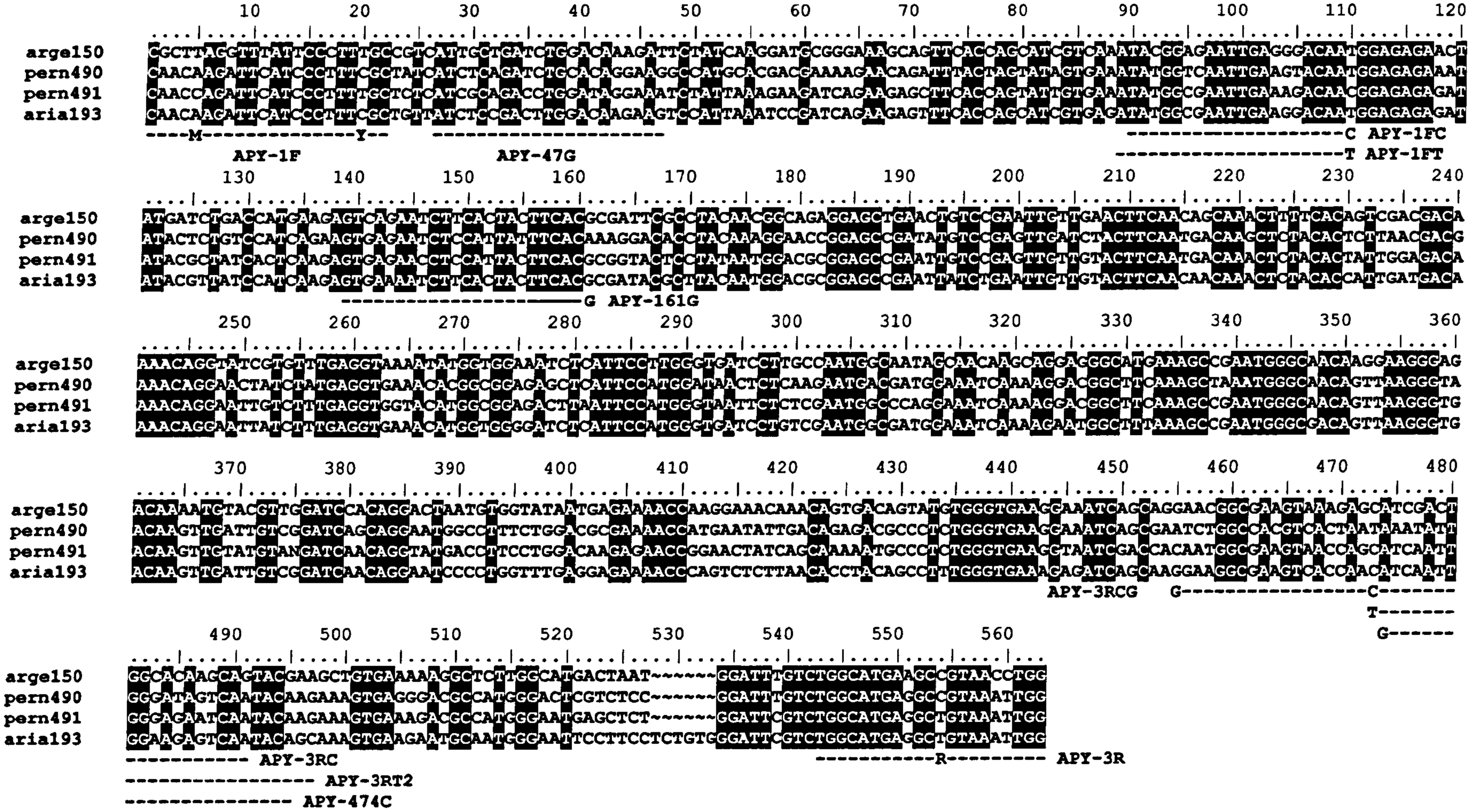


Figure 3.2 Nucleotide alignments of the 563 bp apyrase fragment targeted by conserved primer pair APY-1F and APY-3R, at three *Phlebotomus* sandflies. *P. argentipes* (arge150, DQ136150), *P. perniciosus* (pern490, DQ192490), (pern491, DQ192491), *P. ariasi* (aria193, AY845193). Conserved nucleotides filled in black. Positions of all apyrase allele specific primers are shown, additional information give in Table 3.2.



denaturation step at 94°C, followed by a single set (35 cycles) or two sets (5, 35 cycles) of denaturation (30 sec at 94°C), annealing (30 sec at specific temperature, see Table 3.2), and extension step (90 sec 72°C). A final 72°C extension step for 10 min terminated the amplification. Concentrations of PCR reagents were standard as described for cyt b and EF-1α in Chapter 2. However, $MgCl_2$ concentration was optimized for individual primer pairs, which in addition to varying annealing temperatures and multiplexing different combinations of conserved and allele specific primers, aimed to achieve maximum specificity and efficiency of PASA (Table 3.2).

3.2.2.2 Cloning of apyrase in *Phlebotomus*

The conserved apyrase fragment had to be cloned for four *Larroussius* species with duplicate loci: *P. kandelakii*, *P. perfiliewi, P. perniciosus* and *P. tobbi*. Ten clones were sequenced from each species' library, built using DNA from two specimens and the TOPO TA cloning® kit (Invitrogen™). Here the manufacturer's protocol was followed and DNA plasmids isolated from bacterial colonies by alkaline lysis during miniprep purification. The cloned sequences in the purified plasmids were amplified by PCR using kit primers T7 and T3 (0.5 μM final concentration) with standard PCR concentrations and thermocycling parameters (35 cycles at 47°C annealing temperature) as described for apyrase conserved primers (section 3.2.2.1). Sequencing in a single direction used 1 pmol of primer T3: primer T3 targets a fragment of the plasmid which lies outside of the cloned fragment; this circumvents the need to sequence in both directions to gain the full sequence of the PCR product. Full protocols for TOPO TA cloning® kit (Invitrogen™) and minipreps are in Appendix 3.1.

3.2.3 Data analyses

3.2.3.1 Inter-species phylogenetic analysis and divergence

Alignments of *Phlebotomus* apyrase direct nucleotide sequences were made using the automated algorithm in MAFTT [default settings and 2 iterations] (v6 for Mac OS X: http://align.bmr.kyushu-u.ac.jp/mafft/software/macosx.html), and post-processed manually in BIOEDIT (Hall, 1999). Phylogenetic relationships among alleles were reconstructed by Bayesian estimation (MRBAYES v3.1.2; Ronquist and Huelsenbeck, 2003), using MRMODELTEST (v2.3; Nylander, 2004) to select nucleotide substitution models for each codon position (as described in Chapter 2). Various outgroup species sequence combinations were used to ascertain (i) the orthology of apyrase alleles

characterized; (ii) the phylogenetic location of gene duplication events if observed; (iii) input phylogenies for PAML to test for branch and site based positive selection; (iv) to conclude whether apyrase in *P. ariasi* is single or multilocus; and, (iv) to identify cryptic sibling species.

Divergence was estimated from pairwise alignments of *Phlebotomus* apyrase amino acids by percentage similarity and identity (MATGAT v2.01 for Windows using the BLOSUM62 scoring matrix; Campanella *et al.*, 2003), and nucleotide divergence (K), based on the average number of nucleotide substitutions per site between species (Nei, 1987), using Jukes-Cantor correction (DNASP v4.90.1; Rozas *et al.*, 2003).

3.2.3.2 Intra-species genealogy

A parsimony network was reconstructed in TCS (Clement *et al.*, 2000) as described in Chapter 2, to show the genealogical relationship between all *P. ariasi* apyrase nucleotide sequences, to identify intra-specific apyrase lineages, phylogeographic allele associations and signals of demographic events, i.e. population expansions.

3.2.3.3 Apyrase protein structure assessment

Protein structural analyses (MACVECTOR v11.0; MacVector, Inc.) assessed whether phylogenetic or genealogical amino acid substitutions at known apyrase functional sites (cation binding or ADPase activity) were associated with changes in protein secondary structures. Changes in beta sheets, alpha helices and turns according to Chou-Fasman, Robson-Garnier and their consensus were plotted.

3.2.3.4 Detecting selection on branches of the *Phlebotomus* apyrase phylogeny

The CODEML program of Phylogenetic Analysis by Maximum Likelihood (PAML v4.2; Yang, 2007) tested for heterogeneity in selection pressure and positive selection on the apyrase phylogeny, based on nonsynonymous/synonymous substitution rate changes (ω). Branch lengths of Bayesian topologies were re-estimated based on the number of nucleotide substitutions per codon, and a likelihood ratio test (LRT) under a chi-squared distribution selected either the use of no clock (unrooted phylogeny, each branch having an independent rate) or global clock (rooted phylogeny, all branches having the same rate).

Branch models were implemented as described in Chapter 2. This method avoids averaging ω over long periods of time. However, it is a conservative test of positive selection, because averaging over the whole gene dilutes the signals of positive selection at particular sites with those of strong purifying selection acting on most sites of a functionally constrained protein (Yang, 2002). Therefore, heterogeneity of sites [codons] was tested using:

(1) Fixed-site codon models A and E (according to control file of Yang and Swanson, (2002)), to test for positive selection in codons partitioned into *a priori* classes of: (i) buried (class 1) and cation binding sites (exposed class 2); (ii) buried and known ADPase functional sites; (iii) buried and putative epitope sites. Model A assumes a homogeneous model, whereas Model E assumes different transition/transversion rate ratio (k), codon frequencies, ω, and proportional branch lengths, for the two class partitions.

(2) Random-site models M1a versus M2a, and M7 versus M8, to test for positive selection at particular sites assuming no prior knowledge of site partitioning. In null model M1a (nearly neutral) two site classes are designated, conserved sites $\omega < 1$ and neutral sites $\omega = 1$. M2a (selection) introduces a third site class $\omega > 1$. Null model M7 (beta; neutral) allows ω to vary among sites according to a beta distribution, that is restricted between 0 to 1, whereas model M8 (beta and ω; selection) adds a discrete ω class that is free to be estimated > 1. Bayes Empirical Bayes (BEB) method was implemented for models M2a and M8 (selection) to calculate the posterior probability that each site was from a particular site class: sites with high posterior probabilities from class $\omega > 1$ with $Pr > 0.95$ were inferred as codons to be under positive selection (Yang *et al.*, 2005).

3.2.3.5 Detecting selection on apyrase within the *P. ariasi* lineage

Codon usage bias can be indicative of strong selection, estimates obtained using DNASP (v4.90.1) included: the Codon Bias Index (CBI93) (Morton, 1993); Effective Number of Codons (ENC) (Wright, 1990); and, the G+C content at synonymous third coding positions (G+C3s) to directly quantify usage bias in the fraction of third positions in codons that are G or C. Additionally, selection was investigated using population based neutrality tests. The McDonald-Kreitman test (MK) (1991) and the associated Neutrality Index (NI) (Rand and Kann, 1996) were implemented to seek evidence of, and assign direction to, selection based on estimates of nonsynonymous to

synonymous polymorphisms and divergence. Computation in DNASP (v4.90.1) of MK, NI and appropriate outgroup choice were described in Chapter 2. To detect weaker and more recent signals of selection, three intra-population tests based on skews in the allele frequency spectrum from neutral expectation were implemented (ARLEQUIN v3.11; Excoffier *et al.*, 2005). The Fu and Li *D* statistic is the only one to consider the timing of selection by incorporating an outgroup sequence. It measures departures from neutral expectation using θ (based on K) derived from the total of singleton mutations on derived branches compared with the total on ancient branches in a phylogeny. Directional selection increases the number of derived mutations, whereas balancing selection causes a deficiency, giving rise to negative and positive *D* values, respectively. Tajima's *D* measures the difference between estimates of $\theta\pi$ (based on the average pairwise nucleotide differences between sequences) and θ_S (based on the number of segregating sites) relative to their standard errors: positive values arise from an excess of alleles at intermediate frequencies and are consistent with balancing selection; and, negative values arise when an excess of low frequency alleles which inflates θ_S and indicates directional selection. The Ewans-Watterson (EW) test identifies recent selection by assessing the deviation of the observed homozygosity from that expected based on sample size and number of alleles: negative values arise from a deficiency of homozygosity and indicate balancing selection; whereas positive values arise from an excess of homozygosity and signal directional selection.

Natural populations, such as those being examined here, often violate assumptions of the neutral model, where demographic processes can cause changes in the expected values of S, K and π leading to neutral deviations from mutation-drift equilibrium. It follows that demographic processes can mirror signals of selection, leading to false inferences of the latter. In general, selective sweeps and population expansion mimic signals of directional selection, and population size decreases and sub-division mimic balancing selection. Population expansion was investigated using the Fu *Fs* test (Fu, 1997) (in ARLEQUIN v3.11): significantly large negative deviations from neutral expectations arise from an excess of recent singleton mutations and reveal recently expanding populations (or selective sweeps). For all tests, significant ($P < 0.05$) departure from neutral expectation was calculated using 16,000 coalescence simulations and when significant, multiple tests were manually corrected for familywise Type 1 errors by applying the sequential Bonferroni correction of Holm (1979) at α 0.05.

In addition, deviations from Hardy-Weinberg equilibrium (HWE) were assessed to reveal whether selection occurred in the present generation (e.g. balancing selection drives an excess of observed heterozygotes, whereas directional selection causes deviation towards the fixation and thus excess of homozygotes). LD was tested for non-random association of alleles between nuclear-nuclear or cyto-nuclear loci (GENEPOP v4.0), which can indicate epistatic selection for gene combinations (Lewontin, 1964), or a selective sweep (Kim and Nielsen, 2004). In a simple neutral model effects of selection are locus specific, whereas demographic effects are genome wide, therefore genetic pressures of neutral versus selective processes on apyrase in *P. ariasi* populations were also distinguished by conducting comparative statistical analyses cyt b and EF-1α which are not under positive directional or balancing selection (Chapter 2).

3.2.3.6 *P. ariasi* nucleotide sequence composition & recombination at the apyrase locus

Extent of intra-population DNA polymorphism for apyrase of *P. ariasi* was measured by haplotype diversity (*h*) (Nei, 1987) and average pairwise nucleotide diversity per site (π) (Nei, 1987), as well as, independently for synonymous (π_s) and non-synonymous sites (π_n). Spearman's rank correlation coefficient was used to evaluate whether within gene recombination rate was significantly correlated to nucleotide diversity (π). Estimates of recombination parameter R (=4*Nr*) between adjacent sites, where *N* is the effective population size and *r* is the recombination rate per generation between the most adjacent nucleotide sites, were estimated according to Hudson (1987). The minimum number of recombination events to occur along the apyrase sequence were identified by the parameter Rm, calculated using the four gamete model (Hudson and Kaplan, 1985). All tests were calculated in DNASP (v4.90.1).

3.2.3.7 Population genetics

The following were implemented as first described in Chapter 2: F_{ST} estimates of genetic distance between population pairs; dependence between genetic distance $[F_{ST}/(1-F_{ST})]$ and geographical proximity of population pairs (Rousset, 1997); Analysis of Molecular Variance (AMOVA in ARLEQUIN v3.11) to evaluate the amount of haplotype diversity correlated with different nested levels of hierarchical population sub-division; distance-based redundancy analysis (dbRNA, DISTLM v.5) to examine the extent to which genetic differentiation is correlated to geographical regionality, beyond that explained by geographical distance, to identify barriers to gene flow.

3.3 Results

3.3.1 *Phlebotomus* apyrase gene structure and lineages

Novel conserved primers APY-1F with APY-3R successfully amplified the apyrase fragment from all *Phlebotomus* species targeted, including clones of species *P. kandelakii*, *P. perfiliewi*, *P. tobbi* and *P. perniciosus*. As only unambiguous alleles were included in the species phylogeny, the single ambiguous nucleotide in *P. perniciosus* (DQ192491) (nt 375, Figure 3.2) was inferred as a guanine (G), this being invariant across *Phlebotomus*. An alignment of all unique apyrase amino acid alleles identified in this thesis, in addition to *Phlebotomus* GenBank sequences published up to 01/09/2009 is given in Appendix 3.2, and an alignment of all unique apyrase nucleotide alleles obtained in this thesis is given in Appendix 3.3.

The initial Bayesian phylogeny was reconstructed based on a multi-species 154-amino acid alignment without introns and indels, starting on nucleotide 166 in GenBank accession AY845193 (*P. ariasi*). This phylogeny included GenBank sequences from the subgenera *Phlebotomus* (AF261768, *P. papatasi*; DQ834331/5, *P. duboscqi*) and *Euphlebotomus* (DQ136150, *P. argentipes*), but the absence of congruence with other gene trees (Chapter 2) indicated the inappropriateness of these distant outgroups. This incongruence may be due to the presence of paralogous apyrases. However, as *Phlebotomus* and *Euphlebotomus* are distant outgroups in this phylogeny, a resolution of orthology versus genetic distance will require a more extensive species' sampling than the current one. Therefore, to be conservative these sequences were removed from all subsequent analyses. A further conservative choice for all subsequent Bayesian reconstructions was the removal of *P. brevis*, because it's apyrase also showed incongruent branching with other gene trees.

The following Bayesian reconstructions used all the available alleles for each species except for *P. ariasi*, for which 31 out of 47 alleles were selected by pruning terminal branches. Strong support (posterior probability, pp, 1) was found for treating species of the subgenera *Adlerius* and *Transphlebotomus* as outgroups for *Larroussius* (Figure 3.3a). In *Larroussius*, the monophyly both of *P. major* and *P. neglectus* was supported (pp 1). *P. ariasi* was also monophyletic in the Bayesian tree and the 95% TCS parsimony criterion reconstruction, this supports the conclusion in Chapter 2 that the current samples did not contain cryptic sibling species. A duplicate lineage paralogous to that of *P. ariasi* indicated paraphyly of four species *P. kandelakii*, *P.*

perfiliewi, *P. perniciosus* and *P. tobbi*: each grouping on two well supported independent lineages (pp 0.96 and 1). This result suggests the occurrence of a single gene duplication event, limited within *Larroussius*, prior to the speciation of *P. kandelakii* and its sister clade. The duplicate lineages are consistent with the result of Anderson *et al.* (2006) who characterized two *P. perniciosus* alleles each of which grouped in two independent lineages. Accordingly, these lineages will be referred to as either pern490 or pern491[2]. Taking branch length as a measure of molecular distance (mutations per site), an episodic period of rapid evolution was observed immediately after the gene duplication event in pern490 lineage (Branch A in Figure 3.3a).

Many gene relationships remained ambiguous. On branches A/B *P. perniciosus, P. tobbi* and *P. perfiliewi* formed an unresolved tricotomy (pp < 0.7). The basal branch of *Larroussius* was not consistent, dependent on the apyrases included: phylogenies that included pro-orthologues [sequences pre-dating the gene duplication event] putative orthologues and duplicated lineages (Figure 3.3a), or pro-orthologues and orthologues to *P. ariasi* only (Figure 3.3b), were concordant with the nuclear gene elongation factor-1α with the *P. major* complex as basal (Chapter 2) but failed to group the two members of the *P. perniciosus* complex (*P. perniciosus* and *P. tobbi*). Conversely, the phylogeny reconstructed using pro-orthologues and the putative paralogous pern490 lineage only (see next section), supported *P. ariasi* as being basal within *Larroussius* followed by the *P. major* complex (Figure 3.3c). This is consistent with the gene tree for mitochondrial cyt b (Chapter 2). Concordant apyrase phylogenies were reconstructed excluding functional sites, and including or excluding putative epitopes (data not shown).

3.3.2 Divergence, structure and selection of apyrase lineages of *Phlebotomus*

The duplicate lineage paralogous to that of *P. ariasi* and the three basal species was identified as branch A/B (Figure 3.3a), in part based on the lower amino acid similarities/identities (81.8-83.8/62.8-64.9%) compared with the orthologous duplicate lineage (89.6-91.6/77.3-83.8%; branch C). Moreover, nucleotide divergence (K) was lower between these putative orthologues than with the paralogues, 0.194-0.24 and 0.296-0.316, respectively (Table 3.3). BLASTx searches were conducted to assess the

---

[2] The terminology of the two *P. perniciosus* apyrases of Anderson *et al.* (2006) of SP01 and SP01B were not followed in this study, as the label of the GenBank sequences given was not consistent with the text and phylogeny of their publication. Therefore, the last three numbers of the GenBank accession were definitive; *P. perniciosus* DQ192490 (pern490) and DQ192491 (pern491).

Figure 3.3 Bayesian phylogenies of the 462 nucleotide apyrase fragment, including all alleles of each *Phlebotomus* species except *P. ariasi* (set pruned of APY alleles > 1 step from modes in TCS network). Species of the subgenera *Transphlebotomus* and *Adlerius* are sister to the subgenus *Larroussius*, which contains vectors of *L. infantum*. (Posterior probabilities > 0.7 indicate statistically supported nodes. Solid ellipse marks the gene duplication event. Uppercase letters refer to branches tested in PAML models). (a) Complete apyrase gene tree including pro-orthologues and post-duplicate lineages, (b) putative orthologous apyrases only, and (c) pro-orthologues and the duplicate lineage paralogous to *P. ariasi*. Scale bars are in units of nucleotide substitutions per site.

(a)

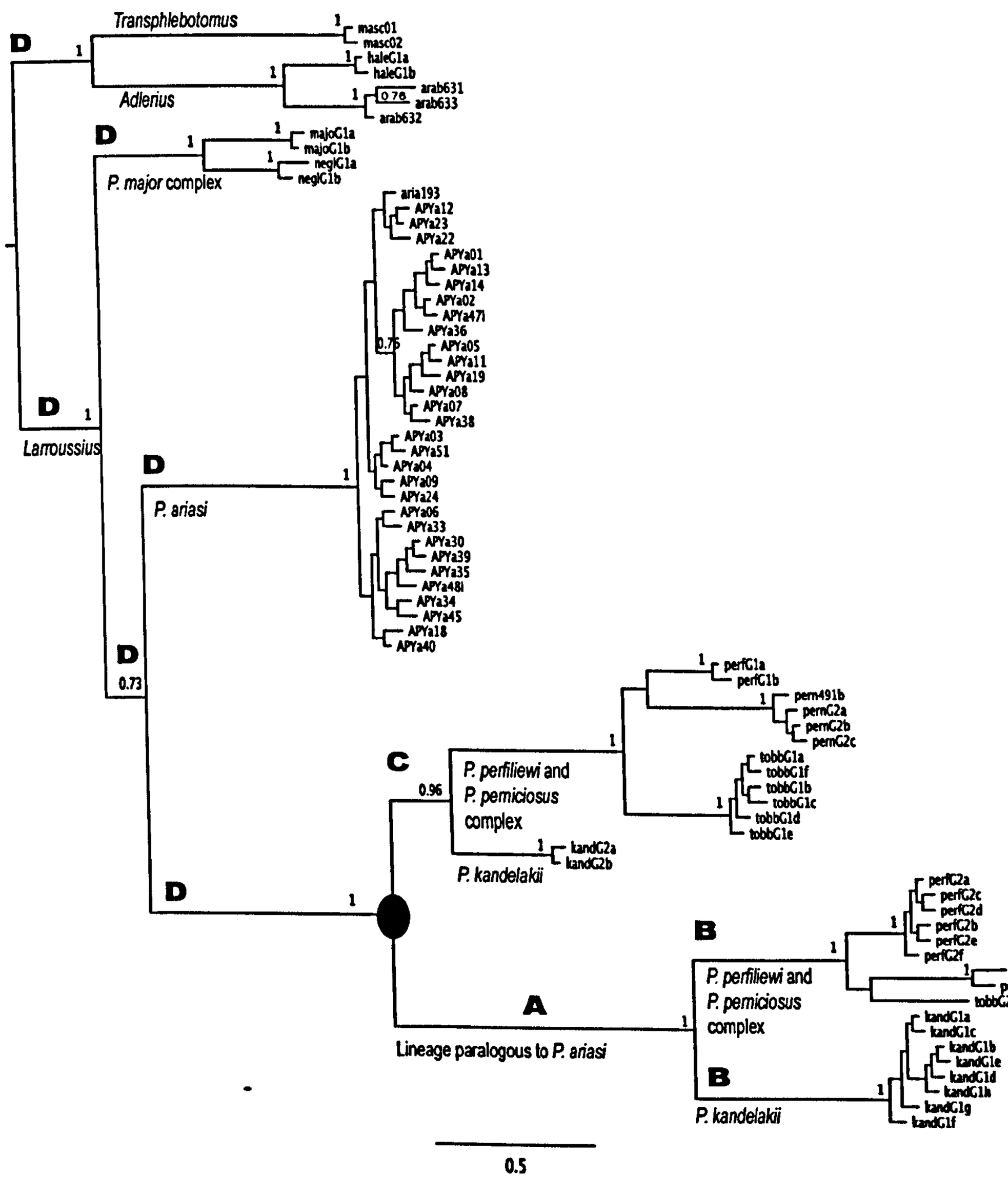
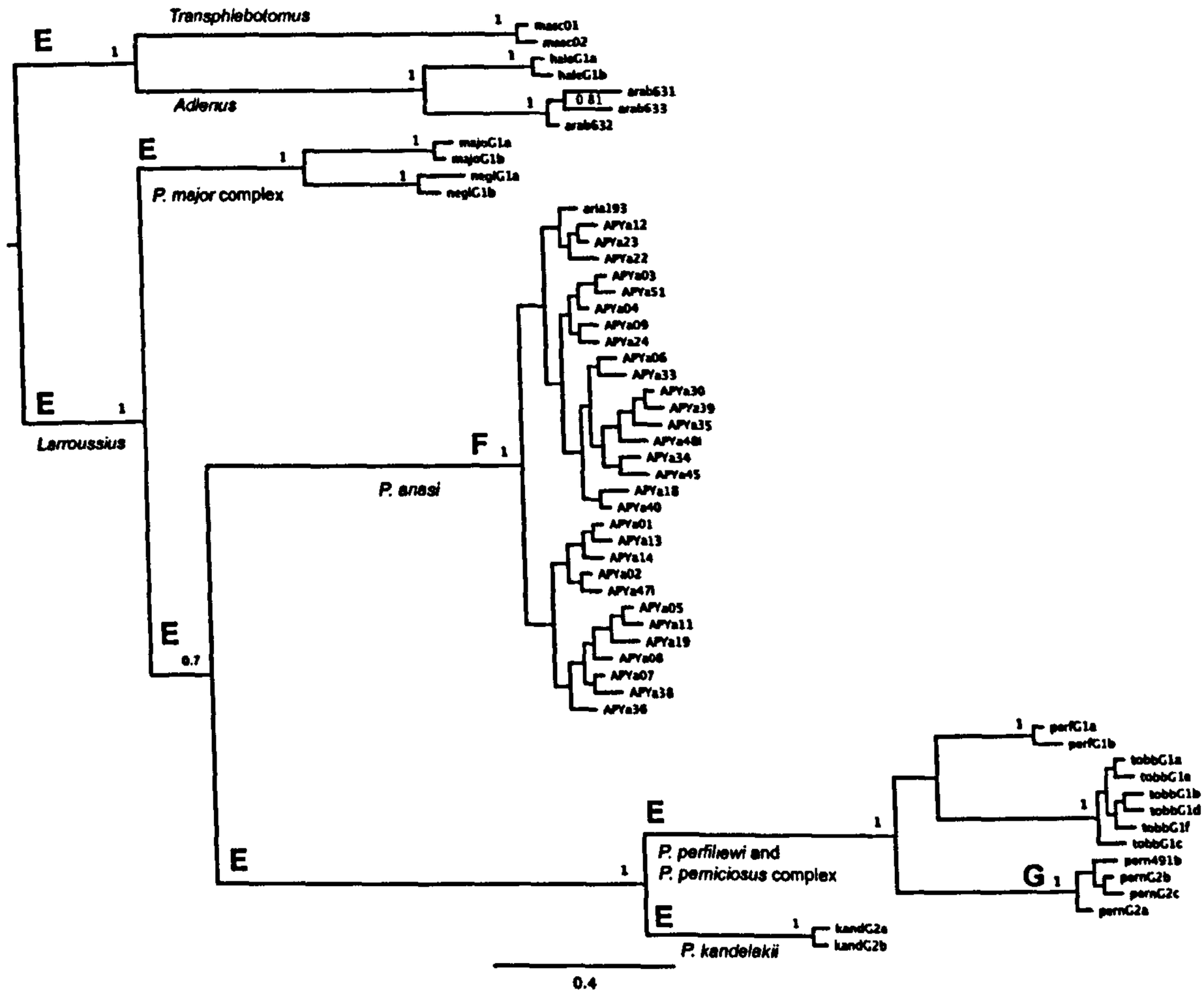


Figure 3.3 Continued.

(b)



(c)

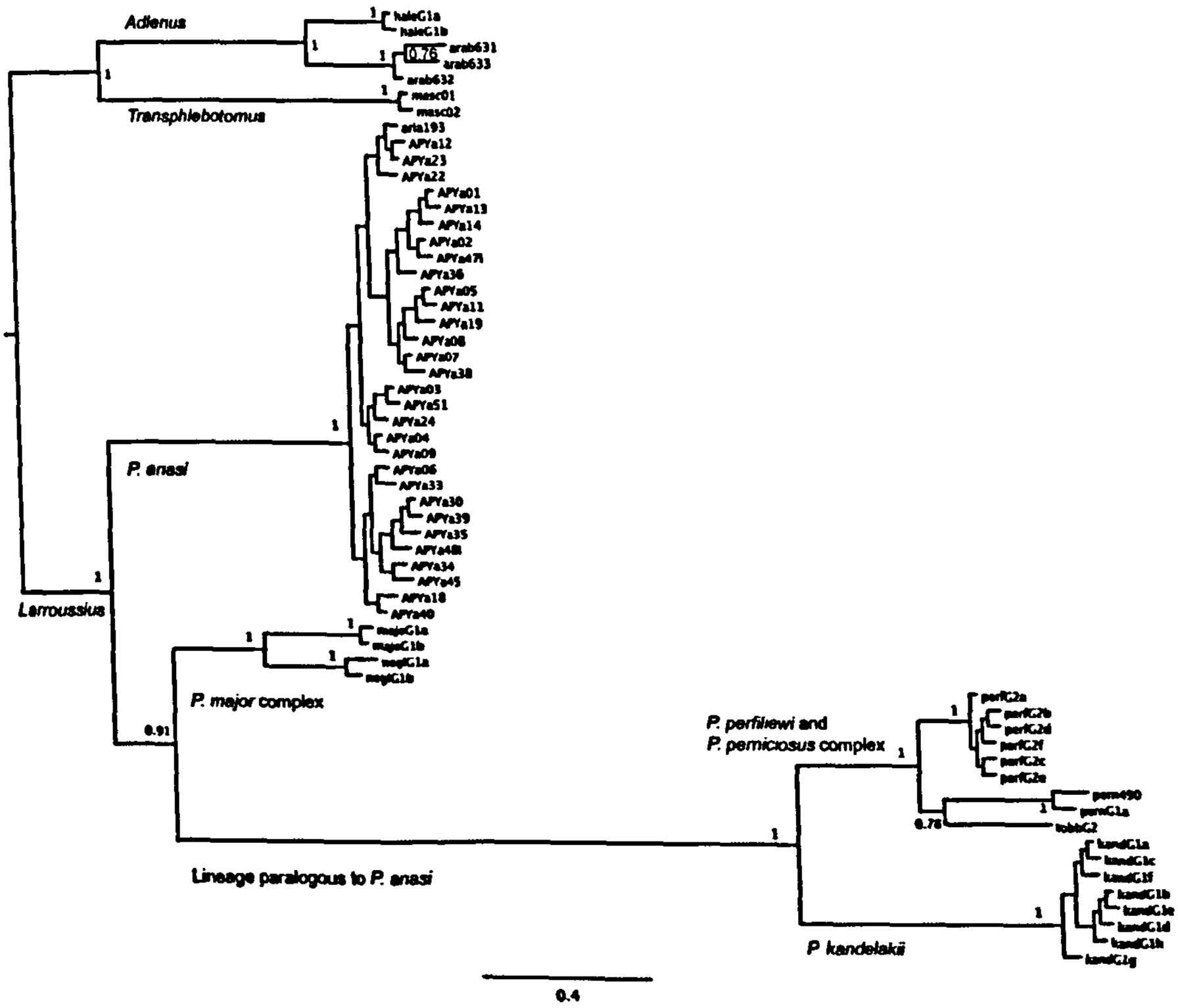


Table 3.3 Quantitative sequence variation among pairwise alleles of the apyrases of *Phlebotomus* species. Above the diagonal: the average number of nucleotide substitutions per site, K (Nei, 1987) with Jukes-Cantor correction (DNASP v490.1). Below diagonal: percentage amino acid similarity (and identity) scored using the BLOSUM62 matrix in MATGAT (v11.0). *Phlebotomus* species coded by the first four letters for the formal species name; masc: *P. mascittii*; hale: *P. halepensis;* arab: *P. arabicus* (EZ000632); majo: *P. major*; negl: *P. neglectus*; pern: *P. perniciosus*; tobb: *P. tobbi*; kand: *P. kandelakii*; perf: *P. perfiliewi*. *P. ariasi* = APYa02. The comparisons between putative orthologues are given in bold type, and between the duplicate lineages (pern490 and pern491) are italicized.

	masc01	haleG1a	arab632	majoG1b	neglG1b	APYa02	perfG1a	pernG2a	tobbG1e	kandG2b	perfG2f	pernG1a	tobbG2	kandG1f
masc01	-	**0.156**	**0.161**	**0.177**	**0.164**	**0.2**	**0.258**	**0.274**	**0.271**	**0.234**	0.325	0.293	0.319	0.363
haleG1a	**98.7 (88.3)**	-	**0.049**	**0.167**	**0.175**	**0.183**	**0.3**	**0.032**	**0.293**	**0.268**	0.339	0.319	0.332	0.37
arab632	**98.7 (87.7)**	**100 (95.5)**	-	**0.161**	**0.18**	**0.18**	**0.264**	**0.277**	**0.286**	**0.243**	0.329	0.309	0.322	0.349
majoG1b	**96.8 (89)**	**96.1 (87.7)**	**96.8 (89)**	-	**0.054**	**0.135**	**0.214**	**0.217**	**0.217**	**0.186**	0.271	0.271	0.261	0.299
neglG1b	**95.5 (87)**	**95.5 (87.7)**	**95.5 (87)**	**96.8 (93.5)**	-	**0.135**	**0.208**	**0.202**	**0.208**	**0.18**	0.261	0.255	0.252	0.289
APYa02	**96.8 (87.7)**	**96.8 (87.7)**	**96.1 (89)**	**95.5 (89)**	**95.5 (89.6)**	-	**0.24**	**0.231**	**0.234**	**0.194**	0.299	0.3	0.296	0.316
perfG1a	**89.6 (77.3)**	**87.7 (76)**	**87.7 (79.2)**	**89.6 (78.6)**	**89.6 (76.6)**	**90.3 (78.6)**	-	**0.063**	**0.061**	**0.114**	*0.268*	*0.252*	*0.234*	*0.237*
pernG2a	**89 (77.6)**	**88.3 (76)**	**88.3 (77.9)**	**89.6 (78.6)**	**90.3 (77.3)**	**90.9 (78.9)**	**94.8 (93.5)**	-	**0.085**	**0.13**	*0.274*	*0.252*	*0.234*	*0.252*
tobbG1e	**88.3 (77.3)**	**86.4 (74.7)**	**86.4 (76.6)**	**88.3 (77.3)**	**88.3 (75.3)**	**89.6 (77.3)**	**98.1 (96.1)**	**94.2 (92.9)**	-	**0.124**	*0.271*	*0.25*	*0.255*	*0.237*
kandG2b	**90.9 (79.9)**	**89.6 (79.2)**	**89.6 (81.2)**	**90.3 (81.8)**	**90.9 (81.2)**	**91.6 (83.8)**	**93.5 (88.3)**	**90.9 (85.7)**	**92.9 (87.7)**	-	*0.211*	*0.217*	*0.191*	*0.189*
perfG2f	83.1 (64.9)	82.5 (64.9)	82.5 (65.6)	81.8 (65.6)	81.8 (64.3)	83.1 (64.9)	*85.1 (64.9)*	*81.8 (64.3)*	*83.1 (64.3)*	*86.4 (68.2)*	-	0.06	0.056	0.135
pernG1a	83.8 (64.9)	83.1 (65.6)	83.1 (65.6)	82.5 (65.6)	82.5 (64.3)	83.8 (64.9)	*85.7 (66.2)*	*82.5 (65.6)*	*83.8 (65.6)*	*87 (69.5)*	99.4 (96.1)	-	0.073	0.145
tobbG2	81.8 (62.3)	81.2 (62.3)	81.2 (63)	80.5 (63)	80.5 (61.7)	81.8 (62.3)	*84.4 (63.6)*	*81.2 (63)*	*82.5 (63)*	*85.7 (66.9)*	97.4 (96.8)	98.1 (94.2)	-	0.143
kandG1f	81.2 (65.6)	80.5 (64.9)	80.5 (66.2)	81.2 (66.2)	81.2 (64.9)	82.5 (66.2)	*84.4 (66.2)*	*80.5 (65.6)*	*83.8 (66.2)*	*87 (70.8)*	92.9 (86.4)	93.5 (85.1)	91.6 (83.8)	-

function of all new sandfly apyrase sequences. The top 20 hits (E value < $9e^{-07}$) matched apyrases [calcium activated nucleotidases, CANs] from *Phlebotomus*, Hemiptera, Coleoptera, humans and mouse amongst others, indicating the conserved function of all alleles on the duplicate lineages and of other *Phlebotomus* species.

Figure 3.4 shows the position of variable amino acids and functional sites, extrapolated from the human CAN, of 44 unique *Phlebotomus* amino acid alleles. Of the 18 single codon functional sites four were conserved across *Phlebotomus*, codons 51, 59, 120 and 138. All but one of the remaining sites varied in one or both duplicate lineages, and none varied in *P. ariasi*. The effect of phylogenetically associated amino acid replacements at these functional sites on secondary protein structure was investigated using Chou-Fasman + Robson-Garnier prediction methods. Any structural changes usually involved only the loss of a single beta-sheet with or without an associated loss of a turn. This was also true for the conservative residue replacements at the two polymorphic sites with no known apyrase function (codons 32, 152) in *P. ariasi*.

Heterogeneity in selection pressure and its direction was tested in a maximum likelihood framework (PAML), where positive selection is assumed when $\omega > 1$ and the LRT comparing two test models is significant ($P < 0.05$). After re-estimation of branch length based on number of nucleotide substitutions per codon, the first input topology (pruned data set of Figure 3.3a) favoured a no clock model (unrooted phylogeny), over a global clock (rooted phylogeny): significant LRT (2*(-3073.90)-(-3273.70); df = 42; $P < 0.001$). Apyrase was concluded to be predominantly under purifying selection, not positive selection, with $\omega < 1$: null model I (ω of branch D set to = $\omega C = \omega B = \omega A$) Positive selection was detected along branch A immediately after the duplication event (of the paralogous lineage): model I (null) $\omega = 0.218$ the average over the phylogeny, versus model II single varying branch A $\omega = 999$ (infinity, nonsynonymous substitutions only), with a significant LRT of $P < 0.01$ (Table 3.4). It followed that a significant LRT ($P < 0.01$) directly supported positive directional selection in branch A, when branch A was fixed to $\omega = 1$ (model III) versus model II. However, no evidence was found to support positive selection across the paralogous lineage (branches A/B Figure 3.3a, model IV Table 3.4), as $\omega < 1$ indicated purifying selection. A highly significant LRT was obtained for this model against the null model I, supporting heterogeneity in purifying selection pressures reflecting a two-fold difference in the

Figure 3.4 Alignment of the 154-amino acid apyrase fragment starting on nucleotide 166 in GenBank accession AY845193 (*P. ariasi*) of (a) *P. ariasi* allele APYa02 complete sequence, and (b) polymorphic residues for alleles of each *Phlebotomus* species. Amino acid changes at functional sites are highlighted including: binding (+) and calcium binding (*) in the human homologue (Dai *et al.*, 2004); after *in vitro* mutagenesis of the human homologue, o are essential to APDase activity, ▲ single residue mutation from Glu to Tyr with high associated ADPase nucleotidase activity; and ^ point mutations that convert the wild-type human CAN into 100-fold more potent ADPase which abolishes platelet aggregation (Yang and Kirely, 2004). [] Brackets enclose putative MHC epitopes in the sandfly *P. dubosqci* (Kato *et al.*, 2006).

(a)

```
                                                                                               11111111
       11111112222222222333333333344444444445555555555666666666677777777778888888888999999999900000000
       34567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567
       .....[..........].....................▲..o..+.**.[.......].o+...[.......].[.........].........+
APYa02 SDQKSFTSIVRYGELKDNGERYTLSIKSENLHYFTRYAYNGRGAELSELLYFNNKLYTIDDKTGIIFEVKHGGDLIPWVILSNGDGNQKNGFKAE
                                                                                                     *

       1111111111111111111111111111111111111111111111111
       0011111111112222222222333333333344444444445555555555666666
       8901234567890123456789012345678901234567890123456
       ............++.+.++..............................
APYa02 WATVKGDKLIVGSTGIPWFEEKTQSLNTYSLWVKEISKEGEVTNINWKSQYSKVKNAMG
                   ^^ ^     ^        ^
```

Figure 3.4 (b) see caption on previous page:

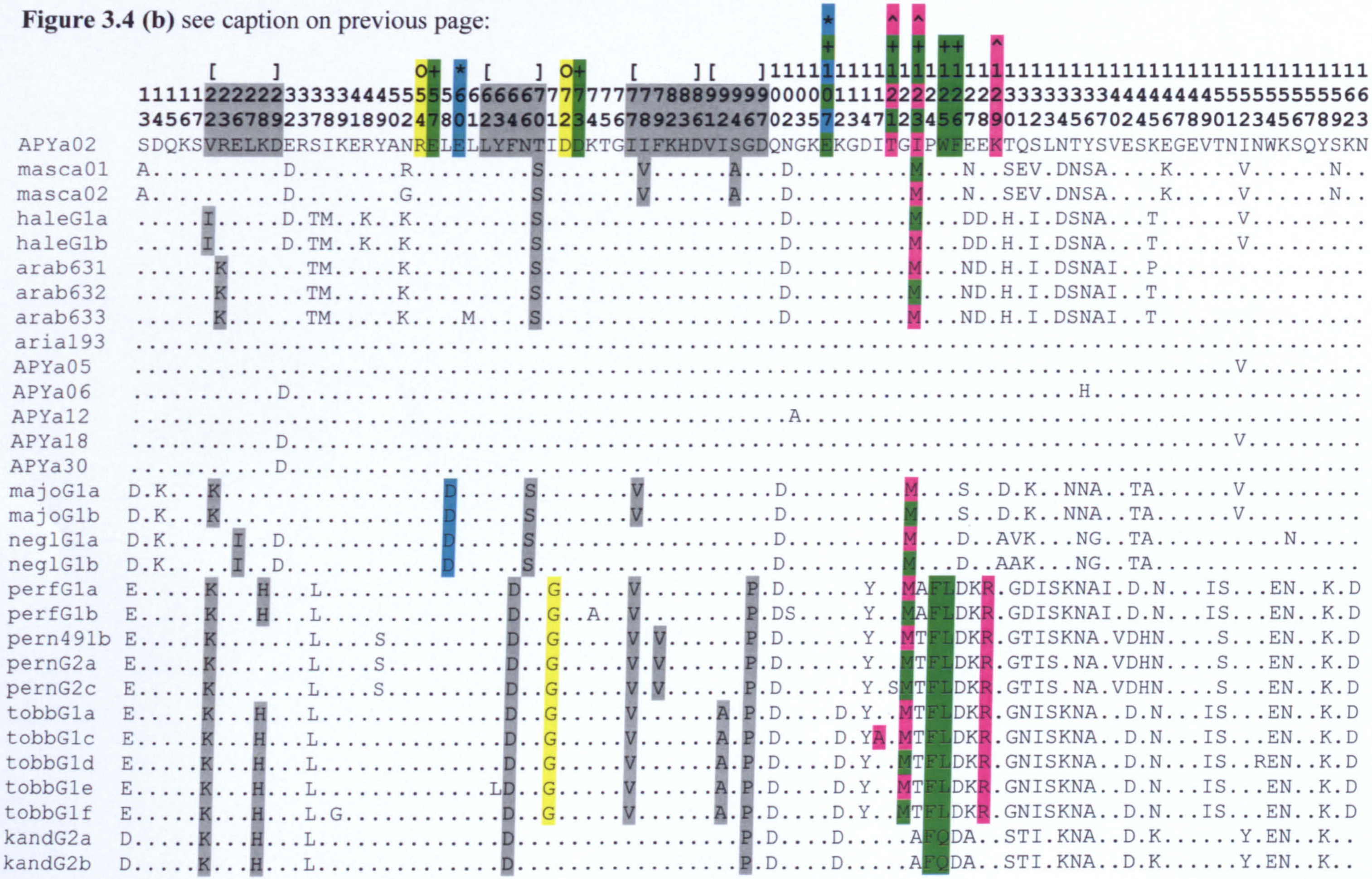
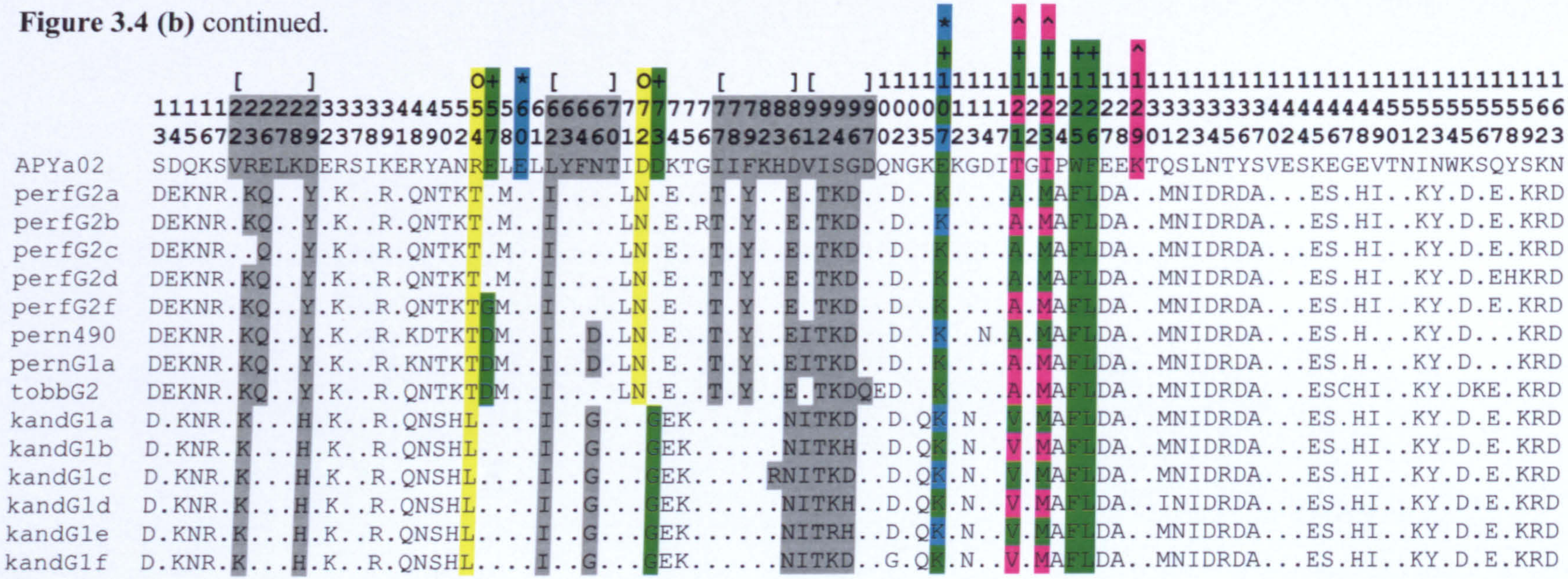


Figure 3.4 (b) continued.



nonsynonymous/synonymous substitution rate change (ω) on branches A/B (ω = 0.375) (less selectively constrained, presumably due to the paralogous nature of duplicate branch A, see discussion) compared to background branches C/D (ω = 0.172). Similar results were obtained when testing for positive selection along both duplicate lineages, using a two ratio model (IV; ωD = ωC ≠ ωB = ωA) versus a three ratio model (V; ωD ≠ ωC ≠ ωB = ωA). Purifying selection was concluded for all branches (ω < 1), and a significant LRT ($P <$ 0.05) indicated heterogeneity in the level of purifying selection pressure between the two duplicate lineages, which were less conserved than the pro-orthologue branches (pro-orthologues ωD = 0.143 < orthologous duplicate pern491 ωC = 0.255 < paralogous duplicate pern490 ωB = ωA = 0.375).

PAML branch models were also used to test for evidence of positive selection along the *P. ariasi* lineage using the phylogeny of putative apyrase orthologue sequences (Figure 3.3b). Again an input no clock model (unrooted phylogeny) was significantly favoured ($P <$ 0.001) over a global clock for this re-estimated topology (LRT: 2*(-2465.29)-(-2311.11); df = 28). A non-significant LRT showed no heterogeneity in selection pressures in the *P. ariasi* branch (F) compared with the background branches (E) ($P > 0.05$) and, as it was accompanied by ω < 1 on all branches in both models, selection was concluded to be purifying not positive (models VI and VII Table 3.4). A further branch test was implemented to detect selection on *P. perniciosus* branch (G), a sympatric vector to *P. ariasi*. Again no significant heterogeneity in selection was supported (LRT $P > 0.05$) and ω < 1 indicated purifying selection (model VIII).

Random-site and Fixed-site models were implemented to assess heterogeneity in selection pressures and positive selection among sites, which may have been masked by averaging codons over branches. For the phylogeny in Figure 3.3a and using model selection M2a versus nearly neutral M1a [see Materials and methods], no sites were found to be under positive selection; proportion of sites $p_2 = 0.00$ at freely estimated $\omega_2 = 11.224$ (Table 3.5). The beta neutral M7 null model showed apyrase to have an exponential beta distribution, most of the sites having ω closer to zero, under purifying selection. The LRT statistic between the models M7 and beta selection M8 whose discrete ω class was free to be estimated > 1, was significant ($0.01 < P < 0.05$; df = 2), suggesting ω to be variable among random-sites. Following, model M8 indicated 3% of sites to be under positive

Table 3.4 PAML parameter estimates and likelihood ratio test statistics, for detecting selection on branches (uppercase letters as given in Figures 3.3a and b) of *Phlebotomus* apyrase phylogenies. * Significant heterogeneity in selection pressure between models.

Model	Parameter estimates	lnL
I. 1 ratio	$\omega D = \omega C = \omega B = \omega A = 0.218$	-3073.90
II. 2 ratio	$\omega D = \omega C = \omega B = 0.182, \neq \omega A = 999$	-3051.104
III. 2 ratio	$\omega D = \omega C = \omega B \neq \omega A = 1$	-3056.45
IV. 2 ratio	$\omega D = \omega C = 0.172, \neq \omega B = \omega A = 0.375$	-3067.17
V. 3 ratio	$\omega D = 0.143, \neq \omega C = 0.255, \neq \omega B = \omega A = 0.375$	-3064.55
VI. 1 ratio	$\omega E = \omega F = \omega G = 0.1640$	-2311.11
VII. 2 ratio	$\omega E = \omega G = 0.170, \neq \omega F = 0.127$	-2310.81
VIII. 2 ratio	$\omega E = \omega F\ 0.157, \neq \omega G = 0.286$	-2309.93

Model compared	2ΔlnL	df	χ^2 *P*
I. and II.	2*((-3051.10)-(-3073.90))	1	< 0.001*
I. and III.	2*((-3051.10)-(-3056.45))	1	< 0.01*
I. and IV.	2*((-3067.17)-(-3073.90))	1	< 0.001*
IV. and V.	2*((-3064.55)-(3067.17))	1	< 0.05*
VI. and VII.	2*((-2310.81)-(-2311.11))	1	> 0.05
VI. and VIII.	2*((-2310.81)-(-2309.93))	1	> 0.05

Table 3.5 PAML parameter estimates and likelihood ratio test statistics, for detecting selection of *Phlebotomus* apyrase under the Random-sites models.

Model compared	Parameter estimates	lnL	d_N/d_S	PSS
Pro-orthologues and both post-duplicate lineages (Figure 3.3a)				
M1a	$p_0 = 0.860, p_1 = 0.140$: $\omega_0 = 0.147, \omega_1 = 1$	-3043.76	0.267	N/A
M2a	$p_0 = 0.860, p_1 = 0.140, p_2 = 0.000$ $\omega_0 = 0.147, \omega_1 = 1, \omega_2 = 11.225$	-3043.76	0.267	131
M7	$p = 0.651$, $q = 2.113$	-3042.06	0.233	N/A
M8	$p_0 = 0.970, p_1 = 0.030,$ $p = 0.908, q = 3.555, \omega = 1.452$	-3039.02	0.239	131,132,145
Orthologues (Figure 3.3b)				
M1a	$p_0 = 0.845, p_1 = 0.155$: $\omega_0 = 0.08, \omega_1 = 1$	-2272.74	0.222	N/A
M2a	$p_0 = 0.844, p_1 = 0.150, p_2 = 0.006$ $\omega_0 = 0.08, \omega_1 = 1, \omega_2 = 2.625$	-2273.01	0.234	131, 132
M7	$p = 0.260, q = 1.155$	-2266.57	0.181	N/A
M8	$p_0 = 0.976, p_1 = 0.021,$ $p = 0.339, q = 1.852, \omega = 2.087$	-2264.55	0.193	131, 132, 145

PSS = Positive Selected Sites, where $Pr\ (\omega > 1) < 0.95$, so not significant. (Yang *et al.*, 2005)

selection with ω = 1.452. Similar results were obtained on the orthologous apyrase phylogeny (Figure 3.3b): LRT non-significant ($P > 0.05$) comparing models M1a and M2a or M7 and M8, with less than 1% and 2.1 % of sites with ω > 1 (positive selection), respectively. For both phylogenies, although M8 supported some sites under positive selection, BEB did not identify any specific site [Pr (ω > 1) < 0.95]. One interpretation of this result is that BEB is confident that positive selection among sites exists, but cannot identify their position. The three sites that were inferred to be under positive selection, but not statistically supported, included codons 131, 132, and 145, these were not functionally important.

For both phylogenies being tested, the results for Fixed-site models indicated no codon usage differences or heterogeneity in selection pressure between buried versus exposed site classes (binding sites, ADPase sites and epitopes). Model E partitioned various exposed sites independently from buried site classes, where it did not give a significantly different log likelihood value from the null homogeneous model A ($P > 0.05$). Furthermore, positive selection was not supported for any site class because ω < 1 (purifying selection). It was observed that the binding site class (ω_2) had a near two-fold increase in the nonsynonymous/synonymous substitution rate change and a lower transition/transversion rate (k_2) relative to buried sites (ω_1 and k_1, respectively) that was attributed to the presence of the paralogous lineage (Table 3.6).

Table 3.6 PAML parameter estimates and likelihood ratio test statistics, for detecting selection of *Phlebotomus* apyrase under the Fixed-sites models.

Model compared	lnL	ω	k	r_2
Pro-orthologues and both post-duplicate lineages (Figure 3.3a)				
MA – null model	-3073.90	0.218	k = 2.036	1
ME - binding sites	-3079.32	ω_1 = 0.211, ω_2 = 0.3843	k_1 = 1.983, k_2 = 0.211	0.885
ME - activity sites	-3067.00	ω_1 = 0.216, ω_2 = 0.112	k_1 = 2.155, k_2 = 0.926	1.908
ME - epitope	-3082.22	ω_1 = 0.230, ω_2 = 0.135	k_1 = 1.969, k_2 = 0.974	0.765
Orthologues (Figure 3.3b)				
MA – null model	-2311.11	0.164	k = 1.837	1
ME - binding sites	-2314.60	ω_1 = 0.165, ω_2 = 0.152	k_1 = 1.836, k_2 = 1.923	0.909
ME - activity sites	-2301.24	ω_1 = 0.162, ω_2 = 0.096	k_1 = 1.923, k_2 = 1.097	2.137
ME - epitope	-2309.00	ω_1 = 0.180, ω_2 = 0.109	k_1 = 1.787, k_2 = 1.746	0.660

Legend k = transition/transversion rate; r_2 = the substitution rate of the second site partition relative to the rate of the first partition (r_1 = 1).

3.3.3 Scoring apyrase genotypes of individual *P. ariasi*

459 (out of 471) *P. ariasi* were successfully amplified and directly sequenced using conserved primers APY-1F with APY-3R (520 bp) from 20 spatio-temporal natural populations. As no prior knowledge of the allelic variation in apyrase of *P. ariasi* was known, extensive PCR optimization experiments using the PASA method were implemented to accurately score genotypes. This study found, concurring with the literature (Kwok *et al.*, 1990) that T 3' terminating oligonucleotides were the only allele specific primers to misprime. However, misprimed alleles were of lower amplitude than the target alleles in sequence chromatograms, so correct genotypes were confidently scored. For *P. ariasi*, 47 nucleotide alleles (Table 3.7) and 86 nucleotide genotypes (Table 3.8) were recorded, which gave 15 deduced amino acid alleles (Table 3.9). In total, 4 novel alleles [16 flies from north Africa, NE Spain and France] had to be inferred using the algorithm described in Chapter 2, section 2.2.4. They showed no more than two mutational steps from a modal allele (allele with more than one derived allele or a frequency over 10) in a TCS network.

3.3.4 *P. ariasi* apyrase genealogy and recombination

35 base positions segregated in *P. ariasi* apyrase alleles, of which 22 and 13 substitutions were synonymous and nonsynonymous, respectively, and each had a single open reading frame. Nineteen segregating sites were singletons and four base positions (88, 100, 430, 451) had a transition rate between 25-45% among alleles (Figure 3.5). The network (Figure 3.6) showed reticulate loops (alternative most parsimonious pathways) caused either by recurrent mutations or recombination between the predominating *P. ariasi* apyrase alleles in Europe (modes APYa01, 02 and 03). These loops were mainly attributed to mutations at nucleotides 88, 100, 430, 451 in addition to bases 306 and 433. The four gamete test (Rm) identified no recombination events in nine populations, one to two events in 11 populations and four recombination events occurring in the history of the samples between nucleotides 100-139, 139-352, 433-451, 451-475. Population recombination parameter R between adjacent sites ranged between 0-0.0301, and nucleotide diversity (π) between 0.00078-0.00663, and in populations where $R > 0$, showed positive significant correlation ($r_s = 0.69$, P two-tailed < 0.01, df = 12). For comparison, no such correlation was found between R and π at nuclear EF-1α, ($r_s = 0.50$, P two-tailed 0.08, df = 12).

Table 3.7 Apyrase nucleotide allele frequencies characterized from 20 natural populations of *P. ariasi.* AA = associated amino acid allele. Dark grey and lighter grey hatched shading highlight the two most frequent amino acids in Europe (AA02) and north Africa (AA01), respectively. N = sample size; i = inferred alleles.

																Massif Central and Rhone valley, France					
Population region		**Morocco**	**Portugal**	**NW Spain**	**C. Pyrenees, France**		**Eastern Pyrenees, France**								**NE Spain**	**MC(s)**		**MC**	**Rhone**	**Lot, France**	
Allele / POP	**AA**	**AGH**	**CHR**	**CSP**	**HP1**	**HP2**	**PAS**	**PLB**	**IRL07**	**TUL**	**ARQ06**	**ARQ08**	**MLQ**	**CAT**	**TRJ**	**CTU**	**SPV**	**SAM13**	**DRAz4**	**LNP**	**RME**
APYa02	02		0.043	0.217	0.278	0.167	0.391	0.396	0.114	0.229	0.136	0.217	0.229	0.25	0.239	0.229	0.37	0.438	0.273		
APYa03	02			0.065	0.315	0.389	0.239	0.313	0.25	0.208	0.159	0.283	0.313	0.219	0.413	0.25	0.152	0.083	0.045	0.071	0.038
APYa01	02				0.185	0.417	0.163	0.208	0.523	0.521	0.614	0.391	0.313	0.469	0.065	0.458	0.435	0.458	0.591	0.929	0.962
APYa08	02		0.087	0.522			0.033	0.021		0.021			0.063		0.022		0.022		0.023		
APYa05	07		0.152	0.022	0.019	0.028	0.011						0.021		0.022						
APYa04	02		0.065											0.031							
APYa15	02			0.022	0.056		0.054	0.042	0.045		0.045	0.022		0.031	0.109		0.022				
APYa23	02			0.022			0.011					0.022	0.021			0.021					
APYa20	04			0.065	0.019		0.033		0.023		0.023				0.043						
APYa12	04				0.019											0.021					
APYa18	06				0.037										0.022						
APYa17	02				0.019										0.022						
APYa21	02						0.043			0.021					0.022	0.021					
APYa19	07						0.022		0.023						0.022						
APYa29 i	02								0.023		0.023										
APYa28	02											0.022									
APYa07	02	0.176	0.087																		
APYa33	01	0.029																			
APYa30	01	0.382																			
APYa31	02	0.059																			
APYa48 i	01	0.029																			
APYa34	01	0.029																			
APYa35	01	0.029																			
APYa36	02	0.029																			
APYa37	01	0.029																			
APYa38	02	0.029																			
APYa39	03	0.029																			
APYa40	01	0.029																			
APYa25	01	0.029																			
APYa32	01	0.059																			
APYa26	01	0.029																			
APYa06	08		0.304																		
APYa10	08		0.13																		
APYa09	09		0.087																		
APYa11	10		0.022																		
APYa45	11		0.022																		
APYa22	02			0.022																	
APYa24	12			0.043																	
APYa46 i	02				0.019																
APYa16	02				0.037																
APYa51	18							0.021													
APYa27	02											0.022									
APYa47 i	05											0.022									
APYa44	02												0.042								
APYa49	02																	0.021			
APYa13	13																		0.045		
APYa14	14																		0.023		
N		**17**	**23**	**23**	**27**	**18**	**46**	**24**	**22**	**24**	**22**	**23**	**24**	**16**	**23**	**24**	**23**	**24**	**22**	**21**	**13**

Table 3.8 Predominant apyrase nucleotide genotype frequencies characterized from 20 populations of *P. ariasi.*

																Massif Central and Rhone valley, France					
Population region	Morocco	Portugal	NW Spain	C. Pyrenees, France		Eastern Pyrenees, France								NE Spain	MC(s)		MC	Rhone	Lot, France	
Genotype / POP	AGH	CHR	CSP	HP1	HP2	PAS	PLB	IRL07	TUL	ARQ06	ARQ08	MLQ	CAT	TRJ	CTU	SPV	SAM13	DRAz4	LNP	RME
N	17	23	23	27	18	46	24	22	24	22	23	24	16	23	24	23	24	22	21	13
0203			0.0435	0.1481	0.2222	0.1957	0.2083	0.0455	0.1250	0.0455	0.1304	0.0833	0.0625	0.1304	0.0417	0.1304		0.0417		
0202			0.0435	0.0741		0.1522	0.2083			0.0455	0.0435	0.0417	0.1250	0.0435	0.0833	0.1739	0.1364	0.2500		
0208			0.2609			0.0435						0.0417		0.0435		0.0435				
0215			0.0435	0.0370		0.0435								0.0870		0.0435				
0308			0.0870			0.0217	0.0417		0.0417			0.0833								
0101				0.0741	0.1667	0.0435	0.0417	0.1818	0.2083	0.4545	0.1304	0.0833	0.1250		0.1667	0.3043	0.3636	0.2083	0.8571	0.9231
0103				0.1111	0.4444	0.1304	0.1667	0.3182	0.2500	0.0909	0.2609	0.2500	0.3750	0.0870	0.2917	0.0870	0.0909	0.1250	0.1429	0.0769
0102				0.0741	0.0556	0.0652	0.1667	0.1818	0.3333	0.1364	0.1739	0.2083	0.1875	0.0435	0.2500	0.1739	0.2273	0.3333		
0303				0.1111	0.0556	0.0435	0.0417	0.0455		0.0455	0.0870	0.0833		0.1739	0.0417	0.0435				
0315				0.0370			0.0833			0.0909				0.0870						
0120				0.0370				0.0455		0.0455										
0218				0.0370										0.0435						
0205					0.0556							0.0417								
0321						0.0217								0.0435	0.0417					
0320						0.0217								0.0435						
0220						0.0435								0.0435						
0115						0.0435		0.0909			0.0435		0.0625							
0508		0.0435	0.0435																	
0730	0.1765																			
3032	0.1176																			
0506		0.1739																		
0820			0.1304																	
0808			0.1739																	
Low frequency and private	0.7059	0.7826	0.1739	0.2593	0	0.1304	0.0417	0.0909	0.0417	0.0455	0.1304	0.0833	0.0625	0.1304	0.0833	0	0.1818	0.0417	0	0
No. low frequency and private	12	16	3	7	-	4	1	2	1	1	3	2	1	3	2	-	4	1	-	-

N = sample size. Low frequency and private, and No. low frequency and private, denote the sum of genotype frequencies where constituent genotypes were < 0.1 in any population, and the number of genotypes that constitute this value, respectively.

Table 3.9 Geographical variation in the frequencies of the amino acid (AA) alleles of the apyrase of *P. ariasi*, showing the near fixation of allele 02 in France and northern Spain (dark grey shading), and polymorphism involving different alleles in Morocco and Portugal (lighter, hatched shading). MC(s) = southern Massif Central.

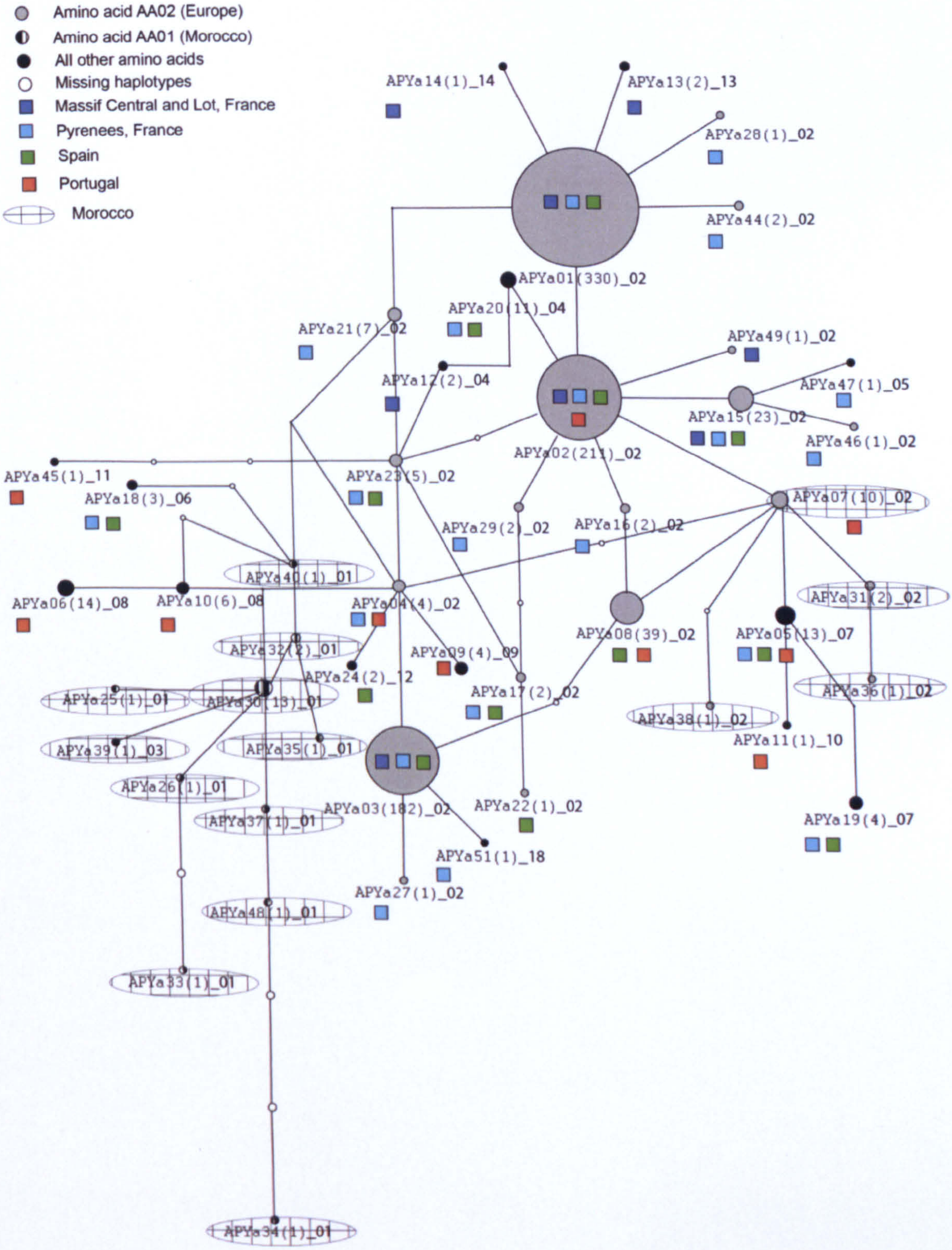
															Massif Central (MC), Rhone and Lot valleys, France					
Population region	Morocco	Portugal	NW Spain	C Pyrenees, France		Eastern Pyrenees, France								NE Spain	MC(s)		MC	Rhone	Lot, France	
AA allele/ POP	AGH	CHR	CSP	HP1	HP2	PAS	PLB	IRL07	TUL	ARQ06	ARQ08	MLQ	CAT	TRJ	CTU	SPV	SAM13	DRAz4	LNP	RME
02	0.294	0.283	0.870	0.907	0.972	0.935	0.979	0.955	1	0.977	0.978	0.979	1	0.891	0.979	1	1	0.932	1	1
07		0.152	0.022	0.019	0.028	0.033		0.023				0.021		0.043						
04			0.065	0.037		0.033		0.023		0.023				0.043	0.021					
06				0.037										0.022						
01	0.676																			
03	0.029																			
08		0.435																		
09		0.087																		
10		0.022																		
11		0.022																		
12			0.043																	
18							0.021													
05											0.022									
13																		0.045		
14																		0.023		

Figure 3.5 35 variable nucleotide positions in the 520 bp fragment of apyrase [starting on nucleotide of GenBank starting nucleotide 110 in GenBank accession AY845193], observed in 47 unique alleles characterized from 20 populations of *P. ariasi*.

```
          1111222223333333444444444455
       2588903890134800033456003355677801
       5508409665751716746721340301 2345151
APYa01 GCACGCGACCCTGGGGCGGACACTCGACAGGCGCG
APYa13 .......T.............................
APYa44 ...............G.....................
APYa28 ....................C................
APYa14 ............................T........
APYa19 ....................A....TG.........
APYa36 ......C.....................T........
APYa02 ........................T............
APYa29 .........T..............T............
APYa49i ..........T.............T............
APYa20 ...............C........T............
APYa47i ..C.....................T........A.
APYa15 .........................T........A.
APYa46i ....................T...T........A.
APYa16 ........................TA...........
APYa08 ........................TA.T.........
APYa07 ........................T..T.........
APYa31 ......C.................T..T.........
APYa11 ................T......T..TG........
APYa05 ........................T..TG........
APYa38 ........................T..T...TC..
APYa21 ...T.T..............................
APYa23 ...T.T.....................T..........
APYa22 ...T.T...T.................T....A.....
APYa17 ...T.T...T.................T..........
APYa12 ...T.T.........C........T..........
APYa45 ...TTT.............G....T.......T...
APYa18 ...TTT.................A....TG........
APYa32 ...TTTC.....................T........
APYa35 ...TTTC.......A............T........
APYa40 ...TTT.....................T........
APYa03 ...T.T.....................TA.T........
APYa27 .T.T.T.....................TA.T........
APYa51 ...T.T............A.....TA.T........
APYa06 A..TTT.................CT..T........
APYa30 ...TTTC................T..T........
APYa48i ...TTTC................T..T...TA..
APYa37 ...TTTC................T..T....A..
APYa25 ...TTTC.....A..........T..T........
APYa39 ...TTTC......A.........T..T........
APYa26 ...TTTC............T...T..T........
APYa24 ...T.T..A..............T..T........
APYa33 ...TTT.....C........T...T..T........
APYa10 ...TTT................CT..T........
APYa34 ...TTT................T..T...TC.T
APYa09 ...T.T................T..T..A....
APYa04 ...T.T................T..T........
```

Legend Nucleotide one begins on the 3rd base position of codon 28 in Figure 3.1. Bold typeface highlights the three most common (modal) alleles in *P. ariasi*; grey highlights the four common varying base positions; i, denotes an inferred allele. 19 singleton mutations are indicated by a lighter font.

Figure 3.6 Parsimony network (TCS v1.21) showing the genealogical relationships between the 47 apyrase alleles (APYaNN) from 459 *P. ariasi*, with a 9 step 95% connection limit. These alleles are shown as filled circles with sizes proportional to their frequency of occurrence. Open circles denote missing alleles. Figures in parentheses = number of flies, followed by numbers in bold = associated amino acid allele. Nucleotide allele geographical distributions are coded as given in the key.



3.3.5 Types of selection within the *P. ariasi* apyrase

Maintenance of diversity of nucleotide alleles was observed in North Africa, Iberia and France (four allele frequencies > 0.05) (Table 3.7). A similar pattern was seen in amino acid allele frequency in populations from Morocco (two allele frequencies > 0.15) and Portugal (three allele frequencies > 0.15), in contrast to those populations from northern Spain and southern France where allele AA02 predominated (frequency > 0.87) (Table 3.9). No strong selection for codon usage was shown by three measures: low values of the Codon Bias Index (0.32-0.35; where 0 = unbiased, 1 = extreme bias), high values of the Effective Number of Codons (57.7-59.0, where 61 = unbiased, 20 extreme bias) and 51.4-54.0% GC at synonymous third codon positions. A global test pooling all *P. ariasi* (N = 459), and a separate test pooling flies from Spain and France (N = 419), supported selective neutrality using the MK test as measured by relative rates of divergence and polymorphism of nonsynonymous/synonymous estimates (Fisher's exact test $P = 1$) and associated NI value (0.951-1.069; where NI of 1 = neutrality): these tests were conducted using a single *P. major* and two *P. neglectus* apyrase alleles as outgroups. Furthermore, irrespective of environment, no *P. ariasi* population showed significant departure from neutral expectation using the MK test ($P > 0.05$) (Table 3.10). This test was considered valid for intra-specific populations of *P. ariasi*, as no paralogous genes were identified by the apyrase phylogeny in this ingroup and the outgroup to *P. ariasi*, *P. major*, was an appropriate choice for neutrality based tests, as d_S was unsaturated (< 0.5) (see Appendix 3.4).

After correcting probability values for familywise Type 1 errors by implementing a sequential Bonferroni procedure, no neutrality statistic based on the mutation frequency spectrum (Fu and Li's *D* and Tajima's *D*), showed a deviation from neutral expectation (when $\alpha = 0.05$) (Table 3.10). This result allowed the rejection of the alternative hypothesis of recent selection pressures acting on the apyrase of *P. ariasi*. Only population AGH from Morocco showed a demographic signal, where the Fu F_S was significant (after Bonferroni correction), indicating the occurrence of a population expansion (or selective sweep). This concurred with the result found for both cyt b and EF-1α (Table 3.11).

Furthermore, no support for current generation/recent selection at the apyrase locus was obtained. All populations showed no statistical deviation from HWE ($P > 0.15$), and there was adherence to neutral expectation in allele frequencies for the Ewans-Watterson test ($P > 0.17$; Table 3.10). LD was investigated in two geographical

Table 3.10 Tests showing the absence of selection on the nucleotide alleles of apyrase from different geographical populations of *P. ariasi*. $P < 0.05$ = significant[#], after sequential Bonferroni correction in bold. * Tests requiring a proven outgroup (*P. major*). N = sample size. S = number of segregating sites. *h* = number of alleles. Ds, Ps, Dn, Pn = the number of synonymous (s) and non-synonymous (n) substitutions per site that are fixed (D) or polymorphic (P). MK = McDonald-Kreitman test. NI = neutrality index (NI < 1 for positive selection; NI > 1 for purifying selection; NA = not applicable). EW = Ewans-Watterson test. Rm = number of recombination events as revealed by the four gamete model (Hudson and Kaplan, 1985).

Population	N	S	*h*	Ds	Ps	Dn	Pn	*MK Fisher *P*-value	*NI	*Fu and Li *D*	*Fu and Li *D* *P*-value	EW *P*-value	Tajima *D*	Tajima *D* *P*-value	Fu *Fs*	Fu *Fs* *P*-value	Rm
AGH	17	13	15	33	11	23	2	0.1133	0.261	-0.5716	0.2871	0.985	-0.420	0.381	-6.546	**0.0020[#]**	2
CHR	23	12	10	33	5	22	6	0.5069	1.8	-0.8757	0.2083	0.176	0.615	0.771	-0.126	0.5303	1
CSP	23	10	9	33	6	22	3	1	0.75	-0.1334	0.4035	0.879	-0.517	0.343	-1.567	0.2292	1
HP1	27	12	11	32	8	22	3	0.5087	0.545	-0.0177	0.4945	0.674	0.184	0.634	-1.109	0.3593	2
HP2	18	6	4	32	5	22	1	0.3914	0.291	1.1518	0.8978	0.171	2.031	0.979	3.730	0.9408	0
PAS	46	9	10	32	6	22	2	0.4675	0.485	1.2178	0.8969	0.506	0.759	0.799	-0.352	0.4865	2
PLB	24	7	6	32	5	23	1	0.3885	0.278	0.1258	0.5229	0.324	1.226	0.891	1.598	0.7990	1
IRL07	22	10	7	32	7	22	2	0.4625	0.416	-0.8555	0.2195	0.722	0.046	0.577	0.675	0.6681	1
TUL	24	5	5	32	5	23	0	0.146	NA	1.0188	0.8539	0.379	1.787	0.963	1.931	0.8336	1
ARQ06	22	8	6	32	6	23	1	0.2322	0.232	-0.5465	0.3011	0.753	-0.120	0.506	0.540	0.6435	0
ARQ08	23	9	8	32	7	23	1	0.1406	0.199	-1.1351	0.1616	0.622	0.594	0.759	0.199	0.5841	0
MLQ	24	7	7	32	6	22	1	0.238	0.242	1.2052	0.8933	0.290	1.502	0.932	1.011	0.7174	1
CAT	16	6	5	32	5	23	0	0.146	NA	1.0668	0.8433	0.350	1.133	0.882	1.570	0.7928	0
TRJ	23	11	11	32	7	22	3	0.7275	0.623	0.6089	0.6931	0.850	0.405	0.706	-1.386	0.2992	2
CTU	24	6	6	32	5	23	1	0.3885	0.278	0.1258	0.5343	0.424	1.601	0.941	1.419	0.7803	1
SPV	23	6	5	32	5	23	5	0.146	NA	1.0236	0.8259	0.325	0.561	0.751	1.308	0.7754	0
SAM13	24	6	4	32	6	23	0	0.0746	NA	0.1258	0.5318	0.278	-0.340	0.421	1.292	0.7686	0
DRAz4	22	7	6	32	5	23	2	0.6911	0.557	0.3340	0.5399	0.762	-0.974	0.181	-1.084	0.2606	0
LNP	21	5	2	32	5	23	0	0.146	NA	1.0339	0.9668	0.636	-1.045	0.165	2.563	0.8639	0
RME	13	5	2	32	5	23	0	0.146	NA	-2.0374	0.0242[#]	1.000	-2.002	0.005[#]	1.073	0.5653	0

Table 3.11 Neutrality based population genetic tests (without an outgroup) and recombination estimates for 20 natural populations of *P. ariasi* for neutral loci mitochondrial cyt b and EF-1α. N = sample size. S = number of segregating sites. *h* = number of alleles. EW = Ewans-Watterson test. # Significant deviation from neutral expectations when $P < 0.05$, after sequential Bonferroni correction in bold. Rm = number of recombination events as revealed by the four gamete model (Hudson and Kaplan, 1985).

Locus	Population	N	S	*h*	EW (*P*-value)	Tajima *D*	Tajima *D* (*P*-value)	Fu F_S	Fu F_S (*P*-value)	Rm
EF-1α	AGH	17	10	14	0.957	-0.427	0.3792	-7.269	**0.0004**#	2
	CHR	23	9	14	0.521	-0.018	0.5428	-6.086	0.0042#	2
	CSP	21	7	8	0.533	-0.461	0.3603	-2.093	0.1207	2
	HP1	25	9	8	0.964	-1.722	0.0201#	-4.128	0.0086#	0
	HP2	18	2	3	0.516	-0.057	0.4386	0.047	0.4176	0
	PAS	48	7	10	0.988	-1.260	0.0957	-6.771	**0.0015**#	2
	PLB	21	5	5	0.702	-0.597	0.3226	-0.607	0.3611	1
	IRL07	22	4	5	0.410	-0.268	0.4356	-0.791	0.2985	0
	TUL	23	4	5	0.487	-0.108	0.4897	-0.579	0.3694	1
	ARQ06	14	4	5	0.644	-0.758	0.2522	-1.534	0.1071	0
	ARQ08	23	4	5	0.376	0.157	0.6152	-0.272	0.4526	0
	MLQ	24	6	8	0.926	-0.904	0.2065	-3.638	0.0159#	1
	CAT	16	6	7	0.847	-0.793	0.2425	-2.397	0.0617	1
	TRJ	23	9	11	0.979	-1.459	0.0543	-7.098	**0.0001**#	2
	CTU	24	3	4	0.212	0.811	0.801	0.670	0.6640	1
	SPV	23	4	6	0.652	-0.086	0.5046	-1.529	0.1905	1
	SAM13	24	2	3	0.444	0.109	0.6466	0.258	0.4915	0
	DRAz4	22	3	4	0.467	-0.282	0.4161	-0.505	0.3394	0
	LNP	24	2	3	0.780	-0.873	0.1997	-1.118	0.1436	0
	RME	23	1	2	0.593	-0.311	0.292	0.162	0.2853	0
Cyt b	AGH	17	13	10	1.000	-1.877	0.0159#	-5.557	**0.0002**#	1
	CHR	24	14	8	0.980	-0.971	0.1734	-0.738	0.3726	0
	CSP	24	13	4	0.384	-0.368	0.3915	3.784	0.9461	0
	HP1	27	23	11	0.998	-0.958	0.1725	-1.087	0.3237	2
	HP2	18	13	4	1.000	-1.240	0.9229	2.332	0.8806	0
	PAS	52	18	12	0.882	1.732	0.9656	1.642	0.7621	1
	PLB	20	17	20	0.996	-2.344	**0.0006**#	-1.547	0.1688	0
	IRL07	22	14	4	0.681	1.672	0.9655	6.731	0.9889	0
	TUL	24	17	7	0.971	-0.138	0.4998	1.700	0.7892	0
	ARQ06	38	26	15	0.993	0.116	0.6096	-1.056	0.3705	3
	ARQ08	23	17	7	0.936	1.064	0.882	2.830	0.8805	0
	MLQ	35	12	8	0.997	-1.185	0.1201	-1.326	0.2578	0
	CAT	16	21	10	1.000	-1.369	0.0776	-2.432	0.0991	0
	TRJ	23	28	13	1.000	-1.856	0.0172#	-4.216	0.0250#	0
	CTU	24	15	10	0.975	-2.029	0.0065#	-4.662	**0.0030**#	0
	SPV	24	16	7	0.693	-2.037	0.0065#	-1.112	0.2585	0
	SAM13	24	4	4	0.247	0.457	0.7084	0.733	0.6898	0
	DRAz4	20	4	4	0.961	-1.638	0.0324#	-1.613	0.0441#	0
	LNP	24	1	2	1.000	-1.159	0.1512	-1.028	0.0706	0
	RME	13	0	1	-	-	-	-	-	-

populations with larger sample sizes (PAS, ARQ) and multiple cyt b haplogroups. No evidence of within population LD was shown between all locus pairs ($P > 0.05$), providing some evidence for an absence of epistatic selection or selective sweeps. The exception to this pattern was between apyrase and EF-1α in population PAS, where $0.01 < P < 0.05$.

Assuming an absence of recombination can lead to conservative inferences from tests for LD, haplotype distribution (Fu's F_S) and to a lesser extent allele frequency spectra (D statistics) (Ramirez-Soriano *et al.*, 2008). Therefore, the support for up to two recombination events at apyrase within some populations may have masked significant and generally positive D statistics (Table 3.10), namely balancing selection. This is contrary to the result for the conserved nuclear EF-1α for which populations with one or more recombination events mostly showed negative D values (Table 3.11). This would have indicated directional selection (most likely purifying) if the tests had been significant.

3.3.6 Phylogeography and population genetics at apyrase of *P. ariasi* support inferences made at other characterized loci

The parsimony network (Figure 3.6) supported no intra-specific lineages in apyrase of *P. ariasi*, consistent with patterns of other nuclear sequences characterized in this thesis (Chapter 2). The network showed no obvious signal of positive selection e.g. a single extensive star-burst structure indicating a selective sweep of a favoured apyrase allele and its near derivatives. This result, together with tests directly concluding against selection, demonstrate that demographic processes might better explain the distinctive patterns of diversity and frequency patterns in the 20 natural *P. ariasi* populations.

Apyrase showed a population genetic structure in western Europe consistent with that given by other loci of *P. ariasi.* Both nucleotide and amino acid allele distribution and frequencies differentiated the Morocco and Portugal populations from each other and the rest of Iberia and France: the two modal nucleotide alleles with the highest frequencies in Spain and France (APYa01 and APYa03) were absent in Portugal and Morocco. In the latter, 10 out of 15 alleles were derivatives of the modal allele APYa30 that, like its deduced amino acid (allele AA01), is absent in Europe, and in Portugal the predominant amino acid allele (AA08) was also private. Pyrenean France and NE Spain were distinct from NW Spain and the Massif Central France, in the following ways: absence of common nucleotide alleles APYa01 and APYa02 in NW

Spain and Lot France, respectively; moderate frequency of APYa08 in NW Spain, but low frequencies elsewhere; low frequencies of APYa03 and complementary increase of APY01 (towards fixation) in the Massif Central, Rhone and Lot valleys, with both alleles found in moderate frequencies in Pyrenean France (Tables 3.7 and 3.9). Overall gene nucleotide diversity (π) in apyrase was not significantly different (overlapping standard deviations) compared to mitochondrial cyt b, but more polymorphic although not always significantly so than that in EF-1α (Figure 3.7). Diversity at nonsynonymous sites (π_n) was lower than at synonymous sites (π_s) at all loci (Figure 3.8). The global means were not significantly different between APY and cyt b (π_s t = 0.5186 ± 0.003, df = 38, P = 0.6070; π_n t =1.5669, df = 38, P = 0.1254), but were significantly higher for APY compared to nuclear EF-1α (π_s t = 8.4841 ± 0.002, df = 38, P = 0.0001): π_n was not calculated for EF-1α as no nonsynonymous changes were observed. At both types of site and overall, there was a loss of diversity at apyrase at the leading-edge of the species range, as shown for cyt b (and EF-1α where applicable).

F_{ST} estimates of genetic differentiation were also informative. They showed 'very great' levels of differentiation between populations, ranging from -0.0169 to 0.5822. Consistent with Chapter 2 results, significant pairwise F_{ST} values were found between populations from Morocco, Portugal and NW Spain or leading-edge populations (Lot, France) and all other populations. There was no significant genetic differentiation among populations in the Pyrenees (Appendix 3.5). Hierarchical AMOVA did support the same regional clustering in France and NE Spain as found for neutral loci (Chapter 2), and like other nuclear loci (but not mitochondrial) within-regions variation was also statistically significant (Table 3.12).

Globally, there was a significant positive correlation between apyrase genetic and geographical distance supporting a model of isolation-by-distance by a Mantel Test, in one or two dimensions, fitting $F_{ST}/(1-F_{ST})$ to distance (a = 0.1247, b = 0.000242; P < 0.001) or to ln distance (a = -0.3027, b = 0.0981; P < 0.001), respectively. In this single regression model the sample correlation was weak (R^2 = 0.1714), and as observed in other nuclear loci (Chapter 2), relatively high variance and statistical outliers were restricted to pairwise comparisons with two leading-edge (putatively bottle-necked) populations from Lot France (triangles Figure 3.9). For further investigation these two populations were excluded, and the remaining populations continued to follow an IBD model but with improved correlation: 56.5% of genetic distance was significantly (P < 0.001) correlated with geographical distance. All pairwise comparisons between

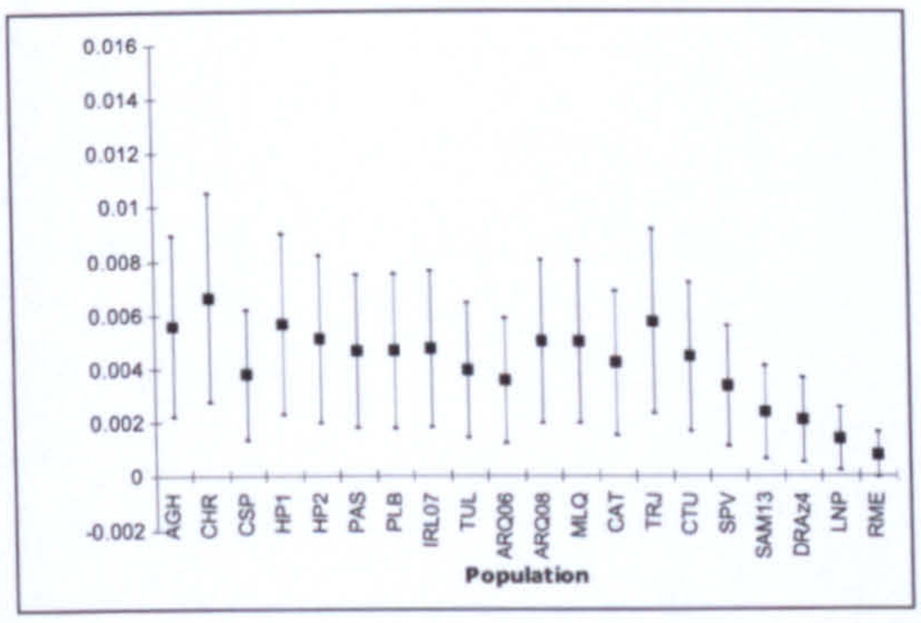
populations north of the Pyrenees (France) and those between south and north of the Pyrenees supported IBD (R^2= 0.126, Mantel test $P = 0.024$ and $R^2 = 0.234$, $P = 0.001$, respectively). Yet no IBD was supported between the outgroup populations ($R^2 = 0.111$, $P = 0.273$): a result that might be supported with higher resolution sampling. Marginal tests in a distance-based redundancy analysis supported a significant relationship between genetic distance ($F_{ST}/(1-F_{ST})$ for both geographical distance (57% variation explained, $P = 0.001$) or geographical region (59%, $P = 0.001$); in the latter data points were categorised as within south/north of the Pyrenees or across the Pyrenees. Furthermore, a conditional test taking into account geographical distance as a covariate in a multiple regression analysis significantly correlated this pairwise categorisation to genetic distance (15%, $P = 0.001$), a result that suggests the Pyrenees is or was recently a barrier to gene flow.

Table 3.12 Hierarchical AMOVA statistics for the apyrase of *P. ariasi*, to demonstrate that the regional clustering of its populations is concordant with neutral locus cyt b. * Significant *P*-values for 16,000 permutations (implemented in ARLEQUIN v3.11).

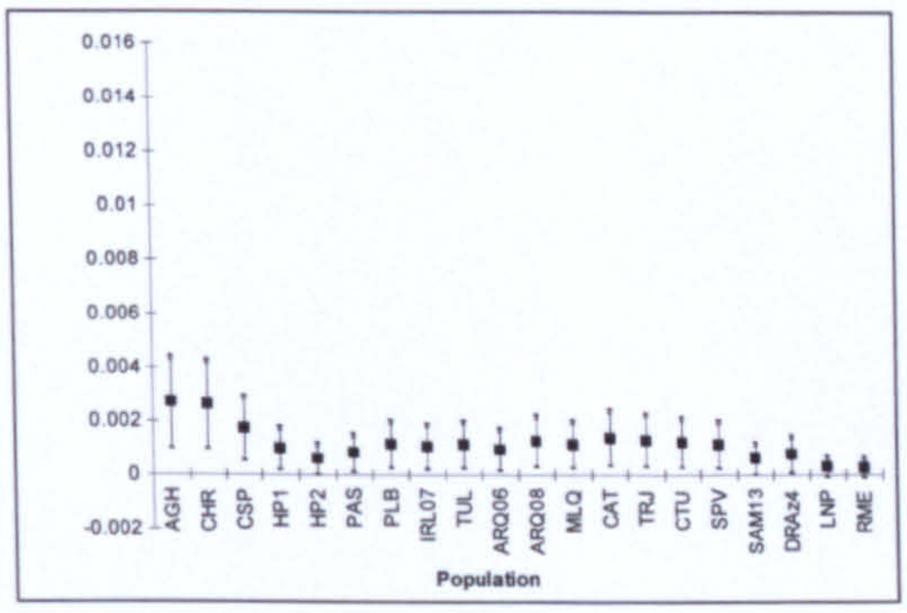
Sub-division tested	df	F Indices	% variation	*P*-value
1. E Pyrenees vs. *Massif Central* vs. *C Pyrenees and NE Spain*				
Among regions	2	0.07725	7.73	<0.001*
Among pops within regions	14	0.03390	3.13	<0.001*
Within pops	775	0.10853	89.15	<0.001*
2. E Pyrenees vs. *Massif Central*				
Among regions	1	0.06287	6.29	<0.001*
Among pops within regions	12	0.04173	3.91	<0.001*
Within pops	642	0.10198	89.80	<0.001*
3. E Pyrenees vs. *C Pyrenees and NE Spain*				
Among regions	1	0.03266	3.27	<0.05*
Among pops within regions	9	0.01459	1.41	<0.05*
Within pops	527	0.04678	95.32	<0.05*
4. Massif Central vs. *C Pyrenees and NE Spain*				
Among regions	1	0.19329	19.33	<0.001*
Among pops within regions	7	0.05501	4.44	<0.001*
Within pops	381	0.23767	76.23	<0.001*

Figure 3.7 Plots of nucleotide diversity (Mean π with standard deviation bars) for three loci characterized from populations of *P. ariasi*. Scale equal on all graphs. Zero diversity of cyt b in population RME was not plotted.

(a) Nuclear apyrase



(b) Nuclear elongation factor-1α



(c) Mitochondrial cytochrome b

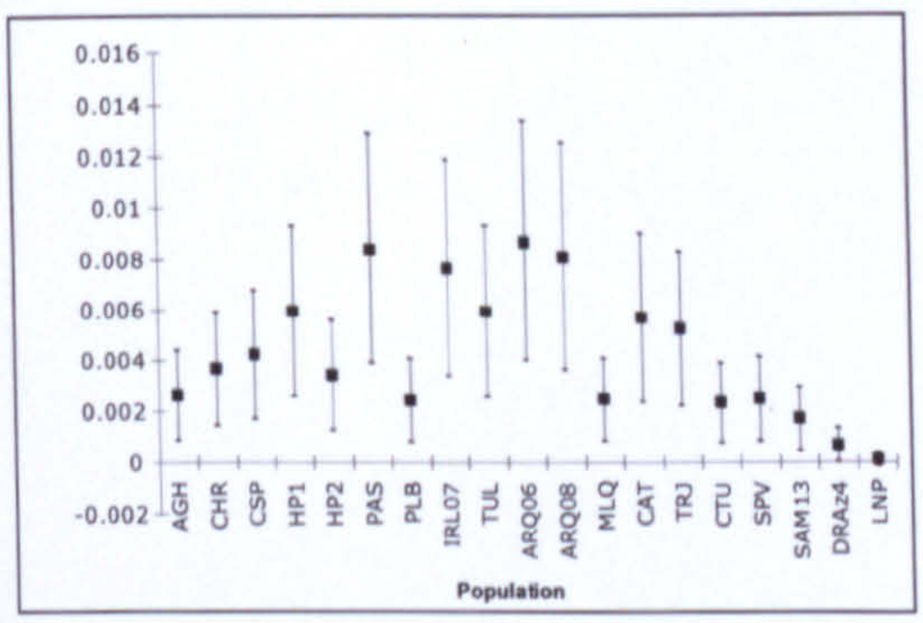


Figure 3.8 Plots of nucleotide diversity π (Pi) for synonymous sites [Pi(s)] and nonsynonymous sites [Pi(n)] for three loci characterized from populations of *P. ariasi*. No nonsynonymous changes were observed in EF-1α.

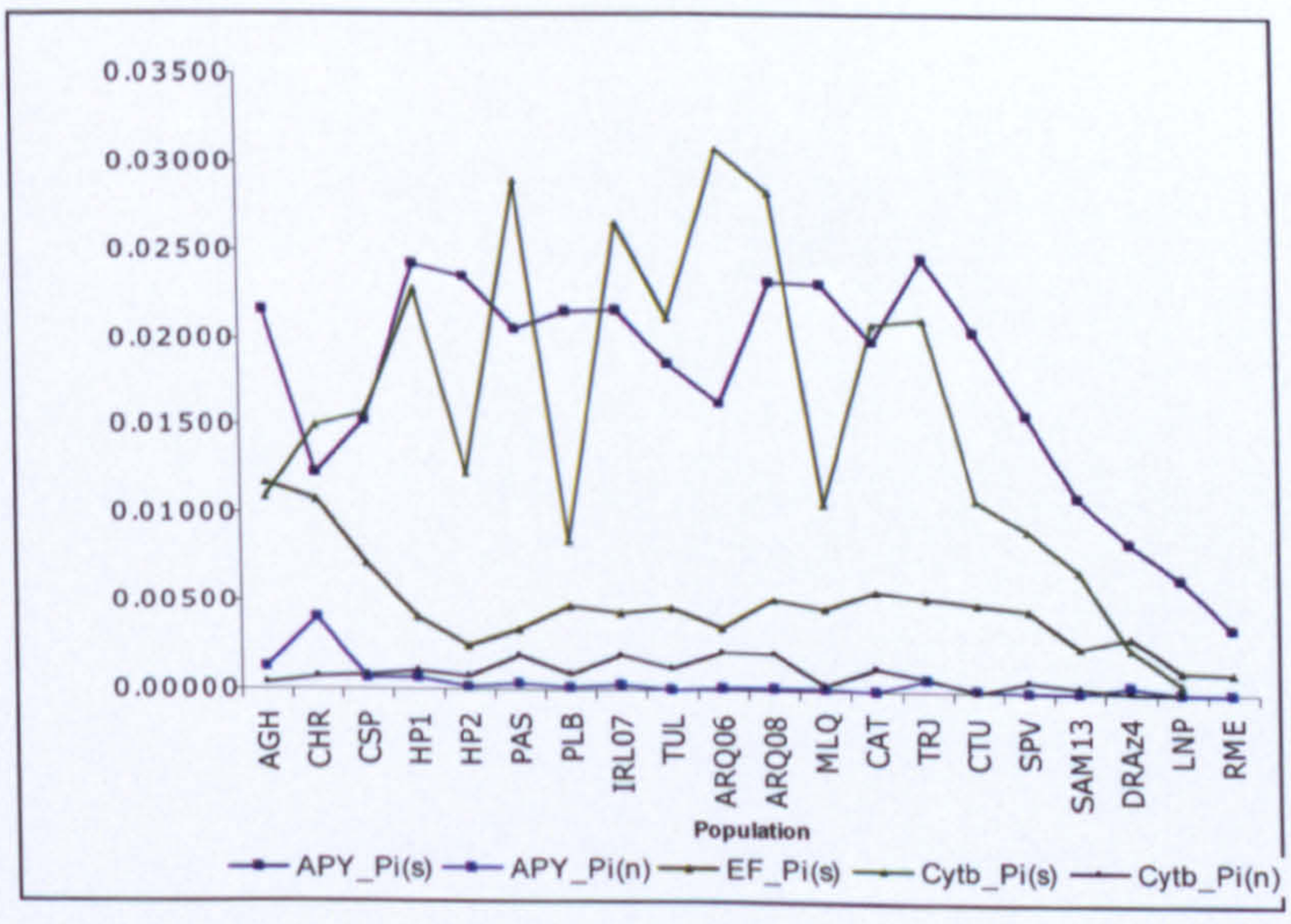
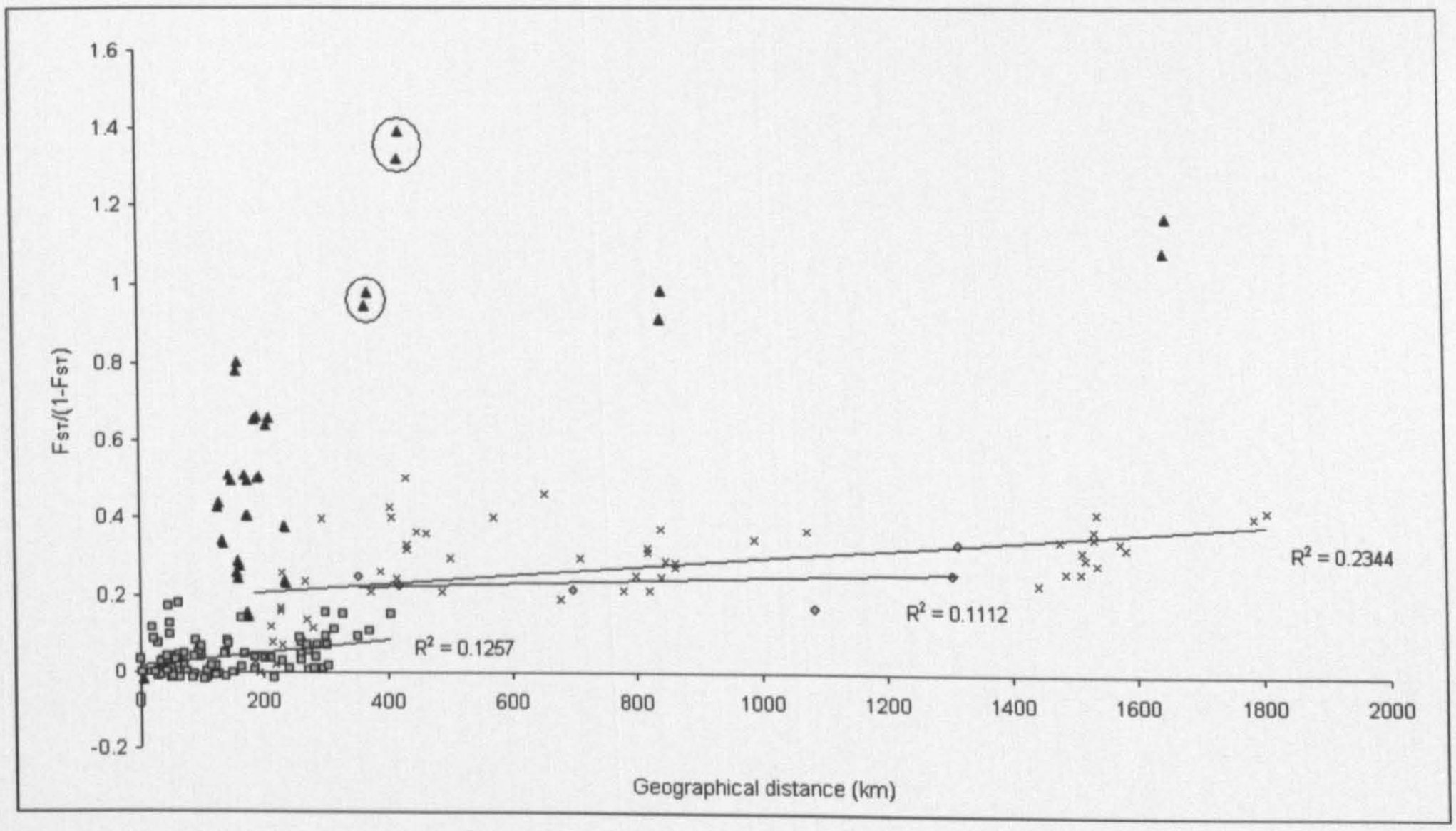


Figure 3.9 Plot showing the association between genetic distance [$F_{ST}/(1-F_{ST})$] and straight-line geographical distance for pairs of populations of *P. ariasi*. An isolation-by-distance (IBD) model was supported for pairwise population comparisons whether supported (z-test) regression outliers (circled data points), attributed to comparisons with bottle-necked Lot France, were included or excluded. Pairwise comparison symbols; triangles: Lot with all other populations; squares: within France (excluding Lot); circles: within outgroups; crosses: across the Pyrenees (between France excluding Lot, and outgroups). (See Table 3.1 for location information). Explained correlation given by R^2 values. With the removal of these bottle-necked populations, a dbRDA conditional test supported regionality to predict genetic distance beyond that explained by geographical distance (covariate), identifying a barrier between populations N and S of the Pyrenees (see text).



3.4 Discussion

This study offered the most detailed evaluation to date of the natural genetic variation of a sandfly salivary peptide. The results were the first to record phylogenetic support for the occurrence of a gene duplication event in a *Phlebotomus* salivary peptide based on direct DNA sequencing of wild sandflies. This study found that the targeted fragment of the salivary peptide apyrase is predominantly under purifying selection across *Phlebotomus*, including the *P. ariasi* lineage. Testing for selection on different taxonomic levels allowed an investigation of the processes affecting apyrase evolution at several evolutionary time-scales. The results did not statistically support persistent or contemporary positive or balancing selection and, therefore, contest the hypothesis of a sandfly peptide-host-parasite meditated arms race on apyrase, a salivary peptide that can putatively protect against *L. infantum* in the western Mediterranean. This study practically presented a molecular protocol for PCR amplification and accurate sequence genotyping of an apyrase fragment for *Phlebotomus* species, which for *P. ariasi* involved an optimized PASA system using a limited number of allele-specific primers.

3.4.1 Evolutionary significance of apyrase gene duplicates in some *Phlebotomus*

This study presented the most extensive *Phlebotomus* phylogeny for apyrase, identifying at least two *Phlebotomus* apyrases, in part by species paraphyly in a Bayesian phylogeny (Figure 3.3a). Orthologous sequences reconstructing the true (e.g. species) phylogeny included the pern491 lineage (Anderson *et al.*, 2006) and identified a paralogous pern490 lineage. The timing of this supported a gene duplication event that included the common ancestor to *P. kandelakii* and the *P. perniciosus* complex, but not earlier within *Larroussius* or its sister subgenera, *Adlerius* and *Transphlebotomus*. This is a result contrary to Anderson *et al.* (2006) who hypothesized, without support, a duplication event in *P. perniciosus* with a subsequent loss in *P. ariasi.* Kato *et al.* (2006) showed the apyrase of two *Phlebotomus* species, *P. duboscqi* and *P. papatasi*, to be closely related and apart from other sandfly apyrases. Along with *P.* (*Euphlebotomus*) *argentipes*, my apyrase phylogenies that included these species were incongruent with both taxonomic and other gene trees, which could be further support for other apyrase paralogues. However, alternative explanations cannot be ruled out, namely the confounding effects of genetic distance (to resolve would require more taxon sampling than the current study), or the low resolving power of the short apyrase fragment utilized (as seen for the gene tree of elongation factor-1α in Chapter 2).

Gene duplication events could be widespread in salivary peptides, having been identified by the phylogenetic analysis of the cDNAs of some multicopy salivary peptides of *Phlebotomus*, e.g. the D7 and SP15 like protein families (Anderson *et al.*, 2006; Elnaiem *et al.*, 2005). Duplicated genes are commonly assumed to evolve under weaker selection than nonduplicated genes, and are fundamental to the process of adaptive evolution (Ohno, 1970; Hurles, 2004). In accordance with previous observations (Lynch and Conery, 2000), acceleration in the evolution of the apyrase paralogue (pern490 lineage) occurred immediately after duplication. Relaxed selective constraints and/or positive selection can cause asymmetric evolutionary rates by accelerated nonsynonymous changes in one duplicate (Zhang *et al.*, 2003; Moore and Purugganan, 2003), leading to a new active site being fixed by drift or selection. This was observed in apyrase: immediately post-duplication positive selection (ω significantly > 1) was supported in the paralogous lineage (Table 3.4), consisting of only nonsynonymous changes, a pattern found in other duplicate systems (Emes and Yang, 2008; Jia *et al.*, 2003). However, this period of adaptation appeared to be a single episode and not persistent, because purifying selection predominated across the entire lineage ($\omega = 0.375$) and in terminal branches. Random site models also supported some 3% of sites under selection, although no specific site could be statistically identified. Increased taxon sampling might permit a clearer conclusion.

The maintenance of an apyrase duplicate over time, in this case in multiple taxa, might indicate a conferred fitness advantage. Although episodic adaptive selection (positive directional) was detected, overall each apyrase was found to be subject to purifying selection maintaining both the apyrases as calcium-activated nucleotidases (CANs) as revealed by BLAST. However, a considerable proportion of amino acid replacements occurred between duplicate lineages, so strict gene functional conservation is unlikely, suggesting possible subfunctionalization (partitioning of ancestral function), which relevant activity experiments would need to confirm. Evolutionarily, the maintenance of duplicates may benefit the sandfly by increasing the expression level or enzyme efficiency of apyrase, to further aid abrogation of host ADP induced platelet aggregation, or by the adaptive improvement of the ancestral apyrase function (Hahn, 2009). There are no reports suggesting that those *Phlebotomus* species with apyrase duplicates are better blood feeders than flies with only a single apyrase.

3.4.2 No supported sandfly peptide-host-*Leishmania* arms race or ecologically meditated adaptive selection in apyrase

When testing for selection, erroneous conclusions result from comparisons of paralogous genes, the presence of cryptic species, non-randomly mating populations, linkage disequilibrium, and demography. To control for these variables: orthologous genes were identified by cloning and optimized PASA systems, as well as genetic models (e.g. phylogenies congruent with other taxonomic and gene trees), and similarity, identity and divergence estimates. Population genetic models assessed samples of *P. ariasi* known to comprise a single phylogenetic and biological species (Chapter 2), used outgroup sequences proven to be orthologous, and disentangled demography versus selection by comparisons with other loci. LD showed ambiguous results, where non-random association of alleles was detected in one of two populations between apyrase and nuclear EF-1α. Through natural selection, LD results either from epistatic selection for gene combinations (Lewontin, 1964) or from selective sweeps involving sites down- or up-stream of the targeted fragment (Kim and Nielsen, 2004). Either selection process is unlikely, as this association was not observed in the other population tested. The alternative of neutral admixture of genetically differentiated populations is more likely (e.g. Stephens *et al.*, 1994), as the population showing LD was composed of equal proportions of two mitochondrially divergent haplogroups (Chapter 2).

Selection was tested for within a maximum likelihood framework, which used the nonsynonymous/synonymous substitution rate ratio (ω) as a measure of selective pressure at the protein level on a *Phlebotomus* phylogeny. At this long time-scale, apyrase was shown to be under predominantly purifying selection, with selection pressure being heterogeneous among branches and orthologues more selectively constrained. Although I cannot reject adaptation common to sandflies, because the analyses did not test across families, Fixed-site tests revealed no evidence of positive selection in codons considered to be functionally important and hypothesized to be evolving under adaptive evolution in sandfly apyrase (Anderson *et al.*, 2006). Moreover, Random-site models failed to identify functional divergence by positive selection in one or a few amino acids, although a low number of sites (~1% to 3%, none with known apyrase associated function), were found to be under such positive selection by BEB. The lack of power of BEB adds further evidence to support the absence of persistent selection on apyrase, because the method fails to detect site-specific positive

selection unless multiple substitutions occur at the same codon position throughout the phylogeny (Yang *et al.*, 2005). However, the method is not robust against intragenic recombination (Anisimova *et al.*, 2003), which was detected in the intra-specific analysis. The implication of a few adaptive sites occurring in apyrase is unclear. In comparison, the tick salivary peptide Salp15 used by the pathogen *Borrelia burgdorferi* to infect their host showed 29% to 54% of sites under positive selection (Schwalie and Schultz, 2009).

The molecular evolution of *P. ariasi* apyrase, sampled from a range of spatio-temporal populations representing different geographical environments that originated from across the species' South-North range, is most likely not to be under positive directional or balancing selection, as tested at different evolutionary time-scales. At the longest-time scale, the PAML analysis using the orthologous *Phlebotomus* apyrase phylogeny revealed purifying selection. No support for adaptive selection was revealed between sister species or among *P. ariasi* populations, which were investigated using population genetic models aimed at detecting longer-term (MK test) or recent selection (*D* statistics after multiple comparison corrections) within *P. ariasi*. This latter conclusion is noted with caution, as I accept the assumption of the absence of recombination may be incorrect, which can lead to conservative inferences of population genetic statistics. Accepting the effects of recombination, an alternative scenario could be proffered for the evolutionary forces acting within *P. ariasi* populations. Per population, but not globally, MK results showed a trend towards longer-term positive selection. Whereas a trend in positive *D* statistics suggested a signal of balancing selection. The latter may have occurred through local recombination, as up to two recombination events were detected in *P. ariasi* populations and estimates of population recombination parameter R were positively correlated with nucleotide diversity. However, balancing selection has not been detected for any dipteran immune peptide and perhaps it should not be expected to act on salivary peptides, because it is usually associated only with parasite-mammal interactions for diseases such as tsetse fly-borne sleeping sickness caused by the antigen-switching *Trypanosoma brucei* (Young *et al.* 2008) and anopheline mosquito-borne malaria caused by *Plasmodium* species with highly polymorphic surface antigens (Tetteh *et al.* 2009).

Phylogeographic variation was observed at both the nucleotide and amino acid levels of diversity, a population genetic structure consistent with that given by other loci

(Chapter 2), which identified plausible biographical barriers to gene flow, proposing the use of apyrase as a neutral population genetic marker for *P. ariasi*.

3.4.3 Apyrase for vaccination against Mediterranean ZVL

My analyses did not support persistent positive directional or balancing selection on apyrase across *Phlebotomus* or within the *P. ariasi* lineage, indicating the absence of an arms race model of molecular evolution driven by sandfly peptide-host-*Leishmania* antagonism. Actually, an arms race between sandfly salivary peptides, *Leishmania* and their vertebrate hosts should not be expected for most transmission cycles, especially for the one involving *P. ariasi*. Amongst others, *P. ariasi* (Guy *et al.*, 1984), *P. perniciosus* (De Colmenares *et al.*, 1995), *P. perfiliewi* (Bongiorno *et al.*, 2003), *P. argentipes* (Palit *et al.*, 2005) and *P. papatasi* (Javadian *et al.*, 1977) are all vectors found to be opportunistic feeders. They take blood meals from whichever host is nearby (e.g. dogs, rodents and birds), with some species ingesting multiple blood meals in a single gonotropic cycle (Guy *et al.*, 1984; De Colmenares *et al.*, 1995; Svobodová *et al.*, 2003). In addition, the mean life expectancy of female *P. ariasi* is only 1.54 ovarian cycles (Dye *et al.*, 1987) and with potentially hundreds of flies biting a single host each day, this makes it unlikely that an arms race will be initiated by any one sandfly-parasite-host system.

VL is usually a zoonosis in the Mediterranean Basin, where dogs are the main reservoirs, and thus is considered both a public and veterinary health problem (Dujardin *et al.*, 2008). With the sandfly being a permanent component in the current transmission cycle, anti-*Leishmania* vaccines targeting vector salivary components with immunomodulatory activities are promising third-generation candidates (Titus *et al.*, 2006; Palatnik-de-Sousa, 2008). Recently, Collin *et al.* (2009) reported that immunization with two salivary peptide-specific DNA plasmids of *L. longipalpis* conferred protection against *L. infantum chagasi* in the natural dog reservoir. This immunization study, like the few others investigating these peptides, involved antibody and T_h1 responses (e.g. Morris *et al.*, 2001; Valenzuela *et al.*, 2001a; Gomes *et al.*, 2008; Oliveira *et al.*, 2008; Collin *et al.*, 2009). By-passing the humoral system response, as apyrase does, may prove advantageous for vaccine development, if the involvement of an antibody response is more likely to lead to an arms race, e.g. maxadilan (Milleron *et al.*, 2004b).

The results of this study suggest caution is required when considering the use of apyrase as a broad-spectrum vaccine candidate, because of the presence of duplicate lineages in some *Phlebotomus*. Findings on intra-specific polymorphism should not be extrapolated to other sandflies, but the methodologies presented are a "proof of principle" indicating how a population genetics approach can distinguish between adaptive and neutral evolution of a salivary peptide. Caution is also required, because the effects of a cell mediated DTH response on *Leishmania* pathogenicity can be contradictory (Oliveira *et al.*, 2008).

CHAPTER 4

Fine-scale spatial genetic structure of *Phlebotomus ariasi* in southwest France: effects of landscape fragmentation on gene flow

4.1 Introduction

The South-North distribution of *Phlebotomus ariasi*, a vector of *Leishmania infantum*, goes from North Africa to 45^0 N in central France. Phylogeographic inferences and a high diversity of mitochondrial haplotypes (including multiple cyt b haplogroups) indicated southwest (SW) France as a putative zone of secondary contact or a region occupied by flies dispersing from one (Chapter 2). The lower slopes of the northeast (NE) Pyrenees offered a gateway for *P. ariasi* migrating from Iberia into southwest (SW) France, and/or a springboard for its northward spread, at the end of the last 1-2 glacial periods. In France the distribution of *P. ariasi* and zoonotic visceral leishmaniasis (ZVL) are associated with land cover types suitable for this vector (Rioux *et al.*, 1980). The aim of the current chapter is to determine if landscape heterogeneity in SW France affects the gene flow of *P. ariasi* using fine-scale spatial genetics. An understanding of the landscape genetics (Holderegger and Wagner, 2008) of *P. ariasi* could be informative for modelling the risk of ZVL spread in the changing environments of western Mediterranean Europe (Ready, 2008).

Habitat fragmentation divides continuous populations into smaller isolated remnants (Foley *et al.*, 2005). The differing levels of connectivity between population clusters (the metapopulation dynamic) in part depends on dispersal capabilities through the landscape (Baguette and Dyck, 2007; Saunders *et al.*, 1991), and in part on the spatial attributes of the landscape including fragment area, its coincidence with edge effects, fragment shape, fragment isolation and matrix structure (Ewers and Didham, 2006). Dispersal is a fundamental process determining the response of a species to landscape changes (Dieckmann *et al.*, 1999), which can be assessed using population genetic tools founded on Wright's (1931; 1943) principles. These are dependent on two main components, neighbourhood size and isolation-by-distance. Addressing the latter permits one to distinguish between two population processes, namely the sub-structuring of populations or individuals into homogeneous gene pools, or the dependence of genetic distance on geographical separation (Guillot *et al.*, 2009; Slatkin, 1995). In the context of habitat fragmentation, landscape genetics (*sensu lato*

Holderegger and Wagner, 2008) aims to resolve the degree to which landscapes facilitate the movement of organisms (landscape connectivity) by relating gene flow patterns to landscape structure. In addition to Wright's classical F_{ST} method to estimate gene flow, developments in population genetics have allowed recent and contemporary barriers to gene flow to be identified, and quantification of their geographical scale (e.g. spatial autocorrelation (Smouse and Peakall, 1999) and the STRUCTURE assignment test (Pritchard *et al.*, 2000)). Indirect (genetic) measures of gene flow tend to be underestimated and difficult to measure in long distance dispersers (Peakall *et al.*, 2003). For *P. ariasi*, however, gene flow should be both detectable and congruent with spatial genetic structure, because of its phalanx or stepping stone geographical spread (Chapter 2) and its restricted local dispersal - mark-release-recapture experiments showed its dispersal distance commonly to be around 1 km, with a maximum of 2.2 km (Killick-Kendrick *et al.*, 1984).

The evolutionary consequences of habitat fragmentation have their principles in island biogeography theory (MacArthur and Wilson, 1967), where decreasing fragment size is accompanied by a decline in species abundance and richness. Genetic content is also affected. Fragmentation can isolate small populations, leading to restricted gene flow and reduced levels of genetic diversity - the likelihood of inbreeding is increased through the accumulation of related individuals within the fragments. Consequently evolutionary potential is lowered, reproductive fitness compromised, and extinction risk is elevated (Couvet, 2002; Spielman *et al.*, 2004). In addition to genetic impoverishment, restricted gene flow with fragmentation alters metapopulation dynamics. It can increase the rate of population differentiation between fragments by genetic drift, and can also affect population or individual behaviour, for example encourage longer-distance dispersers (Dyck and Baguette, 2005). It follows that fragmented landscapes could significantly affect disease epidemiology through vector persistence, creation of new environments that change vector-host encounters and vectorial traits. As estimates of gene flow made using direct measurements of sandfly dispersal are inefficient and labour intensive, fine-scale spatial genetics provides a tractable alternative.

The current study region lies in SW France, including the NE Pyrenees and southern foothills of the Massif Central (SMC), France (Figure 4.1). This is composed of a heterogeneous landscape, the presence of oak or broadleaf forest (Rioux and Golvan, 1969; Ready *et al.*, in prep) is favourable to *P. ariasi*, but is fragmented by

natural features such as rivers and low mountain passes. In addition anthropogenic land cover changes have occurred at specific altitudes, resulting in a matrix of forestry, pastures, orchards, arable crops and vineyards, all associated with scattered small villages, isolated farms and commuter dwellings (Martínez *et al.*, 2007; EDEN project field environmental databasing). In the study area, the contemporary spatial distribution of *P. ariasi* was investigated based on a systematic sampling field strategy unbiased to landscape type. To date, analyses have been conducted using remote sensed data within a Geographical Information System (GIS), with particular focus on the effects of landscape composition and configuration based on vector absence/presence (Martínez *et al.*, 2007) or relative abundances (P.D. Ready, S.S. Mahamdallie and B. Pesson, unpublished observations). In the study area, Martínez *et al.* (2007) confirmed a significant positive association between *P. ariasi* presence and broadleaf forest, and a negative association with the proportion of vineyards, of complex cultivation patterns and other crop types. Furthermore, altitude also contributes to this species' patchy distribution: in the NE Pyrenean foothills *P. ariasi* has an overall range between 120-1,300 m.a.s.l., is relatively more abundant at mid-slope, but found in low numbers (or absent) below 300 m.a.s.l., where unsuitable land covers predominate (urban or agricultural) that fragment the forest (P.D. Ready, S.S. Mahamdallie and B. Pesson, unpublished observations).

Insects are highly susceptible to forest fragmentation (Didham *et al.*, 1996). In Europe, those forests that share a high proportion of their borders with anthropogenic uses are at higher risk of further degradation (Wade *et al.*, 2003), especially where agriculture predominates (Jennersten *et al.*, 1997). In SW France, agricultural policy has caused a change in land use away from traditional cattle farming (blood meal source for sandfly populations) towards the cultivation of maize and sunflowers accompanied by soil drainage changes and increased overgrown fallows (Balent and Coutiade, 1992). Specifically in the study region, Martínez *et al.* (2007) found that broadleaf forest configuration was evolving by an increase in the number of patches over the past 20 years. Only a single study has assessed the population genetic structure of *P. ariasi* in SW France (Chapter 2). Low resolution sampling provided some support for restricted gene flow between populations north of the Pyrenees and Massif Central uplands, as well as, between populations from forested hillsides either side of the Carcassonne corridor (ca. 20 km) - a zone of unsuitable habitat for this species i.e. low-altitude (< 300 m.a.s.l.), deforested, urban and with major road and rail transport routes.

The author is not aware of any published hypervariable microsatellite loci for *P. ariasi*, the preferred markers to characterize population structure (Sunnucks, 2000). Therefore, this investigation characterized the DNA sequences of five loci of *P. ariasi* (Those of Chapters 2 and 3), which at the population level are both polymorphic and not subject to adaptive selection. DNA sequence data offer an advantage over genotyped markers as their assessment can include both divergence and frequency parameters, and moreover, current population structure can be distinguished from historical events through molecular phylogenies (Sunnucks, 2000). Both mitochondrial (cyt b) and nuclear loci (EF-1α and apyrase [protein coding], AAm20 and AAm24 [anonymous loci]) were characterized, which provided independent tests of hypotheses. This study evaluated the fine-scale spatial genetics by two categories of assessment: (i) combining multilocus nuclear genotypes which are labile – a single generation of sexual recombination can destroy a genotype – to infer recent and contemporary metapopulation dynamics and relatedness between individuals; and (ii) using single locus allele frequencies and divergence for longer time-scales, to assess the population neutral processes of genetic drift, gene flow and founder effects (Sunnucks, 2000).

The aim of this chapter was to determine, by exploring patterns of gene flow, the presence and causes of non-random population structure of *P. ariasi* in a study region composed of fragmented forest patches inter-dispersed with other land cover and landscape features.

This chapter's aims were:

1. To estimate diversity and relatedness in geographical populations/sub-regions and determine whether the levels observed can be associated with specific spatial attributes of the landscape.
2. To quantify the spatial scale of genetic connectivity between individual *P. ariasi* in this study region.
3. To infer barriers to gene flow across the study region by modelling the statistical dependence between genetic and geographical distance.
4. To assess if population structure supported the recognition of geographical sub-regions, defined *a priori* by their association with forest separated by the landscape.
5. To use the individual as an operational unit in an assignment test to identify contemporary population clusters.

4.2 Materials and methods

4.2.1 Field sampling information

547 *P. ariasi* from 23 spatio-temporal populations where sampled at relatively high density within a high resolution 70 x 70 km field area in SW France, including the NE Pyrenees (bordered to the west and east by the Ariège and Aude rivers, respectively) and the southern Massif Central in the north. The decimal degree coordinates of the boundaries were: north 43.37426333, east 2.541408333, south 42.745900, and west 1.66310000 (measured at sample site using a TomTom Palm GPS system) (Table 4.1). Figure 4.1 shows sampling locations superimposed over a CORINE land cover data layer (processed by Dr J. Cox, LSHTM). The sample region comprised various levels of broadleaf forest fragmentation amongst a matrix of rural land cover types including urbanization. Peri-domestic sandfly populations (each with 11-52 *P. ariasi*) were sampled by CDC miniature light traps in the same month each year (July), during the season of adult activity. Where practically possible, rural houses/farms were chosen with similar domestic fauna, because this might influence local population size at sampling locations: < 10 dogs and only a few large mammals were present in a few small holdings (Table 4.1).

To investigate landscape features that might restrict *P. ariasi* gene flow, populations in the study region were sub-divided into four *a priori* sub-regions of broadleaf forest separated by other land covers and/or the Aude river: Sub-region 1: two sites in the southern foothills of the Montagne Noire of the Massif Central ("SMC") were grouped apart from all other populations, this divided populations based on their position north or south of the low-altitude, deforested "Carcassonne corridor", a transport route between the Mediterranean and the Atlantic coast (Figure 4.1). South of this corridor: Sub-region 2; populations West of the Aude river ("West Other"), but outside of Sub-region 3; the Fôret de la Malepère ("FDM"), an isolated broadleaf forest patch surrounded predominantly by a matrix of vineyards and other crops unsuitable for *P. ariasi*; and Sub-region 4; "East Aude", to the east of the river, this was chosen following the results of isoenzyme studies that showed disparate allele frequencies for one population on either side of the R. Aude (B. Pesson, unpublished data).

To optimize the genetic information for population structure analyses, it was aimed to collect field samples separated by similar distances: locations were targeted where straight-line geographical distances were comparable within and between sub-regions, i.e. ~10 km within the FDM, and ~11 km between FDM and West Other;

Figure 4.1 Map detailing the location of *P. ariasi* sampling sites from southwest France, including temporal capture information, the 2 km buffer zones around sites, and the position of the low altitude and deforested Carcassonne corridor. Upper figure shows a digital elevation map with pie charts representing the proportion of cyt b haplogroups (A-D) within populations; lower figure superimposes a CORINE land cover map for category 311, the distribution of broadleaf forest (green), and shows the clustering of populations of four *a priori* sub-regions.

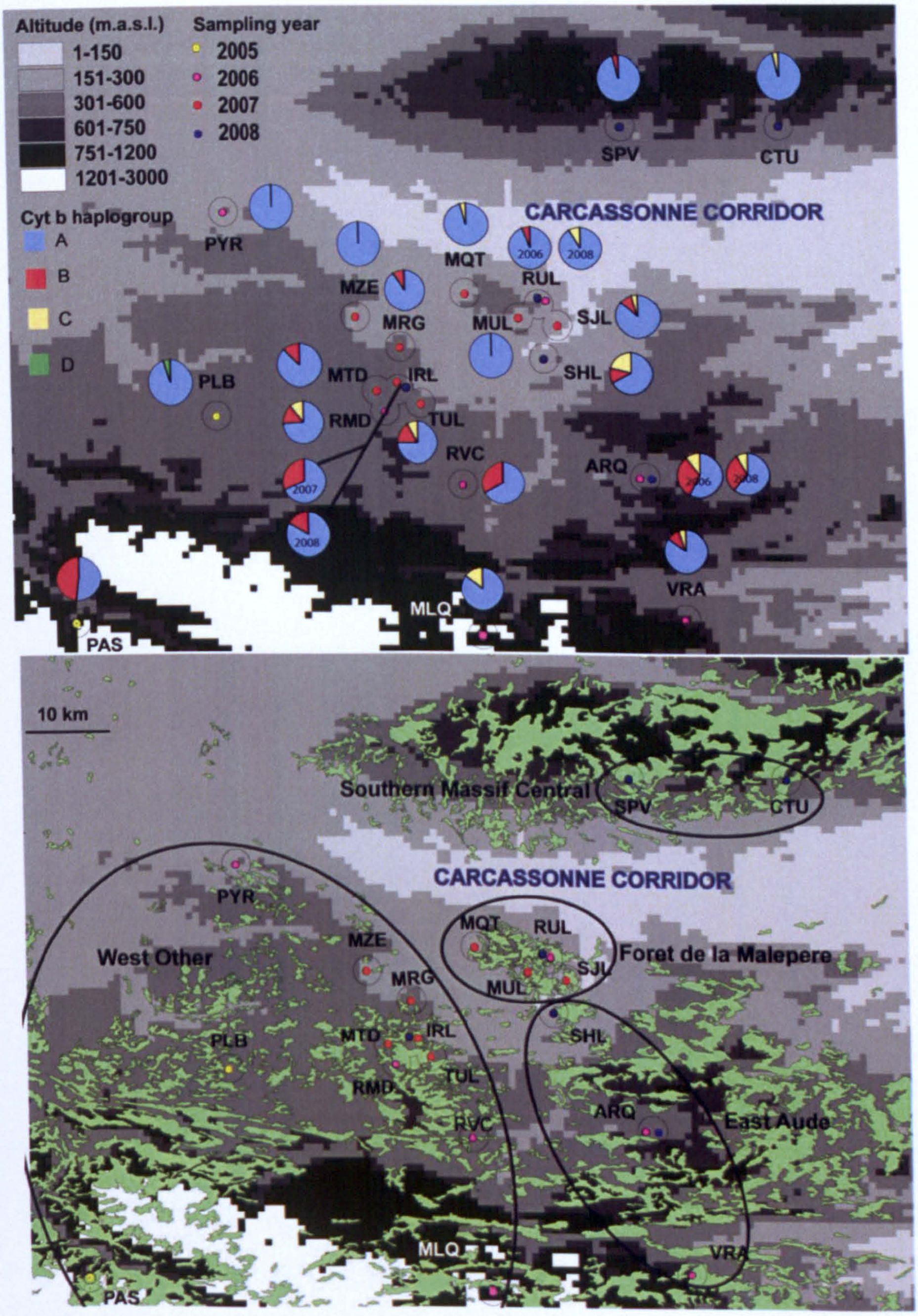


Table 4.1 Geographical locations and associated landscape features, as recorded in a ground-level database or inferred from a CORINE digital land cover map, of *P. ariasi* populations sampled from southwest France.

Population	Latitude (DD)	Longitude (DD)	Capture M/Y	*A priori* assigned sub-region	Field recorded forest fragmentation level	Forest NN (m)†	Forest average D	Field recorded fauna type
PAS	42.76164	1.6631	Jul-05	2. West Other	Continuous	0	0	Canidae
PLB	43.0179	1.8424	Jul-05	2. West Other	Fragmented	460	692.5	Canidae, Leproidae, poultry, Bovinae
PYR	43.26812833	1.855805	Jul-06	2. West Other	Edge-fragmented	1010	1557.5	Canidae, Leproidae
MZE	43.13810833	2.017026667	Jul-07	2. West Other	Edge-fragmented	460	1370	Canidae, Leproidae, poultry
MTD	43.04851	2.041888333	Jul-07	2. West Other	Continuous	0	27.5	Canidae, Bovinae
MRG	43.101675	2.072053333	Jul-07	2. West Other	Edge-fragmented	590	1002.5	Canidae, Leproidae, poultry
RMD	43.02378167	2.051716667	Jul-07	2. West Other	Continuous	0	0	Canidae
RVC	42.93375683	2.146081667	Jul-06	2. West Other	Fragmented	90	455	None
IRL07*	43.04948833	2.068781667	Jul-07	2. West Other	Continuous	40	315	Canidae
IRL08*	43.04948833	2.068781667	Jul-08	2. West Other	Continuous	40	315	Canidae
TUL	43.03310333	2.095721667	Jul-07	2. West Other	Continuous	200	392.5	Canidae, poultry
MLQ	42.7459	2.1672	Jul-06	2. West Other	Continuous	320	345	Canidae, Bovinae
MUL	43.13732	2.218198333	Jul-07	4. Foret de la Malepere (FDM)	Edge-isolated continuous	140	205	Canidae, poultry, Bovinae, Suidae
MQT	43.16749	2.152941667	Jul-07	4. Foret de la Malepere (FDM)	Edge-isolated continuous	600	705	Canidae, poultry
RUL06*	43.15352667	2.24327	Jul-06	4. Foret de la Malepere (FDM)	Isolated continuous	120	185	Canidae, poultry, Bovinae
RUL08*	43.15352667	2.24327	Jul-08	4. Foret de la Malepere (FDM)	Isolated continuous	120	185	Canidae, poultry, Bovinae
SJL	43.12799167	2.26592	Jul-07	4. Foret de la Malepere (FDM)	Isolated continuous	0	167.5	Canidae, Equidae
SHL	43.087815	2.249003333	Jul-08	3. East Aude	Edge-continuous	160	1080	Canidae, poultry, Equidae, Suidae
ARQ06*	42.943805	2.36861	Jul-06	3. East Aude	Continuous	860	1440	Canidae, Bovinae, Equidae
ARQ08*	42.943805	2.36861	Jul-06	3. East Aude	Continuous	860	1440	Canidae, Bovinae, Equidae
VRA	42.76862667	2.414816667	Jul-06	3. East Aude	Continuous	60	135	Canidae, Leproidae, poultry
CTU	43.37426333	2.541408333	Jul-08	1. S Massif Central (SMC)	Continuous	0	0	Poultry [Leporidae near]
SPV	43.37356	2.34766	Jul-08	1. S Massif Central (SMC)	Continuous	0	0	Canidae, Leproidae, poultry

DD = decimal degrees; M/Y = month/year. * Temporal repeats from each of the three southern sub-regions. † Geographical distance from location coordinate to the nearest neighbour (NN) broadleaf forest patch, and average distance (D) to forest within a 2 km buffer: measurements estimated from a CORINE land cover map.

and greater distances within West Other, comparable to across the Carcassonne corridor, ~25-50 km (Figure 4.1).

On a finer-scale, populations were also independently assigned to two categories of broadleaf forest structure, "fragmented" or "continuous", defined by their immediate proximity to a broadleaf forest patch estimated in ARCVIEW (v3.2) from the CORINE land cover map, and field observations. Two proximity measures were considered. Firstly, straight-line distance from a location coordinate centroid to the nearest neighbouring (NN) forest patch. Secondly, by quantifying the extent of forest surrounding a location coordinate centroid by averaging the distance to forest over the four 90° axes within a 2 km buffer zone (automated in ARCVIEW v3.2) (If the nearest forest patch fell outside the buffer, then the maximum distance of 2 km was recorded) (Table 4.1). Only a single study has directly measured the dispersal capability *P. ariasi*, with mark-release-recapture experiments showing fewer than 4% of flies to be recaptured > 700 m from the release points and two thirds of females being recaptured within 250 m (Killick-Kendrick *et al.*, 1984). Accordingly, collection sites were categorized as having: fragmented forest (N = 5), where the nearest forest patch was > 400 m and/or the average distance to forest was > 700 m; and continuous forest (N = 18), where the nearest forest patch was < 400 m and/or the average distance to forest was < 700 m) (Table 4.2). Field studies 'ground truthed' CORINE categories. Consequently, population ARQ was classified as having continuous forest based on field observations that recorded the site to be in close proximity to dense patches of coniferous forest mixed with broadleaf trees - the CORINE map recorded no nearby forest (NN = 860 m).

The contribution of temporal population processes on the level of *P. ariasi* genetic variation was assessed at three sampling locations, one in each southern sub-region: ARQ in 2006 and 2008 (East Aude); IRL in 2007 and 2008 (West Other); and RUL in 2006 and 2008 (FDM). For the repeat collections, CDC traps were placed in identical locations during a similar period in the sandfly season, to minimise the confounding effects of local sampling variables. For some fine-scale analyses nine populations were considered to be outliers if they were separated by moderate geographical distance or associated with changes in bioclimate: populations CTU, SPV, PYR, PLB, PAS, MLQ, ARQ06/08, VRA (Figure 4.1).

4.2.2 Specimen field collection and preparation

Sandflies were collected, preserved and identified as described in Chapter 2 (section 2.2.1). Environmental features surrounding each capture site were recorded in a standardized PALM database (designed by P.D. Ready, J. Cox, C. Davies and the author), or paper questionnaires (designed by P.D. Ready and the author). This extensive database is not presented as many of the records are not pertinent to this study, but some environmental attributes of sampling locations that are relevant to explain results are referred to herein.

4.2.3 Molecular characterization of known neutral loci

DNA extraction methodology from whole or partial sandflies was described in Chapter 2 (section 2.2.2), and protocols to generate direct sequence data from four nuclear loci (17 populations; AAm20, AAm24, apyrase (APY), elongation factor-1 alpha (EF-1α)), and a single mtDNA locus (23 populations; cytochrome b (cyt b)) were described in Chapter 2 (section 2.2.2) and Chapter 3 (section 3.2.2.1).

4.2.4 Data analyses to assess the fine-scale spatial genetic structure of *P. ariasi*

4.2.4.1 Locus genealogies

To disentangle current population structure from historical demographics, parsimony gene networks were reconstructed for each locus, as implemented in TCS (Clement *et al.*, 2000; see Chapter 2, section 2.2.5.2).

4.2.4.2 Description and visualization of the genetic landscape

The Genetic Landscape Shape (GLS) interpolation procedure of Alleles in Space (AIS; Miller, 2005) was used to help visualize population differentiation across the region. Cyt b (nucleotide sequence) and combined nuclear genotypes were analyzed independently in GLS: the program does not handle differences in ploidy or combined nuclear locus sequence data (M. Miller pers. comm.). In all genotype based analyses, binary codes were assigned to unique DNA alleles. The GLS method creates a three dimensional landscape where the X- and Y-axes represent geographical coordinates (UTM, converted from decimal degrees as collected on the field TomTom GPS system) and the Z-axis defines genetic distance whose peak's infer areas of high genetic distance. GLS interpolation proceeds by constructing a Delauney triangulation connectivity matrix between sample sites from whose mid-points inter-individual

genetic distances are calculated and plotted. A 100 x 100 landscape grid was overlaid over the sample sites, and genetic distances between locations estimated (inverse distance-weighted interpolation; using residual genetic distances; distance weight value 1). Repetitions of various grid sizes and distance weight values were investigated to ensure interpolation parameters did not influence the graphical depiction of the genetic landscape.

4.2.4.3 Estimating genetic diversity and relatedness within populations and *a priori* sub-regions

Concordance with Hardy-Weinberg expectation (HWE) (10,000 permutations, ARLEQUIN v3.11; Excoffier *et al.*, 2005) and testing linkage disequilibrium (LD) across multiple unlinked loci (GENEPOP v4.0; Raymond and Rousset, 1995) were implemented to investigate whether there was panmixia in each population and within *a priori* sub-regions. Significant *P*-values of multiple tests were manually corrected for familywise Type 1 errors by applying a sequential Bonferroni correction ($\alpha = 0.05$) (Holm, 1979).

Relative levels of genetic diversity were estimated by: allelic richness (A) correcting for sample size variation (FSTAT v2.9.3.2; Goudet, 2002); gene diversity (ARLEQUIN v3.11) as haplotype diversity (H_d) for cyt b and expected heterozygosity (H_e) in the diploid nuclear loci. The availability of direct sequence data also allowed the molecular index of nucleotide diversity to be calculated (π, with Jukes-Cantor correction in DNASP v4.90.1; Rozas *et al.*, 2003). For each nuclear locus the coefficient F_{IS} measured inbreeding relative to the global population, where positive and negative values represented decreased heterozygosity (inbreeding), and increased heterozygosity (outbreeding), respectively (FSTAT v2.9.3.2).

Significant relatedness was assessed by estimating the mean pairwise relatedness estimator (R) (Queller and Goodnight, 1989) in GENALEX (v6, Peakall and Smouse, 2006), using combined nuclear genotype data. 95% confidence intervals were used to evaluate the significance of R from the expected null distribution of random reproduction across the sample area (999 random genotype permutations). In addition, 95% CI error bars were derived by 999 bootstrap resampling. When error bars fail to overlap the permuted null, population processes are assumed to increase relatedness ("reproductive skew", e.g. inbreeding or genetic drift).

4.2.4.4 Testing for statistical support for regional genetic discontinuity based on *a priori* sub-divisions

Hierarchical Analysis of Molecular Variance (AMOVA) was applied to single locus data or combined nuclear genotypes to test for the support of genetic variance partitioning between *a priori* defined geographical sub-regions. Practically, a standard AMOVA approach was taken using pairwise distances, and statistical significance for each grouping was calculated using 16,000 permutations (ARLEQUIN v3.11).

4.2.4.5 Between population genetic differentiation and testing for statistical dependence between genetic and geographic distances

Restrictions to gene flow across the study region were tested by estimates of population pairwise genetic differentiation as measured by Φ_{ST} estimated in ARLEQUIN v3.11. Statistical dependence between distance matrices supports can either gene flow according to dispersal ability under an isolation-by-distance model (IBD) or the presence of landscape barriers which limit gene flow (Guillot *et al.*, 2009). A Mantel test (GENEPOP v4.0) or marginal tests (DISTLM v5; Anderson, 2004) (See Chapter 2) were implemented to assess whether predictor variables were correlated with genetic distance. Predictor variables included: geographical distance estimated as either straight-line distance or distance along a broadleaf forest line as assessed from a CORINE land cover data layer (ARCVIEW v3.2); and classification of populations according to geographical subregions. Conditional tests (i.e. distance-based redundancy analysis, dbRNA; Anderson, 2004) were implemented to eliminate the effect of IBD on genetic distance, by treating geographical distance as a covariate.

The spatial scale of genetic connectivity as a function of geographical distance was also inferred by the regression of inter-individual pairwise relatedness coefficients on spatial distance as implemented in GENALEX (v6). Spatial autocorrelation is a combined nuclear genotype approach which has been proposed to have greater power and less variance than a single locus assessment (Smouse and Peakall, 1999). It has been applied to animal taxa with restricted dispersal to quantify dispersal behaviour, when gene flow is restricted and selection absent. The latter is applicable to the loci characterized (Chapters 2 and 3). For investigating spatial autocorrelation, estimated inter-individual pairwise genetic distances were transformed to the autocorrelation coefficient *r*, a measure of genetic similarity between pairs of individuals in cumulatively increasing geographical distance classes. As the value at which positive

spatial genetic structure is detected is affected by the distance class chosen, increasing distance classes were chosen as recommended by Peakall *et al.* (2003), namely starting with the maximum dispersal distance of *P. ariasi* (2 km) and going up to the maximum distance of sampling. Individuals are expected to show positive spatial genetic autocorrelation at short distance classes, but values should decline through zero to become negative, preceded by stochastic oscillations of positive and negative values (Smouse and Peakall, 1999; Peakall *et al.*, 2003). Tests for statistical significance included: 999 permutations of randomly shuffled individual genotypes among geographical locations to recompute a null distribution for r assuming no genetic structure (from which 95% CIs define the range about null r), and 999 bootstrap resampling to estimate 95% CIs around mean r by drawing replacements from within relevant pairwise comparisons within each distance class. Following Peakall *et al.* (2003), the null hypothesis of no spatial genetic structure was rejected only when r exceeded the 95% CI derived from the among-population permutation test, and when the 95% CI about r (estimated from bootstrap resampling) do not intercept the X-axis of $r = 0$. If positive spatial genetic structure is found, the first X-intercept provides a quantitative estimate of the spatial limit of non-random (positive) genetic structure.

4.2.4.6 Identifying disruption to gene flow based on a Bayesian clustering approach

A Bayesian clustering model (STRUCTURE v2.3.1; Pritchard *et al.*, 2000) was used to infer if individuals belonged to one or more populations (K clusters). All five loci were included in the analysis, where the second allele of haploid data was coded as missing (J. Pritchard pers. comm.). A cluster is characterized by a set of allele frequencies at each locus attributed to random drift and restricted gene flow, and therefore a cluster represents homogeneous spatial domains. STRUCTURE proceeds by assigning each individual to its appropriate cluster, the number of which is user defined. For each K the log probability of the data (lnP(D)) is estimated that best describes the fit of the data to its respective K. To infer the 'true' cluster number of the data, a series of K clusters was evaluated (1-5), with 100,000 burn-in steps before 1,000,000 MCMC repeats. 10 randomized replicates were made for each K cluster, to ensure stability of posterior probability. Key summary statistics were checked for convergence and therefore a suitable burn-in value, as recommended by Pritchard *et al.* (2009). Evanno *et al.* (2005) reported that, in most cases, the highest estimated lnP(D) does not provide a correct estimation of cluster number. Instead they estimated the true K through an *ad*

hoc statistic, ΔK, based on the rate of change in lnP(D) between successive *K*-values. Both approaches were taken to infer the true *K*, where CLUMPP (Jakobson and Rosenberg, 2007) was then used to summarize and align multiple replicates from this optimal *K*, to estimate the membership coefficient (Q) of individuals to a cluster, which was then visualized as a box plot.

Two ancestry models were implemented, as recommended by Pritchard *et al.* (2009). The admixture model (setting ADMIXTURE = 1) makes no *a priori* assumptions about population clustering, and therefore was initially implemented to learn about population structure using only genetic information. Secondly, sampling location (not spatial) information was used to modify the prior, in order to prefer clustering solutions that correlate with the locations (setting LOCPRIOR = 1; Hubisz *et al.*, 2009). This is recommended to improve STRUCTURE performance in detecting subtle population structure or when data are less informative. This model was considered justified as each 'population' of flies was taken from one or two traps placed in a single property. Because populations may have been connected before forest fragmentation arose, both models implemented the F model for correlated allele frequencies (FREQSCORR = 1; Falush *et al.*, 2003), and alpha (degree of admixture) was estimated independently per population. Other model priors were left as default, i.e. the parameter for distribution of allele sequences, lambda, was fixed at one.

4.3 Results

4.3.1 Locus polymorphism

Individual *P. ariasi* were characterized by direct sequencing at cyt b (N = 533), AAm20 (N = 374), AAm24 (N = 377), apyrase (APY; N = 394) and EF-1α (N = 382). All five loci were polymorphic and thus potentially informative at this geographical scale, where the number of alleles ranged between 68 (cyt b) and 6 (AAm20). Three estimates for genetic diversity were used to assess marker polymorphism *per se*: allelic richness (*A*), gene diversity (H_e, H_d) and nucleotide diversity (π). Table 4.2 shows that cyt b was observed as one of the most diverse markers with the highest *A* (0.8701, corrected for sample size), and the second highest gene and π diversities (0.646 and 0.00552, respectively; the latter corrected for sequence length as a per site calculation). The two anonymous nuclear loci showed the lowest *A* (AAm20 = 2.278 and AAm24 = 3.827) and gene diversity (AAm20 = 0.432 and AAm24 = 0.504), which may be an indirect result of their short fragment length, because their π diversity was high amongst the nuclear loci (e.g. 0.00492, 0.00629, 0.00440, 0.00114, for AAm20, AAm24, APY and EF-1α, respectively).

4.3.2 Evidence of lineages in cyt b only

Parsimony gene networks showed reconstructions concordant with those reported in Chapter 2 (cyt b, AAm20, AAm24, EF-1α), and Chapter 3 (APY). All nuclear networks showed shallow genealogies, with low frequency alleles derived from three or fewer modal haplotypes (allele with more than one derived haplotype or a frequency over 10) with five or fewer mutational steps between them. Summarising: AAm20 connection limit 3, mode 20m01 with four one-step radiations including a second mode 20m02 with a single radiation; AAm24 connection limit 4, three modes 24m06 with five one-step radiations, including mode 24m01 with two one-step radiations including mode 24m07; APY connection limit 9, three modes APYa01, 02, 03 with four to seven one-step radiations and a maximum derived haplotype at five mutational steps from any one mode; and, EF-1α connection limit 12, three modes EF01, 02, 03 with three to seven one-step radiations and a maximum derived haplotype at four mutational steps from any one mode. Cyt b was the only locus to show evidence of lineages, where four haplogroups (A-D) occurred in the study region. Lineage/

haplogroup D was novel, found in a single fly having 10 mutational steps from the predominating haplogroup A, which had the most extensive radiation as observed in Chapter 2. Haplogroup B and C were 14 and 5 mutational steps from haplogroup A, where only these latter two haplogroups were connected by multiple (3) most parsimonious pathways.

4.3.3 Tests supporting within population and sub-region panmixia and linkage equilibrium

Each population at each locus adhered to HWE, after a sequential Bonferroni correction was applied (only SJL was significant before correction, $P = 0.01$ for higher than expected heterozygosity). Nine out of 170 Fisher exact probability tests undertaken (per population for each locus pair) statistically supported linkage disequilibrium (LD) ($P < 0.05$), but none after sequential correction. Moreover, LD was not supported overall between any locus pair (Fisher's exact test $P > 0.05$). A study of *a priori* sub-regions supported panmixia according to HWE and random association of alleles by LD for all loci ($P > 0.05$, after correction). It was therefore considered valid to use all markers for both population and *a priori* sub-region analyses to assess *P. ariasi* statistical spatial genetic structure.

4.3.4 No statistical support for temporal genetic structure in *P. ariasi*

There were three pairs of temporal population repeats and, in two of them each pair shared the same set of common alleles and some rare alleles, both with comparable frequencies. The exception was location RUL, where for cyt b the most common haplotype in haplogroup A (CB25) varied in its frequency between the two years (0.5000 and 0.8824), and the two common alleles in haplogroups B and C were absent in one of the temporal populations. However, temporal genetic structure was not statistically supported for any repeat: for each locus pairwise Φ_{ST} was non-significant ($P > 0.05$) and estimated at ≤ 0.04209 - low genetic differentiation. The spurious result for RUL may be a sampling artefact (N = 12-17); N was not so low in the two other temporal populations (ARQ and IRL with > 22 flies each year).

4.3.5 Some genetic impoverishment associated with fragmented forest

Diversity statistics are informative tools for inferring population structure. For example, allelic richness (A) can decline rapidly in isolated populations because of the

loss of rare alleles through chance events, and gene diversity (H_d, H_e) declines in small populations as a consequence of random genetic drift. Analyses by *a priori* sub-regions revealed neither significant genetic impoverishment at any locus in the putatively isolated FDM sub-region nor higher mean diversity in the main continuous forest West Other sub-region (Figure 4.2a). The FDM was shown only to have a significantly lower mean nucleotide diversity at cyt b than East Aude (t = 4.1735 ± 0.001, df = 7, P = 0.0042), which could be explained by the former's lack of haplogroup B (Table 4.3). Comparing continuous forest populations only, although the isolated FDM populations often had the lowest diversity values compared to like populations from other sub-regions (Table 4.2), their highest diversity values were comparable, e.g. among continuous populations at cyt b: FDM A = 1.923-5.986, H_d = 1.59-0.561, π = 0.0002-0.00345; Others A = 2.765-7.902, H_d = 0.42-0.788, π = 0.00232-0.00868. Furthermore, an evaluation of F_{IS} did not show an inbreeding effect in the FDM (absence of consistent positive values), suggesting its forest patch size is sufficient to maintain an outbreeding population of *P. ariasi*.

An intra-forest analysis within West Other confirmed nucleotide diversity at cyt b was significantly higher in its continuous (N = 4) compared with fragmented (N = 8) forest populations (t = 3.558, df = 10, P = 0.005; P = 0.0061 after Welch correction), the latter reflecting the near fixation of predominating haplogroup A (Figure 4.2b).

Queller and Goodnight's (QG) relatedness estimator for individuals within a population are expected to be ≥0.5 for full sibs, ~0.25 for half sibs, and close to zero for unrelated individuals. In this study, mean pairwise relatedness within each population was low (< 0.25), ranging between -0.101 and 0.177. This evidence did not generally support increased (current) relatedness within populations associated with fragmented forest compared with those from continuous forest. Members of most populations were not more significantly related than expected from the null hypothesis, where genotypes are independently drawn from a panmictic population created across all sample locations - a result expected when migration between populations is sufficiently high and mating is random, which offsets increased relatedness. Two exceptions were: outlier PAS showed significantly more relatedness than expected (R = 0.136, P = 0.015), a result that could be explained by its semi-isolation and lack of migrant exchange from the global mean. Isolation from a forest patch could explain the significant reproductive skew (R = 0.177, P = 0.006) observed in MQT, where

Table 4.2 Genetic diversity statistics by population of *P. ariasi*, arranged by *a priori* sub-regions, at the single mtDNA and four nuclear loci characterized.

	Region		Northeast Pyrenees																			S Massif Ccentral (SMC)	
	Sub-region		West other											Foret de la Malepere (FDM)					East Aude				
	Population		PAS†	PLB†	PYR†	MZE	MRG	MTD	RMD	IRL07	IRL08	TUL	RVC	MLQ†	MQT	MUL	RUL06	RUL08	SJL	SHL	ARQ06†	ARQ08†	VRA†	CTU†	SPV†
	Forest frag category		Cont.	Frag.	Frag.	Frag.	Frag	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Frag.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.
Locus	Statistic	Overall#																							
Cyt b	N	425	52	20	11	17	22	24	19	22	24	24	12	35	23	24	17	12	22	24	38	23	20	24	24
	h	64	12	7	1	3	6	8	5	4	3	7	4	8	5	2	3	6	6	7	15	7	6	10	7
	A	8.701	6.882	5.986	1	2.872	5.393	6.194	4.609	3.512	2.919	5.555	3.993	5.667	4.438	1.923	2.765	5.986	5.013	5.559	8.726	5.865	5.019	7.902	5.982
	H_d	0.646	0.769	0.584	0	0.324	0.641	0.562	0.532	0.541	0.42	0.598	0.636	0.561	0.561	0.159	0.228	0.758	0.532	0.634	0.788	0.664	0.447	0.775	0.754
	π	0.00552	0.00839	0.00242	0	0.00046	0.00351	0.00417	0.00544	0 00764	0.00493	0.00593	0.0084	0.00247	0.00135	0.00022	0.00191	0.00222	0.00345	0.0056	0.00868	0.00808	0.0039	0.0023	0.00248
AAm20	N	374	24	23		17		24		22	23	24		24	23		16	12	23	24	24	23		24	24
	h	6	2	3		2		2		4	3	3		2	2		3	2	3	2	2	3		2	3
	A	2.278	2	2.522		1.995		2		3.091	2.522	2.5		2	1.997		2.75	2	2.522	2	2	2.522		2	2.5
	H_e	0.432	0.467	0.4912		0.214		0.361		0.428	0.382	0.467		0.403	0.264		0.462	0.464	0.405	0.454	0.439	0.478		0.496	0.529
	π	0.00492	0.00519	0.00564		0.00238		0.00401		0.00552	0.00435	0.00534		0.00448	0.00293		0.00562	0.00515	0.00462	0.005	0.00489	0.00548		0.0055	0.00613
	F_{IS}		0.201	-0.153		-0.103		0.195		-0.282	-0.26	0.019		0.072	-0.158		-0.086	-0.082	-0.076	0.27	-0.241	-0.093		-0.007	0.137
AAm24	N	377	24	24		17		24		22	24	24		23	23		17	12	24	24	24	23		24	24
	h	8	5	5		4		5		5	4	5		4	4		3	4	4	4	4	5		3	3
	A	3.827	4.304	4.196		3.9		3.989		4.343	3.495	3.975		3.724	3.474		2.98	4	3.383	3.69	3.358	4.279		2.97	2.984
	H_e	0.504	0.456	0.478		0.59		0.561		0.672	0.573	0.539		0.437	0.432		0.54	0.649	0.46*	0.422	0.332	0.562		0.414	0.441
	π	0.00629	0.00579	0.00646		0.00881		0.00638		0.00887	0.00648	0.00602		0.0053	0.0058		0.00539	0.0088	0.00676	0.0051	0.00426	0.00641		0.0053	0.00544
	F_{IS}		0.087	0.042		-0.1		0.035		0.124	0.203	0.152		-0.098	-0.007		0.243	-0.165	0.004	0.013	-0.134	-0.005		0.096	0.055
APY	N	394	46	24		17		23		22	23	24		24	23		17	12	23	24	22	23		24	23
	h	23	10	6		3		6		7	4	5		7	3		4	5	7	8	6	8		6	5
	A	4.989	6.669	4.755		3		5.197		5.421	3.522	4		5.638	2.997		3.706	5	5.374	5.73	4.883	5.609		4.5	4.04
	H_e	0.707	0.764	0.715		0.67		0.741		0.663	0.6483	0.645		0.761	0.561		0.685	0.667	0.632	0.666	0.59	0.733		0.688	0.665
	π	0.00440	0.00463	0.00464		0.004		0.00506		0.00471	0.00446	0.00393		0.00497	0.00398		0.00395	0.00301	0.00488	0.0034	0.00355	0.00499		0.0044	0.00335
	F_{IS}		0.004	0.009		-0.238		0.003		-0.17	-0.075	-0.233		-0.04	0.229		0.056	0	0.108	-0.194	0.234	-0.008		-0.03	0.285
EF-1α	N	382	48	21		17		21		22	24	23		24	24		15	12	23	24	14	23		24	23
	h	22	10	5		6		10		5	6	5		8	5		5	4	6	4	5	5		4	6
	A	4.957	4.51	4.323		5.037		7.285		4.309	4.383	4.254		5.951	4.633		4.4	4	5.428	3.828	4.571	4.473		3.754	4.719
	H_e	0.606	0.434	0.546		0.595		0.675		0.644	0.617	0.615		0.589	0.635		0.566	0.562	0.718	0.62	0.619	0.649		0.63	0.628
	π	0.00114	0.00082	0.00114		0.00109		0.00155		0.00105	0.00109	0.00112		0.00113	0.0012		0.00096	0.0001	0.00164	0.001	0.00094	0.00125		0.0012	0.00113
	F_{IS}		-0.056	-0.228		0.114		0.012		0.227	-0.221	0.227		-0.063	0.016		-0.311	-0.197	-0.092	-0.285	0.316	0.2		0.143	-0.11

N = sample size; *h* = number of haplotypes/alleles; *A* (allelic richness), corrected using the min. sample size of individuals; Gene diversity as, H_e (diploid) and H_d (haploid); π = nucleotide diversity; F_{IS}, inbreeding coefficient.* Uncorrected significant deviation from Hardy-Weinberg expectation, $P = 0.01$. # Diversity statistics calculated overall for each locus for populations characterized at all five loci only. Forest fragmentation category: Cont. = continuous forest; Frag. = fragmented forest. † Outlier populations as defined by their geographical distance from another population or bioclimate (see Materials and methods).

Figure 4.2a Plotting allelic richness, gene diversity and nucleotide diversity for five loci for each of the four *a priori* sub-regions. Midpoint = mean; boxes = standard error; whiskers = standard deviation. (A comparison of Pyrenean sub-regions revealed FDM to have a significantly lower nucleotide diversity than East Aude; *t*-test $P = 0.0042$).

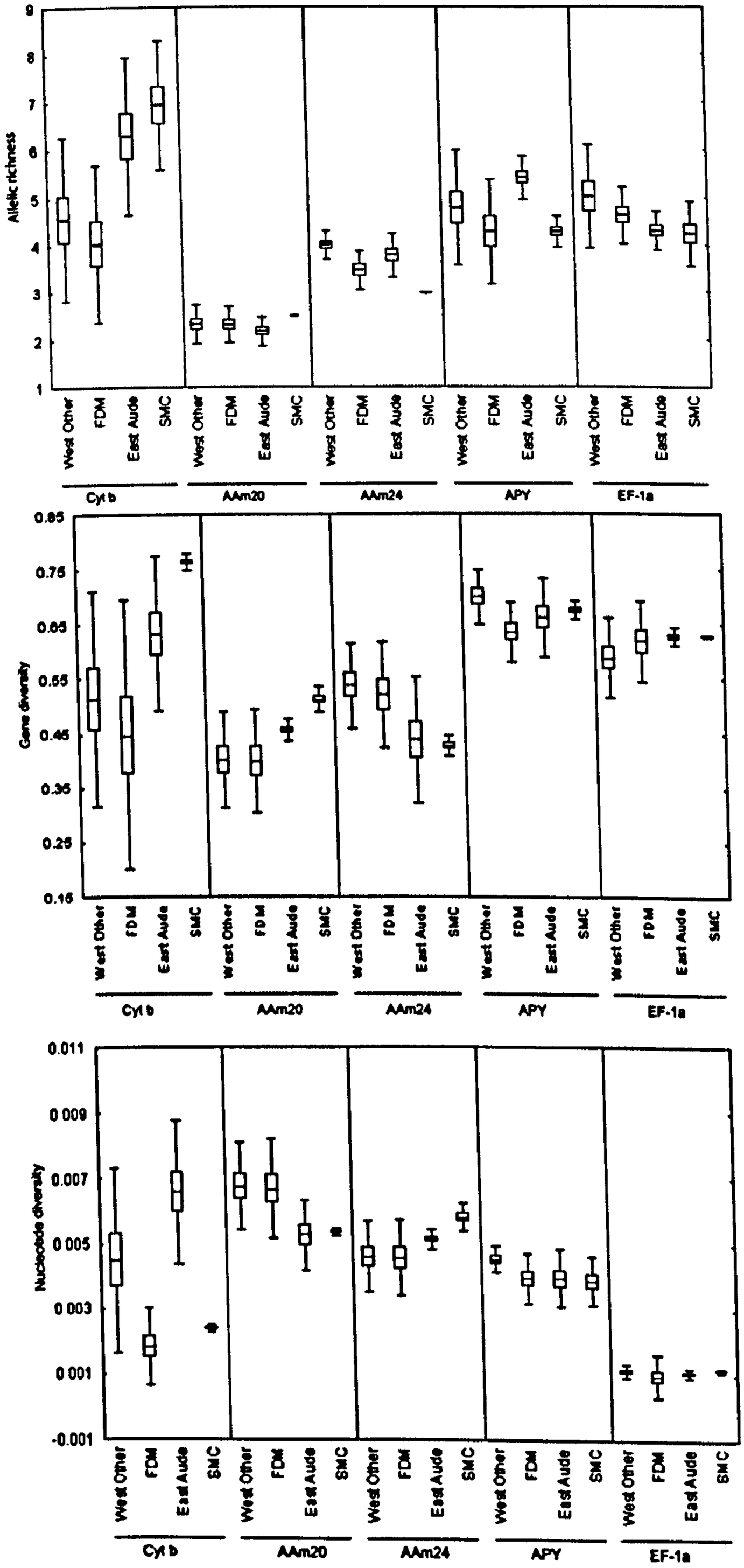


Figure 4.2b Plotting (left to right) allelic richness, gene diversity and nucleotide diversity at locus cyt b, to compare diversity in fragmented (Frag.) and continuous (Cont.) forest populations in sub-region West Other. (*t*-test showed a significantly higher nucleotide diversity in continuous compared with fragmented forest; $P = 0.005$).

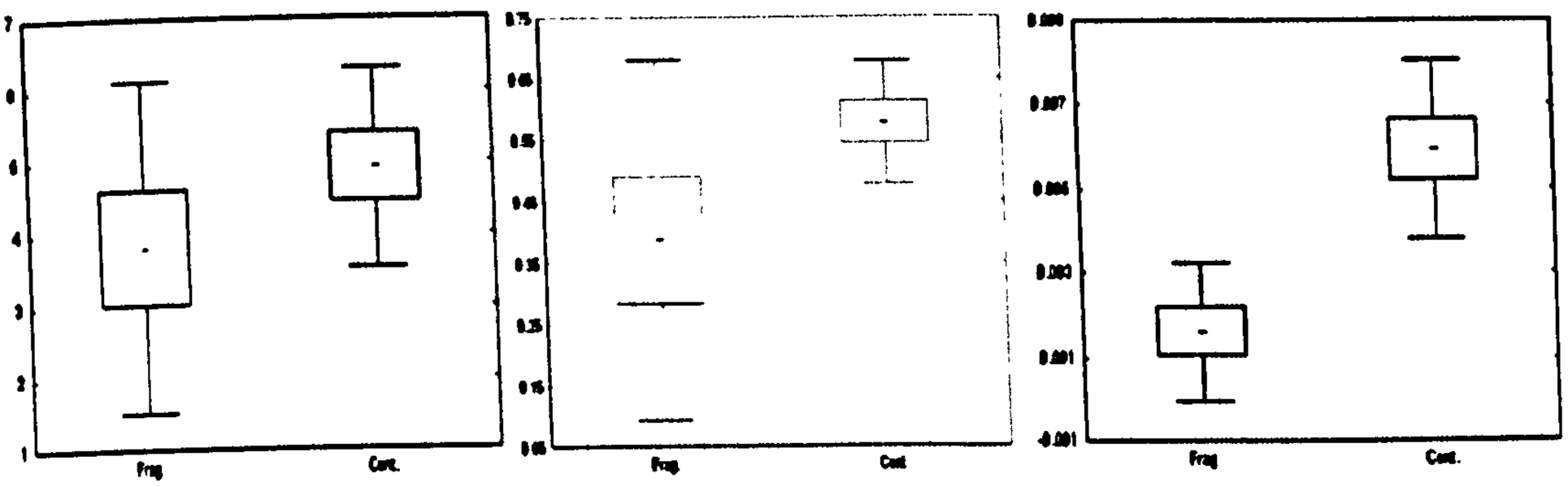
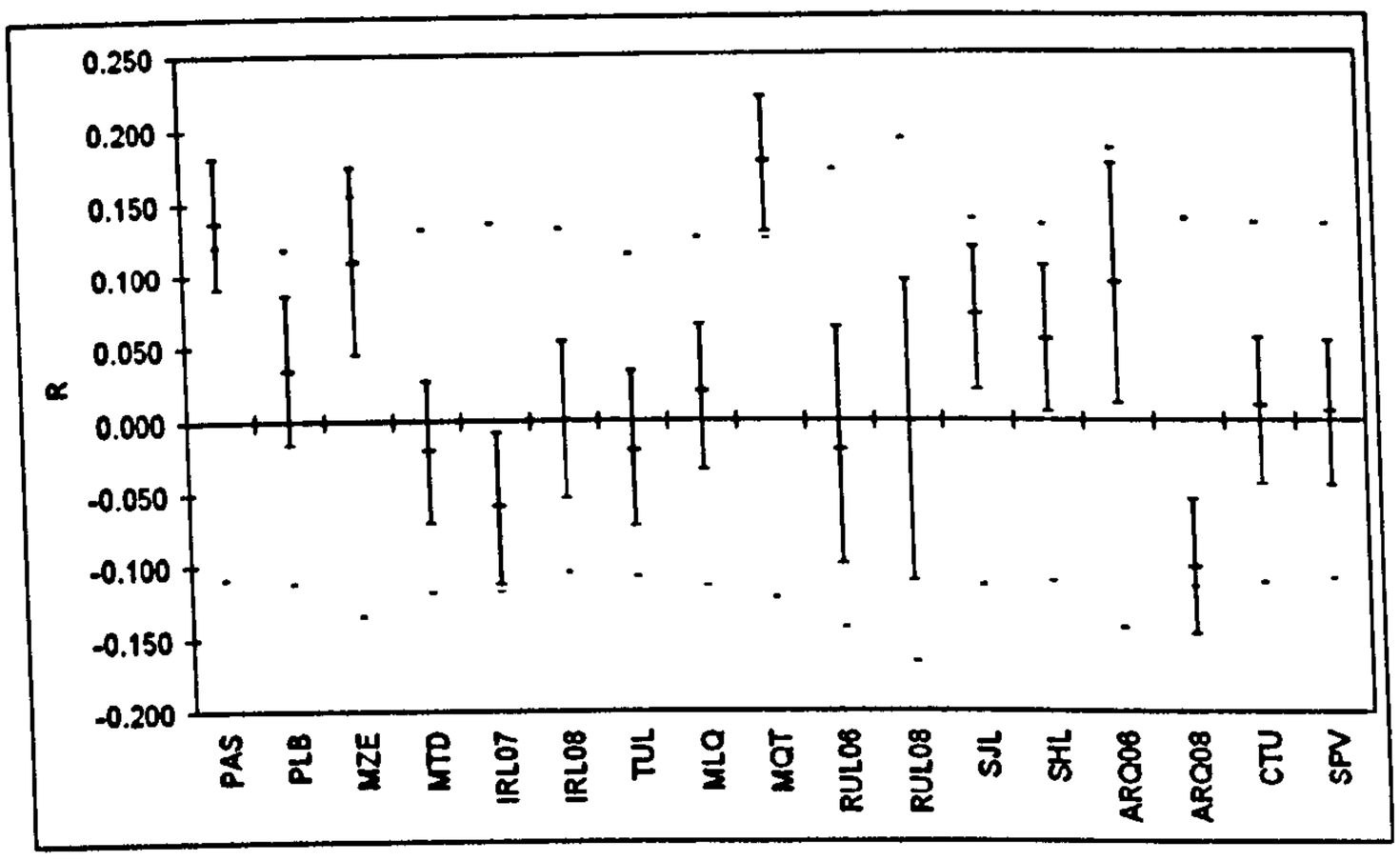
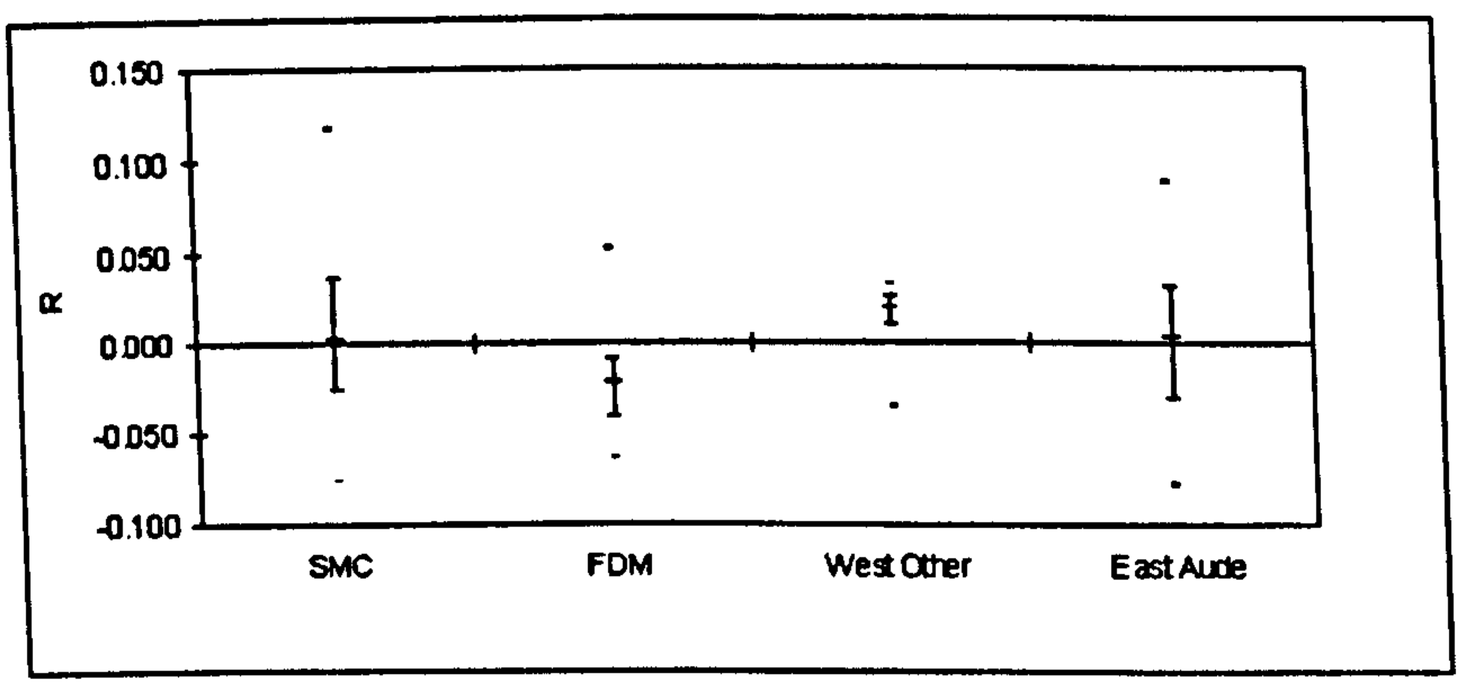


Figure 4.3 Queller and Goodnight's (1989) relatedness estimator (R) for individuals: (a) within *P. ariasi* populations, and (b) within *a priori* sub-regions.

a)



b)

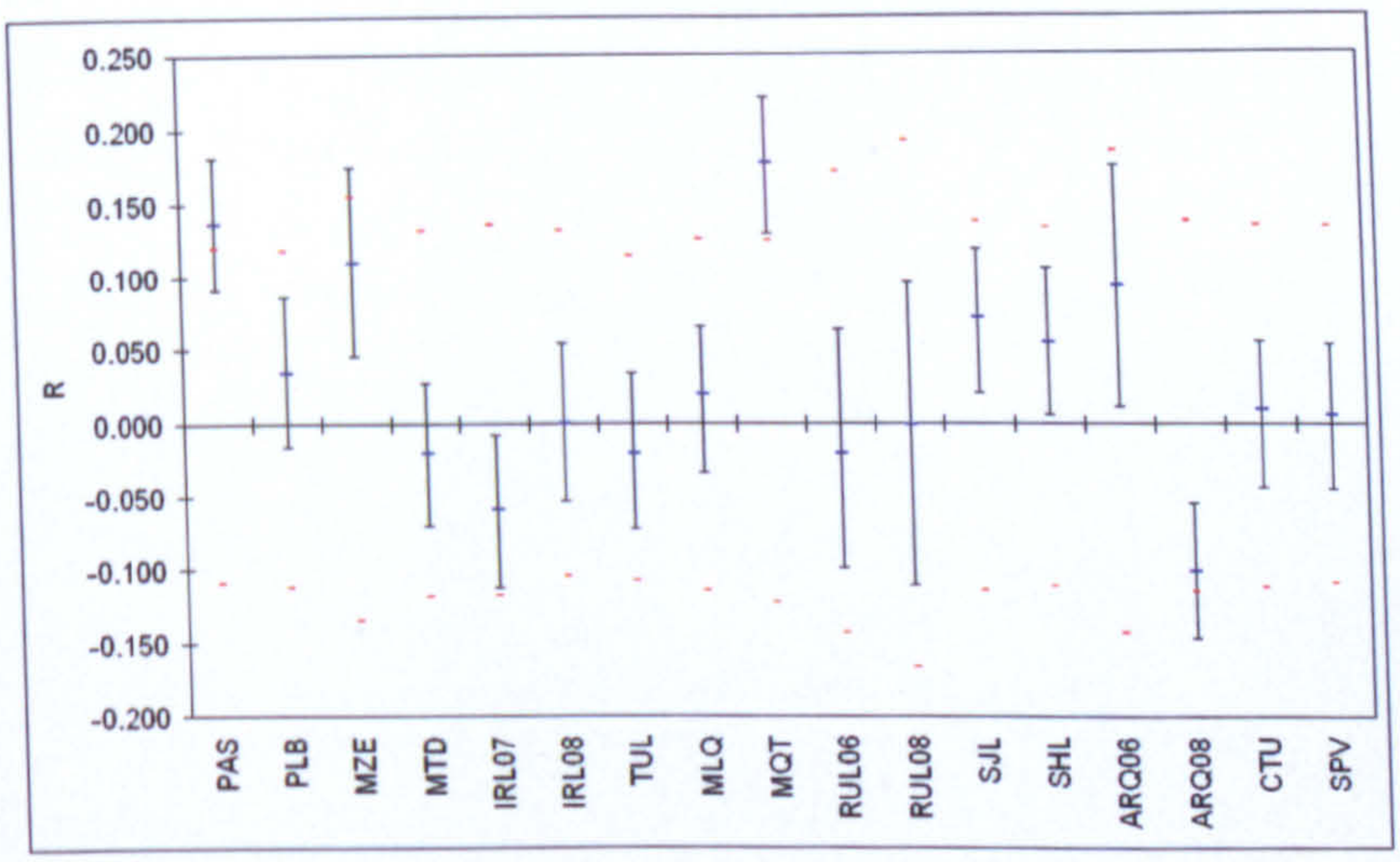


Legend Within population estimates of relatedness are based on the mean inter-individual relatedness (blue lines). (a) Mean relatedness of populations PAS and MQT differed significantly from expectations under a null of population panmixia ($P = 0.015$ and 0.006, respectively; red bars are upper and lower 95% confidence limits of this null). However, only population MQT showed a population mean relatedness whose 95% CI error bars (from bootstrap resampling) fell above the null permuted expectation, indicative of reproductive skew e.g. by inbreeding or random genetic drift. (b) No sub-region showed deviation for relatedness from the global null of panmixia.

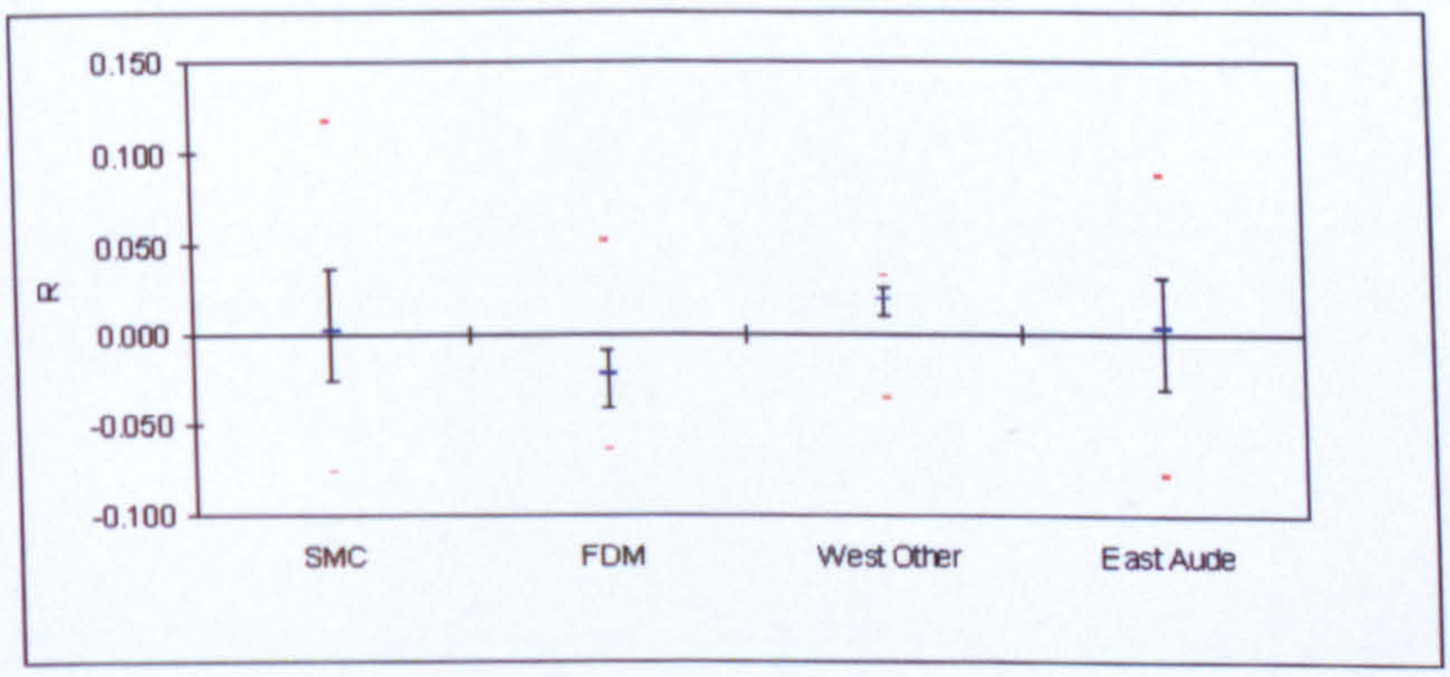
Figure 4.2b Plotting (left to right) allelic richness, gene diversity and nucleotide diversity at locus cyt b, to compare diversity in fragmented (Frag.) and continuous (Cont.) forest populations in sub-region West Other. (*t*-test showed a significantly higher nucleotide diversity in continuous compared with fragmented forest; $P = 0.005$).

Figure 4.3 Queller and Goodnight's (1989) relatedness estimator (R) for individuals: (a) within *P. ariasi* populations, and (b) within *a priori* sub-regions.

a)



b)



Legend Within population estimates of relatedness are based on the mean inter-individual relatedness (blue lines). (a) Mean relatedness of populations PAS and MQT differed significantly from expectations under a null of population panmixia ($P = 0.015$ and 0.006, respectively; red bars are upper and lower 95% confidence limits of this null). However, only population MQT showed a population mean relatedness whose 95% CI error bars (from bootstrap resampling) fell above the null permuted expectation, indicative of reproductive skew e.g. by inbreeding or random genetic drift. (b) No sub-region showed deviation for relatedness from the global null of panmixia.

inbreeding or genetic drift due to population isolation was supported: 95% CI error bars around the population mean relatedness failed to overlap the permuted null (Figure 4.3a). No evidence of increased relatedness from the global null of panmixia was supported within *a priori* sub-regions (Figure 4.3b) (above).

4.3.6 Allele and haplotype distribution patterns match those of forested sub-regions and bottle-necked populations isolated from continuous forest

Restricted gene flow and a genetic bottle-neck were suggested north of the Carcassonne corridor (SMC) by: a high interpolated genetic distance (elevated peaks) geographically positioned at the Carcassonne corridor for the combined nuclear genotype analysis, and to a lesser extent for cyt b (Figure 4.4); the northward loss of haplotypes, with the near fixation of cyt b haplogroup A (Table 4.3; Figure 4.1 pie charts); the absence of private and rare alleles at locus AAm24 (Table 4.5); and the general loss of rare alleles at all loci compared with the main forested West Other sub-region (Tables 4.3-4.7).

A moderate frequency of cyt b haplotype CB04 was found ubiquitously across the Pyrenees in Chapter 2. With the addition of more populations in the NE Pyrenees, it remained present in all continuous forest populations with the exception of two outlier populations, but was lost in four out of five fragmented forest populations and absent in the entire FDM (Table 4.3). This provides some support for the latter's isolation, despite only a ~5-10 km separation from the West Other or East Aude sub-regions. Interpolation of the landscape (using combined nuclear genotype data), showed the FDM to be a region of low genetic distance, indicating few barriers to gene flow within this region (Figure 4.4).

Distinct regional patterns of allele frequencies were not evident south of the Carcassonne corridor at any locus. Genotype distributions/frequencies can reflect short-term population processes (Cornuet *et al.*, 1999) and therefore, their analyses can be informative at this time-scale. However, these were not obviously spatially structured (Appendices 4.1-4.4).

Table 4.3 Haplotype frequencies at locus cyt b for *P. ariasi* originating from southwest France.

Region	Northeast Pyrenees																						
Sub-Region	West Other												Foret de la Malepere					East Aude				S Massif Central		
Pop code	PAS†	PLB†	PYR†	MZE	MRG	MTD	RMD	IRL07	IRL08	TUL	RVC	MLQ†	MQT	MUL	RUL06	RUL08	SJL	SHL	ARQ06†	ARQ08†	VRA†	CTU†	SPV†	Shared
N	52	20	11	17	22	24	19	22	24	24	12	35	23	24	17	12	22	24	38	23	20	24	24	haplotype
Frag category	Cont.	Frag.	Frag.	Frag.	Frag.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Frag.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	level
Haplotype from Haplogroup A																								
CB25	0.4038	0.6500	1.0000	0.8235	0.5909	0.6667	0.6842	0.6364	0.7500	0.6250	0.5833	0.6571	0.6522	0.9167	0.8824	0.5000	0.6818	0.5833	0.4211	0.5652	0.7500	0.4583	0.4167	Multi-region
CB50													0.1304									0.0833	0.0833	Multi-region
CB41															0.0588	0.0833						0.0417		Multi-region
CB75																0.0833						0.1667	0.2917	Multi-region
CB18		0.0500			0.1364		0.0526			0.0417	0.0833								0.0263					Multi-sub-region
CB24	0.0192																		0.0263					Multi-sub-region
CB28	0.0192											0.0286	0.1304											Multi-sub-region
CB33		0.1000				0.0417																		Multi-sub-region
CB47					0.0909	0.0417				0.0417				0.0833		0.0833								Multi-sub-region
CB51						0.0417							0.0435				0.0455							Multi-sub-region
CB21												0.0286									0.0500			Multi-sub-region
CB57																0.1667	0.1364							Sub-region
CB20	0.0192			0.0588																				Sub-region
CB68																						0.0417	0.0417	Region
CB45								0.0455	0.0833															Population
CB31	0.0192																							Population
CB01	0.0385																							Population
CB09		0.0500																						Population
CB13		0.0500																						Population
CB16		0.0500																						Population
CB48				0.1176																				Population
CB46					0.0909																			Population
CB54						0.0417																		Population
CB56						0.0417																		Population
CB61										0.0417														Population
CB29												0.0286												Population
CB27												0.0571												Population
CB30												0.0571												Population
CB70																		0.0417						Population
CB72																		0.0417						Population
CB35																			0.0263					Population
CB40																			0.0263					Population
CB55																			0.0263					Population
CB63																			0.0263					Population
CB79																				0.0435				Population
CB42																					0.0500			Population
CB64																						0.0417		Population
CB66																						0.0417		Population
CB67																						0.0417		Population
CB69																						0.0417		Population
CB107																							0.0833	Population
CB108																							0.0417	Population
Total A	27	19	11	17	20	21	14	15	20	18	8	30	22	24	16	11	19	16	22	14	17	23	23	

Table 4.3 continued. Haplotype frequencies of locus cyt b for *P. ariasi* originating from southwest France.

Region	Northeast Pyrenees																						
Sub-region	West Other												Foret de la Malepere					East Aude				S Massif Central		
Pop code	PAS†	PLB†	PYR†	MZE	MRG	MTD	RMD	IRL07	IRL08	TUL	RVC	MLQ†	MQT	MUL	RUL06	RUL08	SJL	SHL	ARQ06†	ARQ08†	VRA†	CTU†	SPV†	Shared
N	52	20	11	17	22	24	19	22	24	24	12	35	23	24	17	12	22	24	38	23	20	24	24	haplotype
Frag. category	Cont.	Frag.	Frag.	Frag.	Frag.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Frag.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	level
Haplotype code from haplogroup B																								
CB05	0.2308				0.0455		0.1053								0.0588			0.0417			0.0500		0.0417	Multi-region
CB04	0.1538				0.0455	0.0833	0.0526	0.2727	0.1667	0.1667	0.2500							0.0417	0.2105	0.1739				Multi-sub-region
CB52																			0.0263	0.0435				Population
CB32	0.0385																							Population
CB06	0.0192																							Population
CB07	0.0192																							Population
CB17	0.0192																							Population
CB53						0.0417																		Population
CB44								0.0455																Population
CB49											0.0833													Population
CB58																	0.0455							Population
CB59																	0.0455							Population
CB71																		0.0417						Population
CB36																			0.0263					Population
CB60																			0.0263					Population
CB95																			0.0263					Population
CB96																				0.0435				Population
CB97																				0.0435				Population
CB43																					0.0500			Population
Total B	25	0	0	0	2	3	3	7	4	4	4	0	0	0	1	0	2	3	12	7	2	0	1	
Haplotype code from haplogroup B																								
CB26							0.1053			0.0417		0.1143	0.0435			0.0833	0.0455	0.2083	0.0526	0.0870	0.0500			Multi-sub-region
CB62										0.0417														Population
CB34												0.0286												Population
CB37																			0.0263					Population
CB39																			0.0263					Population
CB65																						0.0417		Population
Total C	0	0	0	0	0	0	2	0	0	2	0	5	0	1	0	1	1	5	4	2	1	1	0	
Haplotype code from haplogroup D																								
CB08		0.0500																						Population
Total D	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Legend [Forest] Frag. category: Cont. = continuous forest; Frag. = fragmented forest. † Outlier populations as defined by their geographical distance from another population or bioclimate (see Materials and methods).

Table 4.4 Allele frequencies of locus AAm20 for *P. ariasi* originating from southwest France.

Region	Northeast Pyrenees															S Massif Central	
Sub-Region	West Other								Foret de la Malepere				East Aude				
Pop Code	PAS†	PLB†	MZE	MTD	IRL07	IRL08	TUL	MLQ†	MQT	RUL06	RUL08	SJL	SHL	ARQ06†	ARQ08†	CTU†	SPV†
N	24	23	17	24	22	23	24	24	23	16	12	23	24	24	23	24	24
Frag. category	Cont.	Frag	Frag	Cont	Cont	Cont.	Cont.	Cont	Frag	Cont.	Cont.	Cont	Cont.	Cont	Cont.	Cont	Cont.
Allele code																	
20m02	0.354	0.348	0.118	0.229	0.227	0.217	0.313	0.271	0.152	0.281	0.333	0.239	0 333	0.313	0.326	0.417	0.458
20m01	0.646	0.63	0.882	0.771	0.727	0.761	0.667	0.729	0.848	0.688	0.667	0.739	0.667	0.688	0.652	0.583	0.521
20m03		0.022				0.022	0.021					0.022			0.022		
20m14					0.023					0.031							
20m15					0.023												
20m16																	0.021

Legend [Forest] Frag. category: Cont. = continuous forest; Frag. = fragmented forest. † Outlier populations as defined by their geographical distance from another population or bioclimate (see Materials and methods).

Table 4.5 Allele frequencies of locus AAm24 for *P. ariasi* originating from southwest France. For Table annotations see legend of Table 4.4.

Region	Northeast Pyrenees															S Massif Central	
Sub-Region	West Other								Foret de la Malepere				East Aude				
Pop Code	PAS†	PLB†	MZE	MTD	IRL07	IRL08	TUL	MLQ†	MQT	RUL06	RUL08	SJL	SHL	ARQ06†	ARQ08†	CTU†	SPV†
N	24	24	17	24	22	24	24	23	23	17	12	24	24	24	23	24	24
Frag category	Cont	Frag	Frag	Cont.	Cont	Cont.	Cont	Cont	Frag	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.
Allele Code																	
24m06	0.729	0.708	0.588	0.625	0.477	0.604	0.646	0.739	0.739	0.618	0.5	0.708	0.75	0.813	0.63	0.75	0.729
24m01	0.104	0.083	0.059	0.208	0.159	0.229	0.208	0.13	0.087	0.294	0.125	0.021	0.125	0.063	0.196	0.104	0.125
24m07	0.083	0.146	0.265	0.125	0.295	0.146	0.104	0.043	0.152	0.088	0.333	0.208	0.042	0.104	0.109	0.146	0.146
24m09	0 063	0.042	0.088	0.021	0.045	0.021	0.021	0.087	0.022				0.083		0.022		
24m08					0.023		0.021				0.042	0.063		0.021	0.043		
24m11	0.021																
24m13		0.021															
24m12				0 021													

Table 4.6 Allele frequencies of locus APY for *P. ariasi* originating from southwest France. For Table annotations see legend of Table 4.4.

Region	Northeast Pyrenees															S Massif Central	
Sub-Region	West Other								Foret de la Malepere				East Aude				
Pop Code	PAS†	PLB†	MZE	MTD	IRL07	IRL08	TUL	MLQ†	MQT	RUL06	RUL08	SJL	SHL	ARQ06†	ARQ08†	CTU†	SPV†
N	46	24	17	23	22	23	24	24	23	17	12	23	24	22	23	24	23
Frag category	Cont.	Frag	Frag	Cont	Cont	Cont.	Cont.	Cont.	Frag	Cont.	Cont.	Cont.	Cont.	Cont.	Cont.	Cont	Cont
Allele Code																	
APYa02	0.391	0.396	0.353	0.261	0.114	0.217	0.229	0.229	0.152	0.382	0.375	0.043	0.25	0.136	0.217	0.229	0.37
APYa03	0.239	0.313	0.235	0.391	0.25	0.261	0.208	0.313	0.239	0.206	0.083	0.283	0.104	0.159	0.283	0.25	0.152
APYa01	0.163	0.208	0.412	0.217	0.523	0.5	0.521	0.313	0.609	0.382	0.458	0.543	0.521	0.614	0.391	0.458	0.435
APYa08	0.033	0.021		0.065			0.021	0.063			0.042	0.043	0.042				0.022
APYa15	0.054	0.042			0.045							0.043		0.045	0.022		0.022
APYa23	0.011							0.021							0.022	0.021	
APYa21	0.043						0.021						0.021			0.021	
APYa05	0.011							0.021				0.022	0.021				
APYa20	0.033				0.023									0.023			
APYa29					0.023									0.023			
APYa19	0.022				0.023												
APYa17									0.029				0.021				
APYa28													0.021		0.022		
APYa51		0.021															
APYa42			0.043														
APYa50			0.022														
APYa43						0.022											
APYa44								0.042									
APYa04										0.042							
APYa41											0.022						
APYa27															0.022		
APYa47															0.022		
APYa12																0 021	

Table 4.7 Allele frequencies of locus EF-1α for *P. ariasi* originating from southwest France.

Region	Northeast Pyrenees															S Massif Central	
Sub-Region	West Other								Foret de la Malepere				East Aude				
Pop Code	PAS†	PLB†	MZE	MTD	IRL07	IRL08	TUL	MLQ†	MQT	RUL06	RUL08	SJL	SHL	ARQ06†	ARQ08†	CTU†	SPV†
N	48	21	17	21	22	24	23	24	24	15	12	23	24	14	23	24	23
Frag category	Cont.	Frag	Frag	Cont.	Cont.	Cont.	Cont.	Cont.	Frag	Cont.	Cont.	Cont.	Cont.	Cont	Cont.	Cont.	Cont.
Allele Code																	
EF03	0.146	0.214	0.265	0.167	0.364	0.375	0.304	0.146	0.188	0.3	0.25	0.174	0.333	0.429	0.196	0.229	0.326
EF01	0.74	0.643	0 588	0.548	0.477	0.5	0.543	0.625	0.563	0.6	0.625	0.478	0.521	0.464	0.543	0.542	0.522
EF02	0.031	0.071	0.059	0.071	0.091	0.021		0.083	0.146		0.083	0.087	0.083	0.036	0.152	0.188	0.065
EF05		0.024		0.024	0.045	0.021	0.087	0.042	0.042	0.033		0.152	0.063	0.036	0.087		0.043
EF06	0.021		0.029			0.063	0.043				0.042	0.087				0.042	
EF38								0.021							0.022		
EF10	0.01		0.029	0.024		0.021		0.021									
EF12	0.01			0.071			0.022	0.042									
EF14	0.01			0.024	0.023												
EF08	0.01							0.021									
EF13	0.01	0.048															
EF15	0.01																
EF49			0 029														
EF50				0.024													
EF51				0.024													
EF47										0.033							
EF46										0.033							
EF42											0.022						
EF39														0.036			
EF16																0.022	
EF41																0.022	

Legend [Forest] Frag. category: Cont. = continuous forest; Frag. = fragmented forest. † Outlier populations as defined by their geographical distance from another population or bioclimate (see Materials and methods).

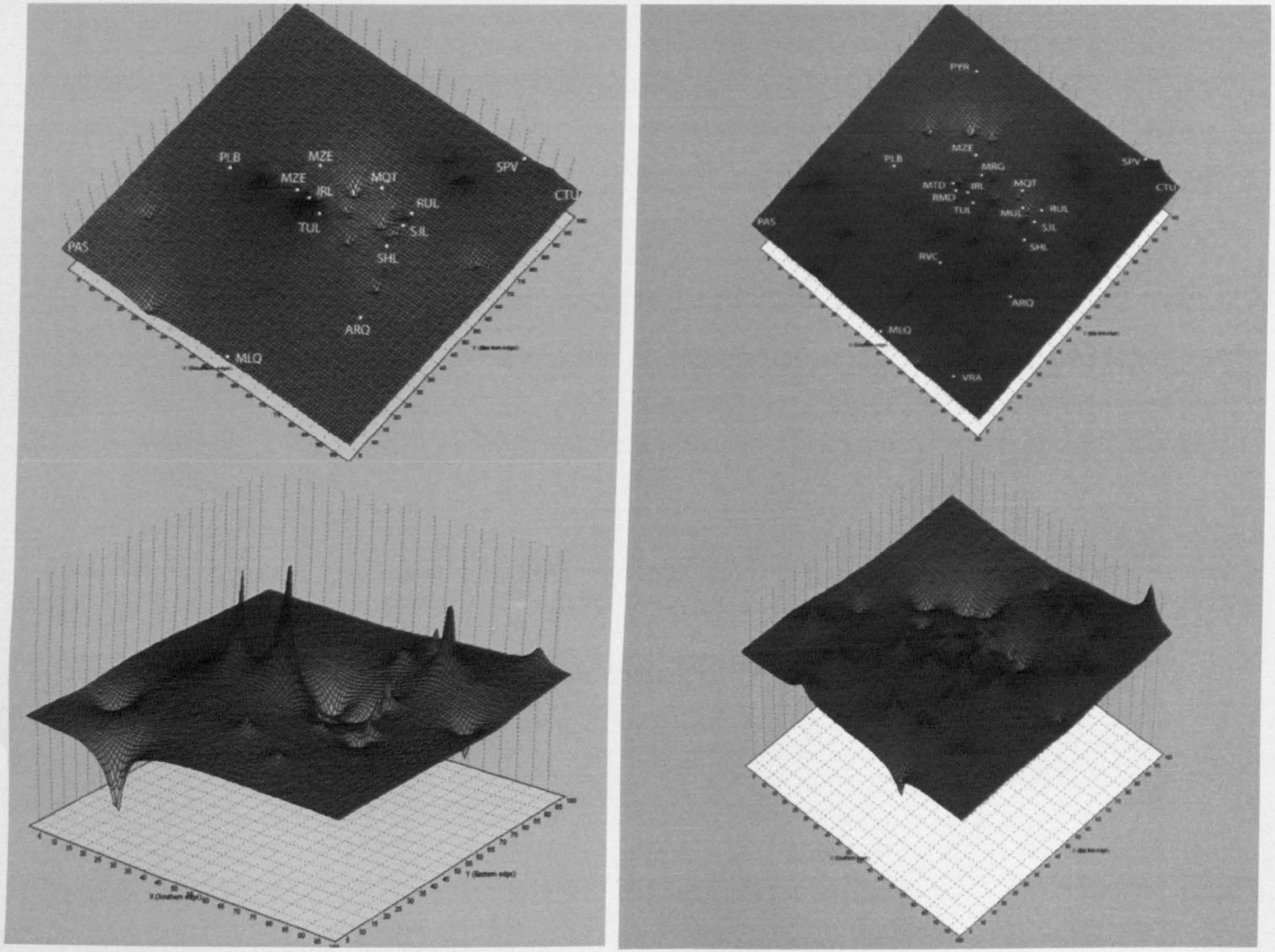


Figure 4.4
Visualizing the interpolated Genetic Landscape Shape (AIS) of *P. ariasi* in southwest France at:
(left) combining nuclear genotypes;
(right) cyt b DNA sequences.

Legend
The X/Y-axis represent geographical coordinates of the *P. ariasi* sample area on which population codes are plotted. The Z-axis represents interpolated genetic distance: elevated and depressed peaks correspond to high and low genetic distances (monochrome scale from white to black), respectively. Interpolation parameters included a 100 x 100 grid and distance weight value = 1. Two corresponding orientations are plotted to show the location of populations (top) and the magnitude of genetic distance (bottom).

4.3.7 Modelling longer-term gene flow using Φ_{ST} for single loci

The presence of genetic divergence caused by drift in isolated populations was assessed by pairwise Φ_{ST} estimates for each locus independently. Levels of differentiation were low in EF-1α (< 0.12966) and moderate to very great in all other loci (range -0.03473 to 0.22661) (Appendix 4.5). Cyt b showed the highest Φ_{ST} values (0.06357 to 0.36248) and, therefore, was the most informative marker for detecting population differentiation. Inter-population pairwise Φ_{ST} estimates did not statistically support differentiation according to *a priori* sub-regions. Alternative explanatory causes for significant results were hypothesized as: (i) environmental, between fragmented forest or outlier populations with others; or (ii) historical ancestry, with only some populations having cyt b haplotypes of mixed ancestry, i.e. from two or more haplogroups (Figure 4.1).

Population pairwise comparisons with any marker categorized as having very great (Φ_{ST} > 0.25) differentiation included: PAS$^{\#\dagger}$ with PLB*, PYR*, MZE*, MQT*, MLQ$^{\#\dagger}$, RUL, MUL, and the SMC; RVC† with MZE*, MQT*, MUL and CTU; MUL with ARQ06$^{\#\dagger}$ and IRL07† [$^{\#}$ = fragmented forest; * = outlier populations; † = mixed cyt b ancestry]. Inter-population estimates which had significant ($P < 0.05$) and great (Φ_{ST} > 0.15) differentiation at cyt b (Table 4.8): PAS$^{\#}$ with all populations of the FDM and SMC, and 6 out of 11 West Other populations; ARQ06$^{\#\dagger}$/ARQ08$^{\#\dagger}$ with 4 out of 5 FDM, SMC, and MLQ$^{\#\dagger}$, MZE*, PYR*, PLB*; IRL07† and RVC† with SMC, PLB†, MZE* and MLQ$^{\#\dagger}$. For nuclear loci, significant and "great" differentiation was only found for APY between MTD and ARQ06$^{\#\dagger}$ or SHL; and for AAm20 (Table 4.8) between SMC and MZE* or MQT*. After sequential Bonferroni correction 7 out of 9 remaining significant comparisons involved population PAS (Table 4.8), supporting its significant relatedness, revealing a lack of migrant exchange from the global mean for combined nuclear loci.

Gene flow was modelled across the study region by testing for inter-population dependence of genetic distance with geographical proximity. At cyt b, consistent with the E Pyrenees (N = 6) result of Chapter 2, the 23 populations characterized globally supported IBD by a Mantel test fitting $\Phi_{ST}/(1-\Phi_{ST})$ to (ln) geographical distance ($P > 0.05$) (Table 4.9). However, the association was shallow: only 11% of the genetic variation was associated with geographical distance with extensive variance for both within and between sub-region comparisons. This variance was equal to, or greater than, comparisons between populations north *vs.* south of the Carcassonne corridor,

Table 4.8 Population pairwise Φ_{ST} values (ARLEQUIN v3.11). Significance after 1,000 permutations: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Bold, significant after sequential Bonferroni correction (α 0.05). Only the two most informative markers are given: cyt b below and AAm20 above the diagonal. Other nuclear loci are given in Appendix Table 4.5.

	PAS	PLB	PYR	MZE	MRG	MTD	RMD	IRL07	IRL08	TUL	RVC	MLQ	MQT	MUL	RUL06	RUL08	SJL	SHL	ARQ06	ARQ08	VRA	CTU	SPV
PAS	-	-0.02065	-	0.11434*	-	0.01663	-	-0.00589	0.02278	-0.01652	-	-0.0049	0.0823*	-	-0.02085	-0.03127	0.00964	-0.02029	-0.0173	-0.01901	-	-0.01289	0.01058
PLB	**0.31089****	-	-	0.10229*	-	0.01222	-	-0.00807	0.01587	-0.01911	-	-0.00725	0.07303*	-	-0.02163	-0.03126	0.00385	-0.02027	-0.01796	-0.02124	-	-0.01102	0.01282
PYR	0.30188**	-0.0286	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MZE	**0.32909****	0.00013	0.00819	-	-	0.01604	-	0.03939	0.00732	0.0728	-	0.04492	-0.0211	-	0.07343	0.10026	0.02014	0 09581*	0.078	0.08392*	-	0.17371**	0.22661***
MRG	0.22799**	0.01417	0.00484	0.03648	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MTD	0.1889**	0.03133	0.02218	0.05329	-0.02523	-	-	-0.01512	-0.02005	-0.0033	-	-0.01656	-0.00237	-	-0.00717	-0.00377	-0.02021	0.00577	-0.00348	0.00191	-	0.05766	0.10486*
RMD	0.11325*	0.09381*	0.0989	0.13563*	0.00782	-0.01302	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
IRL07	0.01757	0.20658**	0.20069	0.24002**	0.10741*	0.06413	0.00828	-	-0.01206	-0.01578	-	-0.01985	0.01863	-	-0.02318	-0.02302	-0.01672	-0.012	-0.01644	-0.01362	-	0.02154	0.0585*
IRL08	0.14257**	0.07031	0.0636	0.09257	-0.00537	-0.02905	-0.02915	0.01691	-	-0.00058	-	-0.01308	-0.00736	-	-0.00255	0 00176	-0.021	0.01146	0.00161	0.00496	-	0.06455*	0.11149**
TUL	0.11329*	0.09031*	0.08514	0.11882*	0.00919	-0.01201	-0.04427	0.00069	-0.03043	-	-	-0.01618	0.04627	-	-0.02372	-0.03031	-0 00937	-0 01934	-0.02029	-0.02135	-	0.00226	0.03352
RVC	-0.0053	0.23412*	0.24328	0.29676**	0.11464	0.06764	0.00102	-0.06367	0.01502	-0.00802	-	-	-	-	-	-	-	-	-	-	-	-	-
MLQ	**0.30727****	0.03488	0.02155	0.05176	0.02689	0.0371	0.05257	0.20398**	0.06254	0.06144*	0.23761**	-	0.02031	-	-0.02034	-0.02252	-0.01799	-0.01186	-0.01699	-0.01358	-	0.02576	0.06629
MQT	**0.32754****	0.01654	-0.00094	0.03475	0.03428	0.04853	0.11299*	0.23617**	0.0885	0.10551*	0.28468**	0.02409	-	-	0.04497	0.06308	0.00191	0.06511	0.04897	0.05613	-	0.13907**	0.19362***
MUL	**0.35722****	0.01736*	-0.0038	0.05115	0.0452	0.07085	0.17326**	0.28494****	0.12375*	0.14724*	0.36248**	0.06622*	0 04713**	-	-	-	-	-	-	-	-	-	-
RUL06	0 2576**	-0.0088	-0.02783	0.00833	-0.02791	-0.02116	0.03476	0.14578**	0.01061	0.0348	0.17287	0.00812	0.00554	0.02129	-	-0.03473	-0.01152	-0.02361	-0.02453	-0.02377	-	-0.00241	0.02788
RUL08	0 27219**	0.00064	0.01	0.04304*	-0.00442	0.01048	0.04513	0.16068	0.04031	0.04483	0.18142	-0 01139	0.01288	0.05426*	-0.02107	-	-0.01016	-0.03226	-0.03122	-0.03158	-	-0 01724	0.01017
SJL	0.22043**	0.02586	0.01777	0.04964	-0.01676	-0.02304	-0.00462	0.10082	-0.00825	0.00143	0.113	0.00699	0.03356	0.07096**	-0.0225	-0.02555	-	0.0002	-0.00764	-0.00511	-	0.04648	0.09042*
SHL	0.13214*	0.12947**	0.13112*	0.16624**	0.05553	0.03385	-0.03446	0.03799	0.01291	-0.01968	0.03457	0.06421*	0.14026**	0.20201**	0.07809	0.06708	0.02786	-	-0.02026	-0.02059	-	-0.00626	0.02185
ARQ06	0.01265	0.19998**	0.18834*	0.21975**	0.11931*	0.08574*	0.01575	-0.02497	0.04546	0.01652	-0.04374	0.18436****	0.21533*	**0.24988****	0.14684*	0.15352*	0.10795*	0.02885	-	-0.02028	-	0.00236	0.03488
ARQ08	0.01029	0.21021**	0.20365*	0.24353**	0.11632*	0.07621	0.00184	-0.03679	0.03164	0.00131	-0.05399	0.19467****	0.23588****	0.28648****	0.15164*	0.15748*	0.10331*	0.01724	-0.0334	-	-	-0.00376	0.02465
VRA	0.19191**	0.02708	0.01993	0.05246	-0.0266	-0.03562	-0.02642	0.06981	-0.02757	-0.01916	0.0759	0.00425	0.03768	0.07614*	-0.02764	-0.00617	-0.03433	0.00822	0.08175*	0.07211	-	-	-
CTU	**0.32535****	0 02587*	0.01505	0.04334*	0.04582*	0.05902	0.11126**	0.22726**	0.09188	0.10741*	0 26129**	0.04842	0.03186	0.06522**	0.01828	-0.00904	0.04738	0.13788**	**0.21469****	0.22843****	0.04597	-	-0.01271
SPV	**0.29885****	0 03592*	0.02801	0 05976*	0 03123	0 03744	0 09072*	0 19609**	0.06656*	0.08749*	0 22516**	0.06342*	0.04688*	0.08027****	0 01524	0 01	0 03752	0 12824**	0.19208**	0.20167**	0 03321	-0.01848	-

Table 4.9 Testing the support for IBD using a Mantel test to fit genetic distance $[\Phi_{ST}/(1-\Phi_{ST})]$ to geographical distance or ln distance. Populations from different landscape categories were systematically excluded to infer causes of population structure (see text).

	Fitting $\Phi_{ST}/(1-\Phi_{ST})$ to distance				Fitting $\Phi_{ST}/(1-\Phi_{ST})$ to ln distance		
Populations tested for IBD; locus	a	b	*P* (c>o)	R^2	a	b	*P* (c>o)
23 (total) populations; cyt b	0.0284185	0.0023873	0.004*	0.1105	-0.0520848	0.0483059	0.004*
18 continuous forest populations; cyt b	0.0101163	0.0024562	<0.001*	0.171	-0.0639136	0.04668168	<0.001
15 continuous plateau populations; cyt b	0.0022513	0.00356603	<0.001*	0.2019	-0.0524322	0.04837520	<0.001*
12 continuous plateau populations; AAm20	-0.0290205	0.00128540	0.006*	0.4244	-0.0373037	0.01375193	0.006*
12 continuous plateau populations; AAm24	0.0313414	-0.00033858	0.732	0.0137	0.0309601	-0.00271968	0.732
12 continuous plateau populations; APY	0.0213108	-0.00016617	0.472	0.0016	0.0218515	-0.00159109	0.474
12 continuous plateau populations; EF-1α	0.0000544	0.00004171	0.519	0.0041	0.0000382	0.00035735	0.513
Continuous West Other & FDM - straightline[#]; cyt b	-0.0103735	0.00552963	0 .009	0.139	-0.033422	0.04174196	0.006
Continuous West Other & FDM - forest line[#]; cyt b	-0.0017682	0.00392172	0.009	0.1251	-0.0285551	0.0371984	0.004

which matches the signal of genetic discontinuity observed in Chapter 2 (Figure 4.5a).

This result suggested that gene flow between populations was similarly restricted within the NE Pyrenees to that across the corridor. Therefore, IBD was used as an explanatory tool to help investigate the underlying cause(s) of this (Guillot *et al.*, 2009). The five fragmented forest populations were found to have levels of genetic differentiation similar to those attributed to larger-scale geographical barriers, as their removal supported IBD ($P < 0.001$), reduced the variance, and slightly increased the level of correlation (by 6%, $R^2 = 0.171$) (Figure 4.5b; Table 4.9). This similarity can be explained by fragmented forest populations containing high frequencies of cyt b haplogroup A (88-100%), as observed in the SMC north of the Carcassonne corridor.

The removal of climatic/geographical outlier populations from the NE Pyrenees (PAS, MLQ, VRA) also reduced the variance, with only a small increase in the correlation supporting IBD (Figure 4.5c, where $R^2 = 0.2019$ and $P < 0.001$). The remaining great differentiation between populations in the NE Pyrenees (< 30 km, $\Phi_{ST} > 0.15$) mostly involved comparisons with population MUL. Again, this population was fixed for cyt b haplogroup A. This sample was not obviously isolated from forest, and so its anomalous position resulted either from a sampling artefact or a bottle-neck event involving a different population process. After excluding MUL the association between genetic and geographic distance matrices was over two-fold greater across the corridor ($R^2 = 0.4216$) as that within the NE Pyrenees (0.1903). A marginal test (DISTLM) found a significant relationship between genetic distance and categorical data that partitioned pairwise comparisons across the corridor apart from those within the NE Pyrenees or SMC (23% variation explained, $P = 0.001$). However, a conditional test that eliminated the effects of geographical distance by taking it as a covariate, showed these categories to be no longer correlated to genetic distance (1%, $P = 0.36$). Furthermore, the two genetic/geographical distance regression coefficients were not significantly different ($t = 1.342$, df = 88, $P > 0.05$) (Figure 4.5c). Modelling this data set for each nuclear locus (population N = 12), showed only AAm20 supporting IBD ($P = 0.001$); the other loci were not informative, as would be expected from their low pairwise genetic distances (Table 4.9). For all IBD Mantel tests, similar results were gained when using ln geographical distance (two dimensional dispersal model).

Models of gene flow were not affected by measuring geographical distance by straight-line distances compared with following the lower boundary of broadleaf forest. Finer-scale analyses used continuous forest populations only to investigate differences between the sub-regions West Other and FDM (including SHL). Both models supported

which matches the signal of genetic discontinuity observed in Chapter 2 (Figure 4.5a).

This result suggested that gene flow between populations was similarly restricted within the NE Pyrenees to that across the corridor. Therefore, IBD was used as an explanatory tool to help investigate the underlying cause(s) of this (Guillot *et al.*, 2009). The five fragmented forest populations were found to have levels of genetic differentiation similar to those attributed to larger-scale geographical barriers, as their removal supported IBD ($P < 0.001$), reduced the variance, and slightly increased the level of correlation (by 6%, $R^2 = 0.171$) (Figure 4.5b; Table 4.9). This similarity can be explained by fragmented forest populations containing high frequencies of cyt b haplogroup A (88-100%), as observed in the SMC north of the Carcassonne corridor.

The removal of climatic/geographical outlier populations from the NE Pyrenees (PAS, MLQ, VRA) also reduced the variance, with only a small increase in the correlation supporting IBD (Figure 4.5c, where $R^2 = 0.2019$ and $P < 0.001$). The remaining great differentiation between populations in the NE Pyrenees (< 30 km, $\Phi_{ST} > 0.15$) mostly involved comparisons with population MUL. Again, this population was fixed for cyt b haplogroup A. This sample was not obviously isolated from forest, and so its anomalous position resulted either from a sampling artefact or a bottle-neck event involving a different population process. After excluding MUL the association between genetic and geographic distance matrices was over two-fold greater across the corridor ($R^2 = 0.4216$) as that within the NE Pyrenees (0.1903). A marginal test (DISTLM) found a significant relationship between genetic distance and categorical data that partitioned pairwise comparisons across the corridor apart from those within the NE Pyrenees or SMC (23% variation explained, $P = 0.001$). However, a conditional test that eliminated the effects of geographical distance by taking it as a covariate, showed these categories to be no longer correlated to genetic distance (1%, $P = 0.36$). Furthermore, the two genetic/geographical distance regression coefficients were not significantly different ($t = 1.342$, df = 88, $P > 0.05$) (Figure 4.5c). Modelling this data set for each nuclear locus (population N = 12), showed only AAm20 supporting IBD ($P = 0.001$); the other loci were not informative, as would be expected from their low pairwise genetic distances (Table 4.9). For all IBD Mantel tests, similar results were gained when using ln geographical distance (two dimensional dispersal model).

Models of gene flow were not affected by measuring geographical distance by straight-line distances compared with following the lower boundary of broadleaf forest. Finer-scale analyses used continuous forest populations only to investigate differences between the sub-regions West Other and FDM (including SHL). Both models supported

Figure 4.5 Plots and regression of genetic distance [Y-axis = $\Phi_{ST}/(1-\Phi_{ST})$] on straight-line geographical distance (X-axis = km) at cyt b: (a) 23 populations, (b) excluding fragmented forest populations, (c) further exclusion of 3 outlier populations except SMC - the two regression lines represent comparisons within (black) sub-regions (excluding bottle-necked MUL), or between (red) north with south of the Carcassonne corridor; there was no significant difference between these regression coefficients.

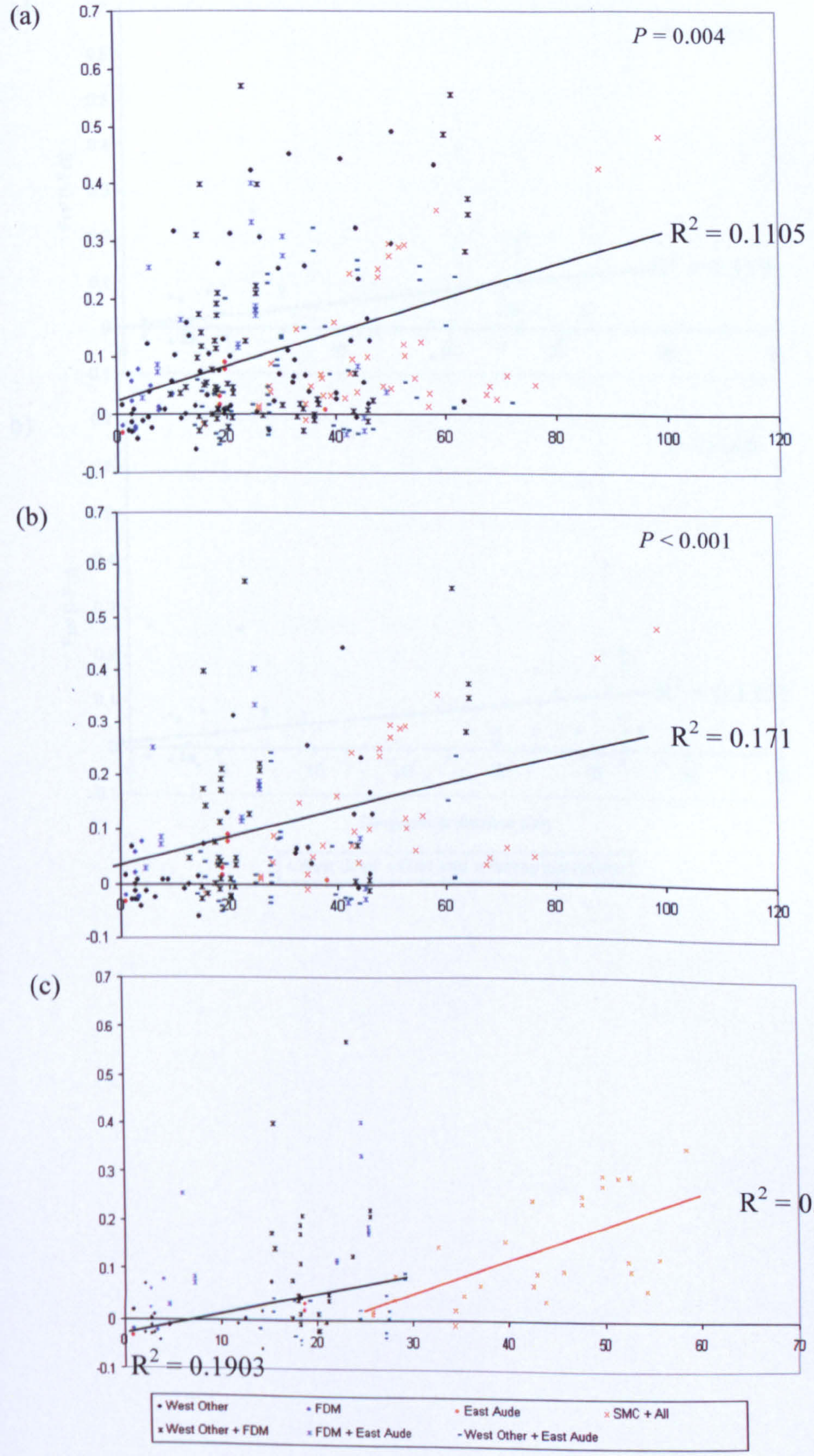
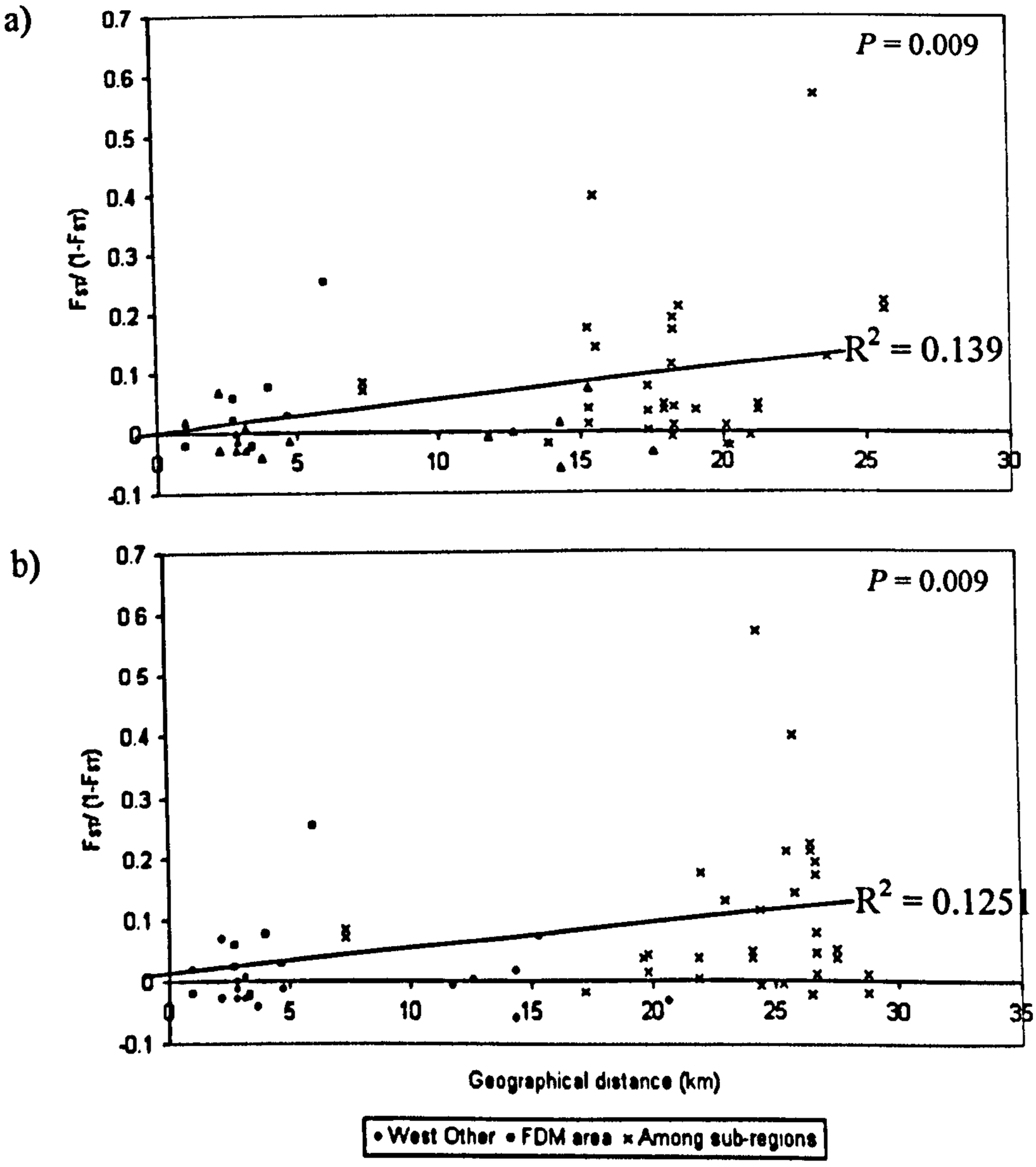


Figure 4.6 Testing for fine-scale IBD at cyt b for continuous forest populations by comparing two sub-regions in the northeast Pyrenees; West Other and FDM. Plots and regressions of genetic distance [$\Phi_{ST}/(1-\Phi_{ST})$] on (a) straight-line geographical distance, and (b) an alternative dispersal route following the lower boundary of continuous broadleaf forest.



IBD ($P < 0.05$), with similar shallow regression slopes (straight-line route $R^2 = 0.139$, forest route $R^2 = 0.1251$) (Figure 4.6; Table 4.9).

4.3.8 Quantifying the short-term spatial scale of genetic connectivity between individual *P. ariasi* using combined nuclear genotype data

Spatial autocorrelation, a combined nuclear genotype approach for evaluating short-term population processes, was used to investigate the limit of non-random gene flow between individual *P. ariasi* by quantifying their scale of spatial (landscape) connectivity (dispersal). Although spatial autocorrelation reveals the scale and pattern of correlation, it does not identify specific location of discontinuities. In principle, the correlation coefficient (r) between geographical and genetic distances reflects the global properties over a sample area, and therefore, only makes sense when the study region is homogeneous e.g. in terms of gene flow patterns (Guillot *et al.*, 2009). Following this approach, a limited data set was used, including only NE Pyrenees populations from continuous forest (excluding the 3 geographically distant outliers on the northern Pyrenees slope (PAS, MLQ, VRA) and bottle-necked MUL).

The pattern in the autocorrelogram was consistent with the signature expected under IBD with spatial genetic structure - initial high positive autocorrelation which declines through zero followed by subsequent oscillation around zero is typical of a restricted gene flow scenario (Peakall *et al.*, 2003). Significant ($P = 0.019$) positive spatial genetic structure (connectivity) was supported for geographically close individuals at distance class 2 km (i.e. within samples only), with a second periodic increase at 14 m ($P = 0.02$) and no autocorrelation at distance classes above this (Figure 4.7). As positive spatial structure was supported, the distance among *P. ariasi* where non-random genetic correlations (gene flow) are expected to cease was estimated at 3.805 km. Distances below this threshold unite populations that share a higher proportion of genes, whereas populations more distant are genetically independent.

4.3.9 Population sub-structure supported *a priori* population sub-division

As would be expected from the Φ_{ST} values, which were modelled to support IBD, hierarchical AMOVA based on all 23 populations characterized at cyt b did not support geographical sub-divisions according to *a priori* groupings (Table 4.10): (1) testing sub-structure across the test region; (2) testing the Carcassonne corridor as a genetic barrier (SMC *vs.* others); (3) testing whether the agriculture belt isolates FDM from West Other; and, (4) testing the Aude river as a genetic barrier (West Other *vs.* East Aude). Tests supported homogeneity among sub-regions ($P > 0.05$), and heterogeneity both among populations within regions and within populations ($P < 0.001$).

This again prompted the exclusion of fragmented forest populations (N = 5) and three outlier populations from the northern slopes of the Pyrenees (PAS, MLQ, VRA), which were either bottle-necked or distant from the main population clusters sampled. Cyt b (population N = 15) was shown to be the most informative population marker, with three of the four hypotheses tested supporting among region sub-division ($P < 0.05$), where two tests were accompanied by within region homogeneity ($P > 0.05$) for all sub-regions (hypothesis 1) and the isolation of the FDM from West Other (hypothesis 3). AAm20 (population N = 12) was the only individual nuclear locus to support population sub-structure, among all sub-regions (hypothesis 1) and the sub-division either side of the Carcassonne corridor (hypothesis 2). No locus supported the hypothesis that the Aude river was a barrier to *P. ariasi* (hypothesis 4) (Table 4.10).

Taking into account the preceding results in concert, support for population sub-division corresponding to the three independent forest regions was tested: SMC, FDM (+ SHL, as the Aude is an unlikely barrier), and West Other (Map Figure 4.1). To avoid confounding these putatively distinct regions with genetic variation explained by other landscape factors, populations excluded were those from fragmented forest (PLB, PYR, MZE, MRG, MQT), one with significant internal relatedness (MQT), one bottle-necked (MUL), and those geographically distant from the main population clusters (PAS, PLB, MLQ, VRA, ARQ). This left the following populations for analysis: MTD, IRL07/08, TUL, RUL06/08, SJL, SHL, CTU and SPV at all loci, and additionally RMD and RVC at cyt b. Supporting the hypothesis of unsuitable *P. ariasi* habitats as barriers to gene flow (section introduction): cyt b supported sub-division of the test region as a whole ($P < 0.01$), with homogeneity within sub-regions ($P > 0.05$) (hypothesis 5); both the SMC (north of the Carcassonne corridor; cyt b and nuclear loci hypothesis 7) and FDM (cyt b

hypothesis 8) significantly ($P < 0.05$) differentiated from the main forested region of the NE Pyrenees, accompanied by homogeneity within sub-regions ($P > 0.05$). However, the SMC and FDM was not a supported sub-division (among regions $P > 0.05$; hypothesis 6).

Using that same regional definitions, at cyt b, to test whether the supported regional effects in AMOVA were generated by barriers to gene flow, the effect of geographical distance on genetic distance (IBD) was eliminated by the application of a dbRDA approach. Marginal tests showed all comparisons to follow IBD and a significant correlation between genetic distance and categorical data for within sub-region or between sub-region pairwise comparisons ($P < 0.01$). However, dbRDA did not support the AMOVA population sub-structure (three sub-regions; SMC or FDM vs. Main NE Pyrenees) because, when geographical distance was taken as a covariate in a multiple regression analysis, categorization into within or between comparisons did not leave a correlation with genetic distance ($P > 0.09$).

Table 4.10 Hierarchical AMOVA to test the support for 8 *a priori* hypothesized population sub-divisions. Categories of populations included varied († Excludes PAS, MLQ, VRA; see text). F. Indices and their level of significance from a null of panmixia given: $*P < 0.05$; $**P < 0.01$; $***P < 0.001$.

Populations included	All pops	Continuous forest populations, excluding Pyrenean slopes†				
Locus	Cyt b	Cyt b	EF-1α	APY	AAm20	AAm24
Sub-division hypothesis tested	F. Indices	F. Indices	F. Indices	F. Indices	F. Indices	F . Indices
1. SMC vs. FDM vs. *West Other* vs. *East Aude*	N = 23	N = 15	N = 12	N = 12	N = 12	N = 12
Among regions	0.04923	0.09569***	0.00076	-0.00644	0.02451**	0.00996
Within regions	0.09530***	-0.00220	-0.00004	0.02373*	-0.01495	0.01291
Within pops	0.13984***	0.09370***	0.00072	0.01745*	0.00993	0.02274*
2. SMC vs. *others*						
Among regions	0.04809	0.07584*	0.00036	-0.01009	0.05261*	-0.00520
Within regions	0.11814***	0.05383**	0.00045	0.02188*	-0.01299	0.02243*
Within pops	0.16054***	0.12559***	0.00081	0.01201*	0.04030	0.01735*
3. FDM vs. *West Other*						
Among regions	0.05389	0.08454**	-0.00090	-0.00837	-0.00213	-0.00482
Within regions	0.12434***	-0.00094	0.00190	0.02481*	-0.01270	0.01872
Within pops	0.17153***	0.08368*	0.00100	0.01665*	-0.01486	0.01399
4. West Other vs. *East Aude*						
Among regions	-0.00450	0.01903	0.00088	0.00382	0.00805	0.02749
Within regions	0.10127***	-0.00246	-0.00703	0.02922*	-0.01487	0.01008
Within pops	0.09723***	0.01662	-0.00614	0.03293**	-0.00670	0.03730*

Sub-division hypothesis tested	Cyt b	Combined nuclear genotypes
5. SMC vs. *FDM* vs. *Main NE Pyrenees*	N = 12	N = 10
Among regions	0.06089**	0.00449
Within regions	0.00717	0.01011*
Within pops	0.06762**	0.01455**
6. SMC vs. *FDM*	N = 6	N = 6
Among regions	0.04919	0.00051
Within regions	0.02515	0.01154*
Within pops	0.07311	0.01205*
7. SMC vs. *Main NE Pyrenees*	N = 8	N = 6
Among regions	0.10844*	0.01715*
Within regions	-0.00073	0.00499
Within pops	0.10779*	0.02205*
8. FDM vs. *Main NE Pyrenees*	N = 10	N = 8
Among regions	0.02929*	-0.00159
Within regions	0.00626	0.01308**
Within pops	0.03536	0.01150**

4.3.10 Bayesian cluster method fails to identify current population sub-division in the study region

The Bayesian assignment test, which combined the genotypes for all loci, did not statistically support population clusters in the study region: the area was considered to be currently connected as a single genetic deme. The standard admixture model was used, which only uses genetic information to cluster populations. It failed to converge even when burn-in and subsequent MCMC runs were substantially increased to 500,000 and 5 million, respectively. To assess whether genuine current population structure occurred in the study region, the Hubisz *et al.* (2009) ancestry model was implemented, which modifies the standard admixture model prior to including location information, and thus can be more sensitive to population structure when the signal is too weak to be detected by the standard model. This model led to virtually the same result for 1-5 K clusters: range of mean lnP(D) between -4073.1 to -4021.0 and ΔK 1.50-11.78 (Figure 4.8). The plot of lnP(D) showed no clear peak among these clusters. The magnitude change of lnP(D) relative to the standard deviation, ΔK, showed a maximum value at K = 2, for which the mean membership plot is presented (averaged over the 10 repeats in CLUMPP; Figure 4.9a). The plot shows the membership of individuals was in fact to a single cluster; no support for population sub-division. In this plot each individual is represented by a single bi-coloured vertical line, the relative lengths of which are proportional to membership of one of the two inferred clusters. From this it can be seen that each individual predominantly belonged to a single K cluster/population (yellow) (Figure 4.9a). Evanno *et al.* (2005) noted that the highest lnP(D) does not always indicate the most likely K, and their *ad hoc* method is based on prior K values and therefore only valid when $\Delta K > 3$. It is therefore not incorrect to conclude the presence of a single genetic deme in the sample area. The same result, rejecting population structure, was similarly found when including only those populations from the continuous forest category (population N = 12), showing this clustering method was not improved by the elimination of potentially bottle-necked populations (Figure 4.9b).

4.4 Discussion

There is little known about the population genetic structure of *Phlebotomus ariasi*. This chapter follows on from the low resolution study of Chapter 2, by assessing the fine-scale spatial structure of this vector in a 70 x 70 km study area which was composed of a landscape of fragmented suitable habitats. The effect of landscape structure was quantified with the aim of revealing areas of restricted *P. ariasi* dispersal and the genetic consequences of this restriction, over both the recent past and contemporary time-scales. The results provided some evidence of gene flow restricted by: (1) moderate geographical distances (ca. 10-50 km) between relatively large forest fragments (>/= ca. 96 sq km); and (2) 'micro-isolation' in habitats greater than 400-700 m from the nearest forest fragment. By excluding category (2) and geographically distant and/or coincident bioclimate outliers, AMOVA supported population sub-structure among three forested regions. However, Bayesian assignment tests supported an uninformative shallow likelihood topology and the presence of a single genetic deme. Overall, the results of this investigation could not provide conclusive evidence supporting contemporary genetic sub-structuring or genetic impoverishment of *P. ariasi* sampled from a fragmented landscape, but the results presented certainly warrant the development of prospective studies to do so.

4.4.1 Can current markers for *P. ariasi* detect fine-scale population sub-structure across a mosaic landscape?

The distribution of molecular genetic variation is partitioned both in time and space and, therefore, inferences made about the biology of individuals through to species must consider the level of molecular change in context. Molecular characters reveal information at various time-scales (Sunnucks, 2000): at the shortest time-scale genotypes assess within population processes or current migration; longer time-scales test for between population processes or population history, e.g. expansions or contractions, using allele frequency data; and at the longest time-scale, information on phylogeography or phylogenetic speciation are understood though DNA sequence evolution. The current fine-scale population study followed on from the earlier investigation of the low resolution historical population biology of *P. ariasi*, and demonstrated that knowledge of the phylogeography and past demographic effects on genetic variation is imperative when using markers and tests sensitive to these past population processes.

Hierarchical AMOVA at cyt b supported the *a priori* hypothesis of population sub-division across the test region and the differentiation of two forest-associated sub-regions, the SMC (north of the Carcassonne corridor) and the FDM, from the western forest that extends along the northern Pyrenean foothills, but not from each other. Additionally a single nuclear locus supported the isolation of the FDM from the western forest region, and a combined analysis of nuclear genotypes supported the north versus main western forest sub-divisions (Table 4.10). The FDM is a dense broadleaf forest patch ca. 12 x 8 km, isolated to the north and southwest by ~10 km of cultivated crop or other non-forest land covers, and in the east by the Aude river and ~1.5 km non-forest transport route. The AMOVA result supported its isolation (with or without population SHL) from southwest forest populations, but not from those to the north. This result could be explained if the founding events establishing these two populations from the parent populations on the Pyrenean slopes were similar.

Assessment of the genetic connectivity of individual *P. ariasi* estimated the limit of positive local genetic structure at ca. 4 km, which is consistent with the support given by AMOVA for sub-regions which are separated by 5-10 km of unsuitable land covers (belts of agriculture or urbanisation). However, at cyt b a dbRDA analysis revealed that IBD could actually generate the AMOVA results. Peakall *et al.* (2003) in their bush rat study showed moderate to extensive gene flow, which was over considerably larger distances than the scale of per generational dispersal. This disparity was a consequence of gene flow measured by evolutionary estimators such as F_{ST} that actually reflect past interconnection. It is plausible that the foothills of the NE Pyrenees and the Massif Central were in the recent past connected by continuous forest that formed no barriers to *P. ariasi* dispersal, explaining the IBD results presented. For cyt b, AMOVA and IBD modelling used DNA sequence information and, therefore, the statistical support of sub-division given by this marker could actually be an artefact of the historical presence and distribution of multiple cyt b haplogroups. Haplogroup A was the most frequent of cyt b haplogroups and, consistent with this, it was found to be at, or near to, fixation in the isolated regions. FDM was found to be mitochondrially indistinguishable from the populations north of the Carcassonne corridor, SMC, but both were mitochondrially and statistically differentiated from the putative multi-haplogroup parent population of the main southern forest.

Longer-term interconnection by gene flow and strong local contemporary genetic structure are not mutually exclusive (Peakall *et al.*, 2003). In the spatial autocorrelogram, the oscillating pattern of *r* (correlation coefficient) in low geographical distance classes mirrors the periodicity patterns expected with contemporary spatial patchiness (Elmer *et al.*, 2007). Yet the Bayesian clustering assignment analysis, which similarly also used genotype data of the individual to infer contemporary population structure, failed to converge without location information as priors, and a model that did include this information still failed to recover more than a single genetic deme. However, this STRUCTURE result could stem either from the lack of suitability of the data inputted or from the real absence of contemporary population structure.

In their literature search (*Molecular Ecology* publications in 2001), Berry *et al.* (2004) showed that on average eight microsatellites are used to study animal population genetics, but noted that the rate of improvement of assignment of additional markers is low if levels of genetic differentiation are low. Therefore, the five markers used for *P. ariasi* may have been too few, especially as they were not rapidly evolving microsatellites. Alternatively, as proposed by Mank and Avise (2004), the "handful" of markers used could have been sufficient for meaningful signals in various frequency-based population assessments (i.e. AMOVA), but not in Bayesian searches because of uninformatively shallow likelihood topologies. Bayesian assignment tests do not perform well for weakly differentiated populations (Mank and Avise, 2004) or when there is IBD (Pritchard *et al*, 2009), so they may not be suitable for this study. Lastly, marker polymorphism was low and therefore could have lacked power to detect existing local population structure. Φ_{ST} values for nuclear markers showed only 1% of inter-population comparisons to have greater than 'moderate' genetic differentiation (> 0.15). However, no microsatellite studies of *P. ariasi* are available for comparison. Polymorphism levels are known for five microsatellites that supported regional population differentiation of *P. perniciosus* within Spain (Aransay *et al.*, 2003), and this sandfly is sympatric with *P. ariasi* and in the same subgenus. Comparison of *P. ariasi* nuclear loci with *P. pernicious* microsatellites showed similar genetic information content. Numbers of *P. ariasi* alleles ranged from six to 23 at each locus, with at least two alleles predominating, whereas *P. perniciosus* had four to nine alleles with one to two of these predominating. Ranges for the expected heterozygosity were: *P. ariasi* loci 0.214-0.764 and, *P. perniciosus* loci 0.052-0.683. Both species showed 'little' within-

region differentiation, with F_{ST} values for *P. perniciosus* being < 0.0414. Summarizing, the absence of local structure for *P. ariasi* is inconclusive, and should be resolved by increasing the number of loci characterized, not only to increase the performance of assignment tests but also to heed an oft-quoted note of caution that population inferences should not be based on a single locus.

4.4.2 Limited genetic impoverishment may be explained by sampling or the properties of the physical landscape

On an evolutionary scale restricted gene flow can at one extreme generate local genetic structure and at the other cause inbreeding and genetic isolation. Understanding the relative contributions of these factors can inform us of the likelihood of a species' persistence: reduced population heterozygosity is associated with reduced population reproductive fitness, inbreeding depression increases extinction risk, and loss of genetic diversity reduces the ability of populations to evolve to cope with environmental change (Spielman *et al.*, 2004). Genetic impoverishment of *P. ariasi* was associated with fragmented land covers, with nucleotide diversity declining in populations more distant from continuous forest. However, haplotype diversity or allelic richness were not reduced, and only a single fragmented forest population showed significant inbreeding. Sampling strategy might explain this result. Only five populations were categorized as fragmented and two of these populations were not characterized at all nuclear loci. Sampling more populations from this category might reveal a negative association between increased forest fragmentation and all genetic diversity statistics and relatedness estimates. However, this poses a practical problem, because adequate sample sizes (15-30 individuals) were only obtained from dwellings or road-side walls in forested areas, and these are not always available. Additionally, categorization as a fragmented or continuous forest population might be inaccurate in light of the results of this study. Categories were delimited based on previous knowledge of direct dispersal distances for this species (Killick-Kendrick *et al.*, 1984), an experimental approach that can be highly inaccurate.

Genetic diversity and relatedness were estimated by several statistics, but none supported consistent reductions associated with *a priori* sub-regions. Lack of diversity and differentiation associated with regional landscape, together with the shallow IBD observed, would be expected of a single genetic population, where gene flow maintains homogeneity. However, I cannot rule out three alternative hypotheses to explain the

limited genetic population differentiation of *P. ariasi*. Firstly, lack of power of genetic markers (previously discussed). Second, fragmentation in the study area may have been recent, and time has not been sufficient for genetic differences to accumulate under a model of restricted gene flow/dispersal. Conversely, it is reasonable to infer that even very recent landscape fragmentation should have restricted dispersal because sexual recombination reorganises genotypes in a single generation and *P. ariasi* breeds at 1 generation per year at temperatures equivalent to those in the study region (Ready and Crosset, 1980). Martínez *et al.* (2007) showed broadleaf forest structure to be evolving in the study region; comparing Landsat data from 1984 to 1992 to 2003 the number of patches has increased. Third, and lastly, gene flow/dispersal might not be restricted by the current level of forest fragmentation in the study region. The current results are now considered in relation to this hypothesis. We see that although forest patch number has increased in the study region, the average shape of patches has remained stable and no net decrease in forest cover has occurred over the past 20 years (Martínez *et al.*, 2007). Population isolation occurs only when habitat loss breaks connectivity, and the degree of connectivity is defined by both the properties of the physical landscape and the dispersal ability of individuals through it (Ewers and Didham, 2006). Therefore, it can be concluded that the forest is not disconnected sufficiently to impede dispersal of *P. ariasi*. This result could be another example indicating that the main determinant of population size and viability is the total amount of habitat in a landscape, not the spatial configuration of that habitat below a threshold (Prugh *et al.*, 2008). Studies on *P. perniciosus* have supported gene flow between contiguous populations over distances up to 500 km (Aransay *et al.*, 2003).

4.4.3 The genetic landscape of *P. ariasi* and disease epidemiology

At the northern limit of *Leishmania* endemicity in Europe, *P. ariasi* is both the predominant *Phlebotomus* species and principal disease vector (Ready, 2008) and, therefore, an understanding of its population differentiation is potentially important for planning intervention strategies and modelling risk of disease spread. Modelling of vector-borne disease spread is becoming increasingly popular (Roger and Randolph, 2000), especially when we acknowledge the potential effects of climate and anthropogenic environmental changes (Patz *et al.*, 2000) on their (re-)emergence. This study did not discover many contemporary restrictions on the dispersal of *P. ariasi*. If I accept suitable marker sensitivity and that sub-division is generated by IBD, then this

study suggests that the landscape matrix north of Pyrenean slopes does not prevent the spread of *P. ariasi* and, therefore, of *L. infantum*. This study showed that forest fragmentation should be considered in the context of net forest reduction and not necessarily of increased patch number. Practically therefore, the maintenance of genetic diversity, a source of genetic evolution, and ability to disperse, highlights the potential for vectorial traits to develop and spread in the NE Pyrenees. However, even with unrestricted dispersal potential, the results are consistent with one monopolization effect, that the populations in the Massif Central could block the genetic spread of *P. ariasi* northward. Positive local genetic structure was detected, but spatial autocorrelation revealed that gene flow through this landscape was restricted to ca. 4 km per generation. This indirect estimate of dispersal is twice as much as directly measured by mark-release-recapture studies, and this should be considered when modelling the spread of leishmaniasis.

CHAPTER 5

General discussion

5.1 Introduction

Europe has seen the recent emergence of new diseases or the re-emergence of existing ones, and both canine and human leishmaniasis follow this trend (Vorou *et al.*, 2007; Dujardin *et al.*, 2008). The occurrence and geographical spread of vector-borne diseases is often associated with changes in their epidemiology, often by environmental change modifying the transmission cycle through the provision of favourable new ecological niches for parasite, host and/or vector (Morens *et al.*, 2004; Patz *et al.*, 2000). The leishmaniases are considered as indicator diseases, sensitive to environmental change, and are the subject of investigations to catalogue European environmental conditions that can influence the spatial and temporal distribution and dynamics of disease agents (e.g. EDEN project, EU FP6: www.eden-fp6project.net/). Environmental change is oft-quoted as affecting the distribution of infectious diseases, but the mode of this change is typically not known for leishmaniasis, for which spatial models can be specific to a particular geographical region (Ready, 2008). The focus of this thesis was the sandfly *Phlebotomus ariasi*, whose predominance as the leading-edge vector in the transmission cycle of zoonotic visceral leishmaniasis (ZVL) makes it an important component in modelling the future risk of northwards spread of this disease. This thesis characterized genetic variation in *P. ariasi* at both the phylogenetic and population levels, to investigate the effects of environmental change on the molecular evolution and spatial distribution of this sandfly in southwest France.

5.2 Identification of a single vector species

Vector control remains the primary measure available to prevent much parasite transmission (Lambrechts *et al.*, 2009). Identification of vector species has a place in this control, as many are members of species complexes of morphologically very similar, or indistinguishable, sibling species (Curtis, 1999). Morphologically indistinguishable species are known in *Phlebotomus*, e.g. males of *P. longicuspis* in the Moroccan Rif (Pesson *et al.*, 2004), which may have implications for their role as vectors of *Leishmania* or the landscape epidemiology of this disease. In Tunisia, the females of *P. ariasi* and *P. chadlii* are indistinguishable (Esseghir *et al.*, 2000) or nearly

so (Chamkhi *et al.*, 2006), and female *P. ariasi* from Morocco (same region as characterized in this thesis) have been reported as morphologically atypical (Boussaa *et al.*, 2009). GenBank sequences arising from others' research on this species complex are few, namely AF161194, AF161195 and AF161196 for cytochrome b (cyt b); and AF160803 and AF160804 for elongation factor-1α (EF-1α). This thesis contributed further DNA sequence accessions at these two loci for morphologically-identified *P. ariasi* from Morocco, Portugal, Spain and France. Furthermore, both phylogenetic and parsimony network reconstructions, in addition to population based tests, confirmed the absence of cryptic sibling species of *P. ariasi* characterized across western Europe and Morocco. This result has two implications: a vector control program for *P. ariasi* could be generic in Europe and, directly relevant to the approach of this study, most natural genetic variation can be attributed to neutral evolution, rather than to reproductive barriers.

5.3 Advances in the molecular tools available for *P. ariasi*

Before the current study, few molecular tools were available or optimized to investigate the population differentiation of *P. ariasi*, and none had been applied. In addition to those mentioned previously, named nucleotide sequences in GenBank include cyt b to 5′ NADH1 (Esseghir *et al.*, 1997; 2000), 5.8S ribosomal DNA (Di Muccio *et al.*, 2000) and various salivary peptide cDNAs (Oliveira *et al.*, 2006). The current study was novel in applying two known markers in population genetic analyses of *P. ariasi* (cyt b and EF-1 α), and contributed a further three markers that showed concordant demographic patterns: anonymous nuclear loci AAm20 and AAm24 adapted from *P. perniciosus* microsatellites (Aransay *et al.*, 2001), and a protocol for the direct sequencing of the salivary peptide apyrase, based on its cDNA (Oliveira *et al.*, 2006).

All nuclear genes directly sequenced showed multiple genotypes, often with more than one polymorphic site. Ambiguous genotypes can be resolved directly through cloning, haplotype-specific extraction (HSE) (Nagy *et al.*, 2007), PCR amplification of specific alleles (PASA) (Sommer *et al.*, 1992), or indirectly by constructing haplotypes from genotypes through statistical programs such as PHASE (Stephens *et al.*, 2001). In this thesis, PASA was chosen for genotype scoring, because it provided high efficiency and accuracy upon optimization. For example, approximately 49% of flies had to be scored using a PASA system for the marker apyrase. This would have been too labour intensive to resolve by cloning, and the PASA approach also circumvented the need to

include priors for recombination or linkage disequilibrium in statistical algorithms, which can lower the accuracy of inference.

The use of molecular markers to estimate levels of genetic variability in a population depends on the assumption that they are selectively neutral. This study confirmed this assumption: within *P. ariasi*, all five loci characterized were shown not to be under positive or balancing selection (Chapters 2 and 3). The selection history of each marker was assessed at different time-scales, e.g. long-term by MK test, recent and current by *D* statistics, HWE and Ewans-Watterson. This approach made use of different genetic characteristics, and so safeguarded against conclusions based on any one test. This thesis also identified appropriate outgroups to *P. ariasi* for phylogenetic analyses, namely *P. chadlli*-like within the *P. ariasi* complex and species in its sister complex, *P. major* complex, all of which showed sufficient divergence without saturation.

5.4 Vector population genetics elucidate the effects of environmental change

Population genetic studies furnish information about the level of gene exchange between populations, where past effects of environmental change can provide information on future tendencies (DeChaine and Martin, 2005). This study searched for the existence of genetic signatures associated with environmental change, focusing on the low resolution spatial distribution of *P. ariasi* associated with Quaternary climate cycles (Chapter 2), and a high resolution spatial assessment on the restrictions to contemporary gene flow attributed to changes in local landscape (Chapter 4). Within the limitations of the data and analyses conducted, this Mediterranean species followed the paradigms for temperate species (Taberlet *et al.*, 1998; Hewitt, 1999) - that oscillating climates during the Quaternary caused repeated shifts in its distribution, evidenced by multiple isolation and re-colonization events that dated to this period.

Mitochondrial DNA revealed strong phylogenetic structure, whereas nuclear genes were less resolved. Of the conclusions reached, those most informative for the vector biology of *P. ariasi* include: the probable location of a glacial refuge north of the Pyrenees; the absence of any strong barrier to gene flow from Iberia into France; and, in contrast, the presence of a barrier to gene flow from the French Pyrenees to the Massif Central, perhaps as a result of land use patterns or "monopolization" (Loeuille and Leibold, 2008) (Chapter 2). Distinct French or northern Iberian mitochondrial lineages have not been observed in *P. perniciosus*, the sympatric vector of *L. infantum* (Esseghir

et al., 2000; Perrotey *et al.*, 2005), perhaps because of its lack of cold tolerance (Rioux *et al.*, 1967; Aransay *et al.*, 2004) prevented survival in France during the late glacials of the Pleistocene. In the context of the transmission dynamics of *Leishmania*, *P. ariasi* is likely to be the more persistent vector in France should there be climate cooling, and spread northwards first following climate warming. The melting pot of genetic diversity in southwest France offers the potential for genetic adaptation, including vectorial traits. This study found the potential for spread across the local environment of the northeast Pyrenees would not be hindered by the current heterogeneous landscape (Chapter 4). As sandflies are currently obligatory vectors for Mediterranean ZVL transmission, the spread of leishmaniasis could be curtailed by rendering vectors incapable of transmitting parasites (Ito *et al.*, 2002). In this respect, the lack of diversity of leading-edge populations in France, which are characterized by the near fixation of a single mitochondrial haplogroup (A), could exploit *Wolbachia*-induced cytoplasmic incompatibility as a mechanism to introduce and spread pathogen-blocking genes to modify vector competence (Hurst and Jiggins, 2000; Benlarbi and Ready, 2003). *Wolbachia* has been detected in *Phlebotomus* in France (Matsumoto *et al.*, 2008) and in *P. ariasi* in the study region (P.D. Ready and A. Cownie, unpublished data).

5.5 Proposing a vaccine candidate against Mediterranean ZVL

Control measures for VL include the early diagnosis and treatment of human cases, reducing the population of the insect vector by massive application of insecticides, and targeting sero-positive dogs (Ready, 2008). Reduction of canine susceptibility to leishmaniasis is proposed to be more effective than vector control in Europe (Dye, 1996). The frequency of some vector-borne diseases of pets is increasing in Europe, CanL among them (Beugnet and Marié, 2009). Therefore, the control of the transmission of canine leishmaniasis in southwest Europe has two potential goals: to reduce the likelihood of human disease and to protect dogs themselves. Control measures include the application of deltamethrin impregnated collars, which have been shown to reduce canine and human ZVL infection incidence by 43–86%. Practically, however, the efficiency of collars can be decreased by their loss or damage (Courtenay *et al.*, 2009), so vaccines provide a desirable alternative.

To date the only a licensed vaccine against CanL, Leishmune, comprises an antigen for *L. donovani* in Brazil (Nogueira *et al.*, 2005). Recently, the vaccine *Li*ESAp-MDP has shown experimental success, but is not commercially available. It is reported

to have an efficacy of 92% in experimentally and naturally infected dogs in France, with protection lasting for 24 months (Lemesre *et al.*, 2007). *Li*ESAp-MDP is based on antigens of *L. infantum* in formulation with muramyl dipeptide (MDP) as adjuvant. As described in Chapter 3, immune genes may be subject to co-evolutionary arms races that can drive the spread of resistant alleles. Salivary peptides are third generation vaccine candidates that show protection against leishmaniasis and are already in experimental trials (Palatnik-de-Sousa, 2008). Vaccine models predominantly target one of two vectors *L. longipalpis* and *P. papatasi* in the New and Old World, respectively. This thesis presented a study that was a "proof of principle", indicating how a population genetics approach can distinguish between adaptive and neutral evolution of a salivary peptide. In this example, the salivary peptide apyrase was shown to be selectively neutral in *P. ariasi*. This peptide does not elicit a host antibody response, but putatively confers protection against ZVL through a DTH cellular response (Oliveira *et al.*, 2006), so has theoretical potential as a vaccine candidate in the Mediterranean *Leishmania* transmission cycle.

5.6 Prospective studies

The importance of understanding sandfly population structure has implications for detecting clinical pleomorphisms and predicting epidemics (Maingon *et al.*, 2007). A knowledge base of the genealogical and phylogeographic relationships among *P. ariasi* populations was produced in this thesis. However, I conclude that the identification of further and more polymorphic makers, including single-locus microsatellites, would further enhance our understanding of *P. ariasi* population sub-structure (Chapter 4). Microsatellites have the potential to be transferred to closely related taxa (Sunnucks, 2000), but this study showed that there is unlikely to be any such transfer from *P. perniciosus* to *P. ariasi* (Aransay *et al.*, 2001; 2003). It was beyond the scope of this thesis to develop these typically hypervariable DNA sequences, but their attributes make them powerful markers for a broad range of population genetic questions. These markers can be used in a multilocus framework to provide information of within-population processes at the shortest time-scale, i.e. individual parentage and relatedness, and the identification of migrants (Sunnucks, 2000).

The effects of environmental change and ecological disturbance on the (re-) emergence of vector-borne diseases (Patz *et al.*, 2000) makes accurate development of

models predicting their impact relevant (Lafferty, 2009). Predictions of the shifts in the geographical distributions of sandflies, and therefore of *Leishmania*, have been generated using ecological niche and species distribution models (Peterson and Shaw, 2006; Ready, 2008). Such a methodological approach combines knowledge on both ecological requirements and current spatial occurrence of species, to predict the location of its fundamental niche - a location which can maintain a population without immigrational subsidy (Holt and Gomulkiewicz, 1996). However, a species may not occupy the entirety of its fundamental niche, as model assumptions are either inaccurate or environmental factors critically influencing species distribution are not modelled, e.g. historical or local constraints on dispersal. Furthermore, statistical models are not always applicable outside their original geographical region (Ready, 2008), so an integrated approach is required. Morin and Lechowicz (2008) have reviewed the factors needed to model the evolutionary ecology of a niche at hierarchical spatial scales (i.e. regional to landscape to local community), with the aim of building species distribution models that are most likely to yield accurate predictions of species occurrence, and thus spread. Variables parameterized included abiotic dimensions such as macro- and micro-climate, and landscape topography, as well as biotic dimensions such as dispersal ability and competition. Such approaches should be used to develop and validate risk models for the northward spread and persistence of *P. ariasi*. In France, at the leading-edge of *Leishmania* distribution, extensive knowledge exists on: the descriptive ecology and biology of *P. ariasi* in relation to the ZVL transmission cycle (publications in the Cévennes by Rioux, Killick-Kendrick and colleagues as previously discussed); its absence/presence and relative abundance both regionally and locally; and, preferred topographical, macro- and micro-environmental data (EDEN partners). This study adds a further dimension, for the first time analysing this species' population genetic structure, contributing information on both historical and contemporary time-scales, as well as providing information on the effects of climate and habitat changes on distribution.

REFERENCES

Alvar, J. and Jimenez, M. (1994) Could infected drug-users be potential *Leishmania infantum* reservoirs? *AIDS*, **8**, 854.

Anderson, J.M., Oliveira, F., Kamhawi, S., Mans, B.J., Reynoso, D., Seitz, A.E., Lawyer, P., Garfield, M., Pham, M.Y. and Valenzuela, J.G. (2006) Comparative salivary gland transcriptomics of sandfly vectors of visceral leishmaniasis. *BMC Genomics*, **7**: 52.

Anderson, M.J. (2004) DISTLM v.5: a FORTRAN computer program to calculate a distance-based multivariate analysis for a linear model. Department of Statistics, University of Auckland, New Zealand. Available from: http://www.stat.auckland. ac.nz/~mja.

Anisimova, M., Bielawski, J.P. and Yang, Z. (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Molecular Biology and Evolution*, **19**, 950-2002.

Anisimova, M., Nielsen, R. and Yang, Z. (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics*, **164**, 1229-1236.

Aransay, A.M., Malarky, G. and Ready, P.D. (2001) Isolation (with enrichment) and characterization of trinucleotide microsatellites from *Phlebotomus perniciosus*, a vector of *Leishmania infantum*. *Molecular Ecology Notes*, **1**, 176-178.

Aransay, A.M., Ready, P.D. and Morillas-Marquez, F. (2003) Population differentiation of *Phlebotomus perniciosus* in Spain following postglacial dispersal. *Heredity*, **90**, 316-325.

Aransay, A.M., Testa, J.M., Morillas-Marquez, F., Lucientes, J. and Ready, P.D. (2004) Distribution of sandfly species in relation to canine leishmaniasis from the Ebro Valley to Valencia, northeastern Spain. *Parasitology Research*, **94**, 416-420.

Ashford, R.W. (1996) Leishmaniasis reservoirs and their significance in control. *Clinics in Dermatology*, **14**, 523-532.

Ashford, R.W. (2000) The leishmaniasies as emerging and reemerging zoonoses. *International Journal of Parasitology*, **30**, 1269-1281.

Avise, J.C. (1994) *Molecular markers. Nautral History and Evolution*. New York: Chapman and Hall.

Avise, J.C. (2000) *Phylogeography: The History and Formation of Species*. Cambridge, Massachesettes: Harvard University Press.

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reed, C.A. and Saunders, N.C. (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, **18**, 489-522.

Avise, J.C., Walker, D. and Johns, G.C. (1998) Speciation durations and Pleistocene effects on vertebrate phylogeography. *Proceedings of the Royal Society of London B: Biological Sciences*, **265**, 1707-1712.

Baguette, M. and Dyck, H.V. (2007) Landscape connectivity and animal behavior: functional grain as a key determinant for dispersal. *Landscape Ecology*, **22**, 1117-1129.

Balent, G. and Courtiade, B. (1992) Modelling bird communities/landscape patterns relationships in a rural area of South-Western France. *Landscape Ecology*, **6**, 195-211.

Ballard, J.W. and Kreitman, M. (1994) Unraveling selection in the mitochondrial genome of *Drosophila*. *Genetics*, **138**, 757-772.

Ballard, J.W. and Whitlock, M.C. (2004) The incomplete natural history of mitochondria. *Molecular Ecology*, **13**, 729-744.

Bates, P. (2007) Transmission of *Leishmania* metacyclic promastigotes by phlebotomine sand flies. *International Journal for Parasitology*, **37**, 1097-1106.

Baum, D. A. and Donoghue, M.J. (1995) Choosing among alternative "phylogenetic" species concepts. *Systematic Botany*, **20**, 560-573.

Bazin, E., Glemin, S. and Galtier, N. (2006) Population size does not influence mitochondrial genetic diversity in animals. *Science*, **312**, 570-572.

Beaudouin, C., Jouet, G., Suc, J-P., Berné, S. and Escarguel, G. (2007) Vegetation dynamics in southern France during the last 30 ky BP in the light of marine palynology. *Quaternary Science Reviews*, **26**, 1037-1054.

Belkaid, Y., Kamhawi, S., Modi, G., Valenzuela, J., Noben-Trauth, N., Rowton, E., Ribeiro, J. and Sacks, D.L. (1998) Development of a natural model of cutaneous leishmaniasis: powerful effects of vector saliva and saliva preexposure on the long-term outcome of *Leishmania major* infections in the mouse ear dermis. *Journal of Experimental Biology*, **188**, 1941-1953.

Bellgard, M.I. and Gojobori, T. (1999) Inferring the direction of evolutionary changes of genomic base composition. *Trends in Genetics*, **15**, 254-256.

Benlarbi, M. and Ready, P.D. (2003) Host-specific *Wolbachia* strains in widespread populations of *Phlebotomus perniciosus* and *P. papatasi* (Diptera: Psychodidae), and prospects for driving genes into these vectors of *Leishmania*. *Bulletin of Entomological Research*, **93**, 383-391.

Berry, O., Tocher, M.D. and Sarre, S.D. (2004) Can assignment tests measure dispersal? *Molecular Ecology*, **13**, 551-561.

Beugnet, F. and Marié, J-L. (2009) Emerging arthropod-borne diseases of companion animals in Europe. *Veterinary Parasitology*, **163**, 298-305.

Beverley, S.M. and Dobson, D.E. (2004) Flypaper for parasites. *Cell*, **119**, 311-316.

Bilton, D.T., Mirol, P.M., Mascheretti, S., Fredga, K., Zima, J. and Searle, J.B. (1998) Mediterranean Europe as an area of endemism for small mammals rather than a source of northwards postglacial colonization. *Proceedings of the Royal Society of London B: Biological Sciences*, **265**, 1219-1226.

Birky, C.W., Jr., Wolf, C., Maughan, H., Herberston, L. and Henry, E. (2005) Speciation and selection without sex. *Hydrobiologia*, **546**, 29-45.

Bongiorno, G., Habluetzelb, A., Khourya, C. and Maroli, M. (2003) Host preferences of phlebotomine sand flies at a hypoendemic focus of canine leishmaniasis in central Italy. *Acta Tropica*, **88**, 109-116.

Bossart, J.L. and Prowell, D.P. (1998) Genetic estimates of population structure and gene flow limitations, lessons and new directions. *Trends in Ecology and Evolution*, **13**, 202-206.

Boussaa, S., Pesson, B. and Boumezzough, A. (2009) Faunistic study of the sandflies (Diptera: Psychodidae) in an emerging focus of cutaneous leishmaniasis in Al Haouz province, Morocco. *Annals of Tropical Medicine & Parasitology*, **103**, 73-83.

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 Sciences USA*, **91**, 6491-6495.

Calvet, M. (2004) The Quaternary glaciation of the Pyrenees. **In**: Ehlers, J. and Gibbard, P.L. (eds.) *Quaternary glaciations; extent and chronology: Pt. 1 Europe (developments in Quaternary science 2)*. Amsterdam: Elsevier Science.

Campanella, J.J., Bitincka, L. and Smalley, J. (2003) MatGAT: An application that generates similarity/identity matrices using protein or DNA sequences. *BMC Bioinformatics*, **4**: 29.

Canestrelli, D., Cimmaruta, R. and Nascetti, G. (2007) Phylogeography and historical demography of the Italian treefrog, *Hyla intermedia*, reveals multiple refugia, populations expansions and secondary contacts within peninsula Italy. *Molecular Ecology*, **16**, 4808-4821.

Carranza, S., Arnold, E.N. and Pleguezuelos, J.M. (2006) Phylogeny, biogeography, and evolution of two Mediterranean snakes, *Malpolon monspessulanus* and *Hermorrhois hippocrepis* (Squamata, Colubridae), using mtDNA sequences. *Molecular Phylogenetics and Evolution*, **40**, 532-546.

Chamkhi, J., Guerbouj, S., Ben Ismail, R. and Guizani, I. (2006) Description de la femelle de *Phlebotomus* (*Larroussius*) c*hadlii* Rioux, Juminer et Gibily, 1966 (Díptera: Psychodidae). D´après un exemplaire capturé aux environs du Kef (Tunisie). *Parasite*, **13**, 299-303.

Champagne, D.E. and Valenzuela, J.G. (1996) Pharmacology of haematophagous arthropod saliva. **In**: Wikel, S.K. (ed.) *The immunology of host-ectoparasitic arthropod relationships*. Wallingford: CAB International.

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

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

Collin, N., Gomes, R., Teixeira, C., Cheng, L., Laughinghouse, A., Ward, J.M., Elnaiem, D.E., Fischer, L., Valenzuela, J.G. and Kamhawi, S. (2009) Sand fly salivary proteins induce strong cellular immunity in a natural reservoir of visceral leishmaniasis with adverse consequences for *Leishmania*. *PLoS Pathogens*, **5**: e1000441.

Coope, G. (1994) The response of insect faunas to glacial-interglacial climate fluctuations. *Proceedings of the Royal Society of London B: Biological Sciences*, **344**, 19-26.

Cornuet, J.M., Piry, S., Luikart, G., Estoup, A. and Solignac, M. (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, **153**, 1989-2000.

Cosson, J.-F., Hutterer, R., Libois, R., Sarà, M., Taberlet, P. and Vogel, P. (2005) Phylogeographical footprints of the Strait of Gibraltar and Quaternary climatic fluctuations in the western Mediterranean: a case study with the greater white-toothed shrew, *Crocidura russula* (Mammalia: Soricidae). *Molecular Ecology*, **21**, 295-305.

Courtenay, O., Kovacic, V., Gomes, P.A.F., Garcez, L.M. and Quinnell, R.J. (2009) A long-lasting topical deltamethrin treatment to protect dogs against visceral leishmaniasis. *Medical and Veterinary Entomology*, **23**, 245-256.

Couvet, D. (2002) Deleterious effects of restricted gene flow in fragmented populations. *Conservation Biology*, **16**, 369-376.

Cupp, M.S., Cupp, E.W. and Ramberg, F.B. (1993) Salivary gland apyrase in black flies (*Simulium vittatum*). *Insect Physiology*, **39**, 817-821.

Curtis, C. (1999) Can molecular biology contribute usefully to vector control? *Schweizerische Medizinische Wochenschrift*, **129**, 1111-1116.

Dai, J., Liu, J., Deng, Y., Smith, T.M. and Lu, M. (2004) Structure and protein design of a human platelet function inhibitor. *Cell*, **116**, 649-659.

Danforth, B.N. and Ji, S. (1998) Elongation Factor-1α occurs as two copies in bees: implications for phylogenetic analysis of EF-1α sequences in insects. *Molecular Biology and Evolution*, **15**, 225-235.

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

Dawkins, R. and Krebs, J.R. (1979) Arms races between and within species. *Proceedings of the Royal Society of London B: Biological Sciences*, **205**, 489-511.

DeChaine, E.G. and Martin, A.P. (2005) Historical biogeography of two alpine butterflies in the Rocky Mountains: broad-scale concordance and local-scale discordance. *Journal of Biogeography*, **32**, 1943-1956.

De Colmenares, M., Portús, M., Botet, J., Dobaño, C., Gállego, M., Wolff, M. and Seguí, G. (1995) Identification of blood meals of *Phlebotomus perniciosus* (Diptera: Psychodidae) in Spain by a competitive enzyme-linked

immunosorbent assay biotin/avidin method. *Journal of Medical Entomology*, **32**, 229-233.

Deffontaine, V., Ledevin, R., Fontaine, M.C., Quéré, J.-P., Renaud, S., Libois, R. and Michaux, J.R. (2009) A relic bank vole lineage highlights the biogeographic history of the Pyrenean region in Europe. *Molecular Ecology*, **18**, 2489-2502.

Delmas, M., Gunnell, Y., Braucher, R., Calvet, M. and Bourlès, D. (2008) Exposure age chronology of the last glaciation in the eastern Pyrenees. *Quaternary Research*, **69**, 231-241.

Dereure, J., Pratlong, F. and Dedet, J.-P. (1999) Geographical distribution and the identification of parasites causing canine leishmaniasis in the Mediterranean Basin. **In:** Killick-Kendrick, R. (ed.) *Canine Leishmaniasis: an update.* Proceedings of the International Canine Leishmaniasis Forum. Barcelona 1999, 18-25.

Desjeux, P. (2001) The increase in risk factors for leishmaniasis worldwide. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **95**, 239-243.

Didham, R.K., Ghazoul, J., Stork, N.E. and Davis, A.J. (1996) Insects in fragmented forests: a functional approach. *Trends in Ecology and Evolution*, **11**, 255-260.

Dieckmann, U., O'Hara, B. and Weisser, W. (1999) The evolutionary ecology of dispersal. *Trends in Ecology and Evolution*, **14**, 88-90.

Di Muccio, T., Marinucci, M., Frusteri, L., Maroli, M., Pesson, B. and Gramiccia, M. (2000) Phylogenetic analysis of *Phlebotomus* species belonging to the subgenus *Larroussius* (Diptera, Psychodidae) by ITS2-rDNA sequences. *Insect Biochemistry and Molecular Biology*, **30**, 387-393.

Donoghue, M.J. (1985) A critique of the biological species concept and recommendations for a phylogenetic alternative. *The Bryologist*, **88**, 172-181.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J. and Rambaut, A. (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biology*, **4**, 0699-0710.

Dujardin, J.-C. (2006) Risk factors in the spread of leishmaniases: towards integrated monitoring. *TRENDS in Parasitology*, **22**, 4-6.

Dujardin, J.-C., Campino, L., Caňavate, C., Dedet, J.-P., Gradoni, L., Soteriadou, K., Mazeris, A., Ozbel, Y. and Boelaert, M. (2008) Spread of vector-borne diseases and neglect of leishmaniasis, Europe. *Emerging Infectious Diseases*, **14**, 1013-1018.

Dyck, H.V. and Baguette, M. (2005) Dispersal behaviour in fragmented landscapes: routine or special movements? *Basic and Applied Ecology*, **6**, 535-545.

Dye, C. (1996) The logic of visceral leishmaniasis control. *The American Journal of Tropical Medicine and Hygiene*, **55**, 125-130.

Dye, C., Guy, M.W., Elkins, D.B., Wilkes, T.J. and Killick-Kendrick, R. (1987) The life expectancy of phlebotomine sandflies: first field estimates from southern France. *Medical and Veterinary Entomology*, **1**, 417-425.

Dye, C. and Reiter, P. (2000) Temperatures without fevers? *Science*, **289**, 1697-1698.

Eldredge, N. and J. Cracraft. (1980) Phylogenetic patterns and the evolutionary process. New York: Columbia University Press.

Elmer, K.R., Dávila, J.A. and Lougheed, S.C. (2007) Applying new inter-individual approaches to assess fine-scale population genetic diversity in a neotropical frog, *Eleutherodactylus ockendeni*. *Heredity*, **99**, 506-515.

Elnaiem, D.-E.A., Meneses, C., Slotman, M. and Lanzaro, G.C. (2005) Genetic variation in the sandfly salivary protein, SP-15, a potential vaccine candidate against *Leishmania major*. *Insect Molecular Biology*, **14**, 145-150.

Emberger, L. (1936) Présentation de la carte phytogéographique du Maroc au 1/1500000. *C. R. Séanc. Mens. Soc. Sci. Nat. Phys. Maroc*, **4**, 8-29.

Emberger, L. (1939) Aperçu général sur la végétation du Maroc. Commentaire de la carte phytogéographique du Maroc 1:1500000. *Veröff Geobot Instit, Eidgen Techn. Hochsch Rübel Zürich*, **14**, 40-157.

Emes, R.D. and Yang, Z. (2008) Duplicated paralogous genes subject to positive selection in the genome of *Trypanosoma brucei*. *PLoS ONE*. **5**:e2295.

Endo, T., Ikeo, K. and Gojobori, T. (1996) Large-scale search for genes on which positive selection may operate. *Molecular Biology and Evolution*, **13**, 685-690.

Erixon, P., Svennblad, B., Britton, T., and Oxelman, B. (2003) Reliability of Bayesian Posterior Probabilities and Bootstrap Frequencies in Phylogenetics. *Systematic Biology*, **52**, 665-673.

Esseghir, S., Ready, P.D. and Ben-Ismail, R. (2000) Speciation of *Phlebotomus* sandflies of the subgenus *Larroussius* coincided with the late Miocene–Pliocene aridification of the Mediterranean subregion. *Biological Journal of the Linnean Society*, **70**, 189-219.

Esseghir, S., Ready, P.D., Killick-Kendrick, R. and Ben-Ismail, R. (1997) Mitochondrial haplotypes and phylogeography of *Phlebotomus* vectors and *Leishmania major*. *Insect Molecular Biology*, **6**, 211-225.

Evanno, G., Regnaut, S. and Goudet, J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611-2620.

Ewers, R.M. and Didham, R.K. (2006) Confounding factors in the detection of species responses to habitat fragmentation. *Biological Reviews*, **81**, 117-142.

Excoffier, L., Laval, G. and Schneider, S. (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**, 47-50.

Excoffier, L., Smouse, P.E. and Quattro, J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479-491.

Falush, D., Stephens, M. and Pritchard, J.K. (2003) Inference of population structure: extensions to linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Fisher, R.A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, **52**, 399-433.

Flemming, N., Bailey, G., Courtillot, V., King, G., Lambeck, K., Ryerson, F. and Vita-Finzi, C. (2003) Coastal and marine palaeo-environments and human dispersal points across the Africa-Eurasia boundary. **In**: Brebbia, C.A. and Gambin, T. (eds.) *Maritime Heritage*. Wessex Institute of Technology, Southampton, UK and University of Malta, Malta. pp. 13.

Foley, J.A., DeFries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., Chapin, F.S., Coe, M.T., Daily, G.C., Gibbs, H.K., Helkowski, J.H., Holloway, T., Howard, E.A., Kucharik, C.J., Monfreda, C., Patz, J.A., Prentice, I.C., Ramankutty, N. and Snyder, P.K. (2005) Global consequences of land use. *Science*, **309**, 570-574.

Ford, M.J. (2002) Applications of selective neutrality tests to molecular ecology. *Molecular Ecology*, **11**, 1245-1262.

Fu, Y-X. (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, **147**, 915-925.

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

Gállego, B.J., Botet, F.J., Gállego, C.M. and Portũs, V.M. (1992) Los phlebotomos de la España peninsular e Islas Baleares. Identificación y corología. Comentarios sobre los métodos de captura. **In**: Memoriam Prof Doc D Francisco de Paula Martinez Gomez. S. Hernandez-Rodrigues, Cordoba, 579-600.

Garrigan, D. and Hedrick, P.W. (2003) Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. *Evolution*, **57**, 1707-1722.

Gibbard, P. and van Kolfschoten, T. (2004) The Pleistocene and Holocene Epochs. **In**: Gradstein, F.M., Ogg, J.G. and Smith, A.G. (eds.) *A Geologic Time Scale 2004*. Cambridge: Cambridge University Press.

Gilbert, S.C., Plebanski, M., Gupta, S., Morris, J., Cox, M., Aidoo, M., Kwiatkowski, D., Greenwood, B., Whittle, H.C. and Hill, A.V.S. (1998) Association of Malaria Parasite Population Structure, HLA, and Immunological Antagonism. *Science*, **279**, 1173-1177.

Gissi, C., Iannelli, F. and Pesole, G. (2008) Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. *Heredity*, **101**, 301-320.

Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, **11**, 725-736.

Gomes, R., Teixeira, C., Teixeira, M.J., Oliveira, F., Menezes, M.J., Silva, C., de Oliveira, C.I., Miranda, J.C., Elnaiem, D.E., Kamhawi, S., Valenzuela, J.G. and Brodskyn, C.I. (2008) Immunity to a salivary protein of a sand fly vector protects against the fatal outcome of visceral leishmaniasis in a hamster model. *Proceedings of the National Academy of Sciences of the USA*, **105**, 7845-7850.

Gómez, A. and Lunt, D.H. (2006) Refugia within refugia: patterns of phylogeographic concordance in the Iberian Peninsula. **In**: Weiss, S. and Ferrand, N. (eds.) *Phylogeography of Southern European Refugia: Evolutionary perspectives on the origins and conservation of European biodiversity*. Netherlands: Kluwer Academic Publishers Group.

Gómez, A., Montero-Pau, J., Lunt, D.H., Serra, M. and Campillo, S. (2007) Persistent genetic signatures of colonization in *Brachionus manjavacas* rotifers in the Iberian Peninsula. *Molecular Ecology*, **16**, 3228-3240.

Goudet J. (2002) FSTAT (version 2.9.3.2) A programme to estimate and test gene diversities and fixation indices. Available from: http://www2.unil.ch/popgen/softwares/fstat.htm.

Grandes, A.E., Gomez-Bautista, M., Novo, M.M. and Martin, F.S. (1988) Leishmaniasis in the province of Salamanca, Spain. Prevalance in dogs and seasonal dynamics of vectors. *Annales de Parasitologie Humaine et Comparée*, **68**, 387-397.

Grimaldi, D. and Engel, M.S. (2005) Evolution of insects. **In**: *Evolution of ectoparasites and blood feeders of vertebrates*. New York: Cambridge University Press.

Gugerli, F., Sperisen, C., Büchler, U., Magni, F., Geburek, T., Jeandroz, S. and Senn, J. (2001) Haplotype variation in a mitochondrial tandem repeat of Norway spruce (*Picea abies*) populations suggests a serious founder effect during postglacial re-colonization of the western Alps. *Molecular Ecology*, **10**, 1255-1263.

Guillot, G., Leblois, R., Coulon, A. and Frantz, A.C. (2009) Statistical methods in spatial genetics. *Molecular Ecology*, **18**, 4734-4756.

Guo, S.W. and Thomson, E.A. (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, **48**, 361-372.

Guy, M.W., Killick-Kendrick, R., Gill, G.S., Rioux, J.A. and Bray, R.S. (1984) Ecology of leishmaniasis in the south of France. Determination of the hosts of *Phlebotomus ariasi* Tonnoir, 1921 in the Cévennes by bloodmeal analyses. *Annales de Parasitologie Humaine et Comparée*, **59**, 449-458.

Habel, J.C., Schmitt, T. and Müller, P. (2005) The fourth paradigm pattern of post-glacial range expansion of European terrestrial species: the phylogeography of the Marbled White butterfly (Satyrinae, Lepidoptera). *Journal of Biogeography*, **32**, 1489-1497.

Hahn, M.W. (2009) Distinguishing among evolutionary models for the maintenance of gene duplicates. *Journal of Heredity*, **100**, 605-617.

Haldane, J.B.S. (1932) *The Causes of Evolution.* London: Longmans, Green & Co., Ltd.

Hall, B.G. (2004) Phylogenetic Trees Made Easy: A How-To Manual, Second Edition. USA: Sinauer Associates, Inc.

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

Hamarsheh, O., Presber, W., Abdeen, Z., Sawalha, S., Al-Lahem, A. and Schönian, G. (2007) Genetic structure of Mediterranean populations of the sandfly *Phlebotomus papatasi* by mitochondrial cytochrome b haplotype analysis. *Medical and Veterinary Entomology*, **21**, 270-277.

Hamasaki, R., Kato, H., Terayama, Y., Iwata, H. and Valenzuela, J.G. (2009) Functional characterization of a salivary apyrase from the sand fly, *Phlebotomus duboscqi*, a vector of *Leishmania major*. *Journal of Insect Physiology*, **55**, 1044-1049.

Hampe, A. and Petit, R.J. (2005) Conserving biodiversity under climate change: the rear edge matters. *Ecological Letters*, **8**, 461-467.

Handman, E. (1999) Cell biology of *Leishmania*. *Advances in Parasitology*, **44**, 1-39.

Hanski, I., Gilpin, M.E. (eds.) (1997) *Metapopulation Biology: Ecology, Genetics, and Evolution*. San Diego: Academic Press.

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

Hart, M.W. and Sunday, J. (2007) Things fall apart: biological species form unconnected parsimony networks. *Biology Letters*, **3**, 509-512.

Hartl, D.E. (1981) A primer of population genetics. Massachusetts: Sinauer Associates, Inc.

Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160-174.

Hedrick, P.W. (2000) *Genetics of Populations* (2nd edition). Sudbury, MA: Jones and Bartlett Publishers.

Hewitt, G.M. (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society*, **58**, 247-276.

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

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

Hewitt, G.M. (2001) Speciation, hybrid zones and phylogeography - or seeing genes in space and time. *Molecular Ecology*, **10**, 537-549.

Hewitt, G.M. (2004a) Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **359**, 183-195.

Hewitt, G.M. (2004b) The structure of biodiversity - insights from molecular phylogeography. *Frontiers in Zoology*, **1**: 4.

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

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

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

Hofman, S., Spolsky, C., Uzzell, T., Cogălniceanu, D., Babik, W. and Szymura, J.M. (2007) Phylogeography of the fire-bellied toads *Bombina*: independent Pleistocene histories inferred from mitochondrial genomes. *Molecular Ecology*, **16**, 2301-2316.

Holder, M. and Lewis, P.O. (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics*, **4**, 275-284.

Holderegger, R. and Wagner, H.H. (2008) Landscape Genetics. *BioScience*, **58**, 199-207.

Holm, S. (1979) A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65-70.

Holt, R.D. and Gomulkiewicz, R. (1996) The evolution of species niches: a population dynamic perspective. **In**: Adler, F.R., Lewis, M.A. and Dallon, J.C. (eds.) Case studies in mathematical modeling: ecology, physiology, and cell biology. Saddle Hill, New Jersey: Prentice-Hall. pp. 25-50.

Hovemann, B., Richter, S., Walldorf, U. and Cziepluch, C. (1988) Two genes encode related cytoplasmic elongation factors-1α (EF-1α) in *Drosophila melanogaster* with continuous and stage specific expression. *Nucleic Acids Research*, **16**, 3175-3194.

Hubisz, M.J., Falush, D., Stephens, M. and Pritchard, J.K. (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322-1332.

Hudson, R.R. (1987) Estimating the recombination parameter of a finite population model without selection. *Genetical Research*, **50**, 245-250.

Hudson, R.R. and 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.

Huelsenbeck, J.P., Larget, B., Miller, R.E. and Ronquist, F. (2002) Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology*, **51**, 673-688.

Huntley, B. (2001) Reconstructing past environments from the Quaternary palaeovegetation record. *Biology and Environment: Proceedings of the Royal Irish Academy*, **101B**, 3-18.

Hurles, M. (2004) Gene duplication: the genome trade in spare parts. *PLoS Biology*, **2**, 0900-0904.

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

Hurst, L.D. and Smith, N.G.C. (1999) Do essential genes evolve slowly? *Current Biology*, **9**, 747-750.

Ibrahim, K.M., Nichols, R.A. and Hewitt, G.M. (1996) Spatial patterns of genetic variation generated by different forms of dispersal during range expansions. *Heredity*, **77**, 282-291.

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

Ish-Horowicz, D. (1982) Personal communication in: Rate of turnover of structural variants in the rDNA gene family of *Drosophila melanogaster*. *Nature*, **295**, 564-568.

Ito, J., Ghosh, A., Moreira, L.A., Wimmer, E.A. and Jacobs-Lorena, M. (2002) Transgenic anopheline mosquitoes impaired in transmission of a malaria parasite. *Nature*, **417**, 452-455.

Jakobsson, M. and Rosenberg, N.A. (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801-1806.

Jalut, G., Dedoubat, J.J., Fontugne, M. and Otto, T. (2009) Holocene circum-Mediterranean vegetation changes: Climate forcing and human impact. *Quaternary International*, **200**, 4-18.

Javadian, E., Tesh, R., Saidi, S. and Nadim, A. (1977) Studies on the epidemiology of sandfly fever in Iran. III. Host-feeding patterns of *Phlebotomus papatasi* in an endemic area of the disease. *American Journal of Tropical Medicine and Hygiene*, **26**, 294-298.

Jennersten, O., Loman, J., Méller, A.P., Robertson, J. and Wideâ, B. (1997) Conservation biology in agricultural habitat islands. *Ecological Bulletins*, **46**, 72-87.

Jenness, J. (2005) Distance Matrix (dist_mat_jen.avx) extension for ArcView 3.x, v.2. Jenness Enterprises. Available from: http://www.jennessent.com/arcview/distmatrix.htm.

Jia, L., Clegg, M.T. and Jiang, T. (2003) Excess of non-synonymous substitutions suggest that positive selection episodes occurred during the evolution of DNA-binding domains in the *Arabidopsis* R2R3-MYB gene family. *Plant Molecular Biology*, **53**, 627-642.

Jiggins, F.M. and Kim, K.W. (2007) A screen for immunity genes evolving under positive selection in *Drosophila*. *Journal of Evolutionary Biology*, **20**, 965-970.

Jung, J., Kim, T-G., Lyons, G.E., Kim, H.-R. and Lee, Y. (2005) Jumonji regulates cardiomyocyte proliferation via interaction with retinoblastoma protein. *The Journal of Biological Chemistry*, **280**, 30916-30923.

Kamhawi, S., Belkaid, Y, Modi, G., Rowton, E. and Sacks, D. (2000) Protection against cutaneous leishmaniasis resulting from bites of uninfected sand flies. *Science*, **290**, 1351-1354.

Kandul, N.P., Lukhtanov, V.A., Dantchenko, A.V., Coleman, J.S.W., Sekercioglu, C.H., Haig, D. and Pierce, N.E. (2004) Phylogeny of *Agrodiaetus* Hübner 1822 (Lepidoptera: Lycaenidae) inferred from mtDNA Sequences of COI and COII and Nuclear Sequences of EF1-α: Karyotype Diversification and Species Radiation. *Systematic Biology*, **53**, 278-298.

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

Kato, H., Anderson, J.M., Kamhawi, S., Oliveira, F., Lawyer, P.G., Pham, V.M., Sangare, C.S., Samake, S., Sissoko, I., Garfield, M., Sigutova, L., Volf, P., Doumbia, S. and Valenzuela, J.G. (2006) High degree of conservancy among secreted salivary gland proteins from two geographically distant *Phlebotomus duboscqi* sandflies populations (Mali and Kenya). *BMC Genomics*, **7**: 226.

Kidd, D.M. and Richie, M.G. (2006) Phylogeographic information systems: putting the geography into phylogeography. *Journal of Biogeography*, **33**, 1851-1865.

Killick-Kendrick, R. (1990) Phlebotomine vectors of the leishmaniases: a review. *Medical and Veterinary Entomology*, **4**, 1-24.

Killick-Kendrick, R., Rioux J.-A., Bailly, M., Guy, M.W., Wilkes, T.J., Guy, F.M., Davidson, I., Knechtli, R., Ward, R.D., Guilvard, E., Perieres, J. and Budois, H. (1984) Ecology of leishmaniasis in the south of France. 20. Dispersal of *Phlebotomus ariasi* Tonnoir, 1921 as a factor in the spread of visceral leishmaniasis in the Cévennes. *Annales de Parasitologie Humaine et Comparée*, **59**, 555-572.

Kim, Y. and Nielsen, R. (2004) Linkage disequilibrium as a signature of selective sweeps. *Genetics*, **167**, 1513-1524.

Kimura, M. (1953) 'Stepping-stone' model of population. *Annual Report National Institute of Genetics, Japan*, **3**, 62-63.

Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature*, **217**, 642-626.

Kimura, M. (1983) The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Kimura, M. and Weiss, G.H. (1964) The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, **49**, 561-576.

Knowles, L.L. and Maddison, W.P. (2002) Statistical phylogeography. *Molecular Ecology*, **11**, 2623-2635.

Knowles, L.L. and Richards, C.L. (2005) Importance of genetic drift during the Pleistocene as revealed by analyses of genomic variation. *Molecular Ecology*, **14**, 4023-4032.

Kocher, T.D., Thomas, W.K., Meyer, A., Edwards, S. V., Pääbo, S., Villablanca, F.X. and Wilson, A.C. (1989) Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proceedings of the National Academy of Sciences of the USA*, **86**, 6196-6200.

Kuhner, M.K. (2009) Coalescent genealogy samplers: windows into population history. *Trends in Ecology and Evolution*, **24**, 86-93.

Kwok, S., Kellogg, D.E., McKinney, N., Spasic, D., Goda, L., Levenson, C. and Sninsky, J.J. (1990) Effects of primer - template mismatches on the polymerase chain reaction: Human immunodeficiency virus type 1 model studies. *Nucleic Acids Research*, **18**, 999-1005.

Lafferty, K.D. (2009) The ecology of climate change and infectious diseases. *Ecology*, **90**, 888-900.

Lambrechts, L., Knox, T.B., Wong, J., Liebman, K.A., Albright, R.G. and Stoddard, S.T. (2009) Shifting priorities in vector biology to improve control of vector-borne disease. *Tropical Medicine and International Health*, **14**, 1-10.

Lane, R.P. and Crosskey, R.W. (eds.) (1993) *Medical insects and arachnids*. London: Natural History Museum.

Lanzaro, G.C., Lopes, A.H.C.S., Ribeiro, J.M.C., Shoemaker, C.B., Warbug, A., Soares, M. and Titus, R.G. (1999) Variation in the salivary peptide, maxadilan, from species in the *Lutzomyia longipalpis* complex. *Insect Molecular Biology*, **8**, 1-9.

Lemesre, J-L., Holzmuller, P., Goncalves, R.B., Bourdoiseau, G., Hugnet, C., Cavaleyra, M. and Papierok, G. (2007) Long-lasting protection against canine visceral leishmaniasis using the *Li*ESAp-MDP vaccine in endemic areas of France: Double-blind randomised efficacy field trial. *Vaccine*, **25**, 4223-4234.

Lewis, P.O. (2001) Phylogenetic systematics turns over a new leaf. *Trrends in Ecology and Evolution*, **16**, 30-37.

Lewontin, R.C. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, **49**, 49-67.

Lisiecki, L.E., and M.E. Raymo. (2005) A Pliocene-Pleistocene stack of 57 globally distributed benthic δ18O records. *Paleoceanography*, **20**: PA1003.

Loeuille, N. and Leibold, M.A. (2008) Evolution in metacommunities: On the relative importance of species sorting and monopolization in structuring communities. *The American Naturalist*, **171**, 788-799.

Lozier, J.D. and Mills, N.J. (2009) Ecological niche models and coalescent analysis of gene flow support recent allopatric isolation of parasitoid wasp populations in the Mediterranean. *PLoS ONE*, **4**: e5901.

Lumaret, R., Mir, C., Michaud, H. and Raynal, V. (2002) Phylogeographical variation of chloroplast DNA in holm oak (*Quercus ilex* L.). *Molecular Ecology*, **11**, 2327-2336.

Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151-1155.

Macarthur, R.H. and Wilson, E.O. (1967) The theory of island biogeography. Princeton: Princeton University Press.

Maingon, R.D.C., Ward, R.D., Hamilton, J.G.C., Bauzer, L.G.S.R. and Peixoto, A.A. (2007) The *Lutzomyia longipalpis* species complex: does population substructure matter to *Leishmania* transmission? *TRENDS in Parasitology*, **24**, 12-17.

Mank, J.E. and Avise, J.C. (2004) Individual organisms as units of analysis: Bayesian-clustering alternatives in population genetics. *Genetics Research*, **84**, 135-143.

Mants, M.J. and Parker, K.R. (1981) Two platelet aggregation inhibitors in tsetse (*Glossina*) saliva with studies of roles of thrombin and citrate in *in vitro* platelet aggregation. *British Journal of Haematology*, **48**, 601-608.

Mar, J.C., Harlow, T.J. and Ragan, M.A. (2005) Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. *BMC Evolutionary Biology*, **5**: 8.

Marcus, A.J. and Safier, L.B. (1993) Thromboregulation: multicellular modulation of platelet reactivity in hemostasis and thrombosis. *The FASEB Journal*, 7, 516-522.

Maroli, M., Rossi, L., Baldelli, R., Capelli, G., Ferroglio, E., Genchi, C., Gramiccia, M., Mortarino, M., Pietrobelli, M. and Grandoni, L. (2008) The northwards spread of leishmaniasis in Italy: evidence from retrospecitive and ongoing studies on the canine reservoir and phlebotomine vectors. *Tropical Medicine and International Health*, **13**, 256-264.

Martínez, S., Vanwambeke, S.O., Ready, P. (2007) Linking changes in landscape composition and configuration with sandfly occurrence in southwest France. *Analysis of multi-temporal remote sensing images, 2007. MultiTemp 2007*, **18-20**, 1-5.

Martínez-Solano, I., Teixeira, J., Buckley, D. and García-París, M. (2006) Mitochondrial DNA phylogeography of *Lissotriton boscai* (Caudata, Salamandridae): evidence for old, multiple refugia in an Iberian endemic. *Molecular Ecology*, **15**, 3375-3388.

Matsumoto, K., Izri, A., Dumon, H., Raoult, D. and Parola, P. (2008) First detection of *Wolbachia* spp., including a new genotype, in sand flies collected in Marseille, France. *Journal of Medical Entomology*, **45**, 466-469.

Mayr, E. (1942) Systematics and the origin of species. New York: Columbia University Press.

Mayr, E. (1963) Animal species and evolution. Cambridge, MA: Belknap Press.

Mbow, M.L., Bleyenberg, J.A., Hall, L.R., Titus, R.G. (1998) *Phlebotomus papatasi* sand fly salivary gland lysate down-regulates a Th1, but up-regulates a Th2, response in mice infected with *Leishmania major*. *Journal of Immunology*, **161**, 5571-5577.

McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*, **351**, 652-654.

Miller, M.P. (2005) Alleles In Space: computer software for the joint analysis of interindividual spatial and genetic information. *Journal of Heredity*, **96**, 722-724.

Milleron, R.S., Mutebi, J.-P., Valle, S., Montoya, A., Yin, H., Soong, L. and Lanzaro, G. (2004a) Antigenic diversity in maxadilan, a salivary protein from the sand fly vector of American visceral leishmaniasis. *American Journal of Tropical Medicine and Hygiene*, **70**, 286-293.

Milleron, R.S., Ribeiro, J.M.C., Elnaiem, D., Soong, L. and Lanzaro, G. (2004b) Negative effect of antibodies against maxadilan on the fitness of the sand fly vector of American visceral leishmaniasis. *American Journal of Tropical Medicine and Hygiene*, **70**, 278-285.

Mills, S., Lunt, D.H. and Gómez, A. (2007) Global isolation by distance despite strong regional phylogeography in a small metazoan. *BMC Evolutionary Biology*, **7**, 225-234.

Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A.G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M.D., Sukernik, R.I., Olckers, A. and Wallace, D.C. (2003) Natural selection shaped regional mtDNA variation in humans. *Proceedings of the National Academy of Sciences of the USA*, **100**, 171-176.

Molina, R., Amela, C., Nieto, J., San-Andres, M., Gonzalez, F., Castillo, J.A., Lucientes, J. and Alvar, J. (1994) Infectivity of dogs naturally infected with *Leishmania infantum* to colonized *Phlebotomus perniciosus*. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **88**, 491-493.

Moore, R.C. and Purugganan, M.D. (2003) The early stages of duplicate gene evolution. *Proceedings of the National Academy of Sciences*, **100**, 15682-15687.

Moore, W.S. (1995) Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear gene trees. *Evolution*, **49**, 718-726.

Morens, D.M., Folkers, G.K. and Fauci, A.S. (2004) The challenge of emerging and re-

emerging infectious diseases. *Nature*, **430**, 242-249.

Morin, X. and Lechowicz, M.J. (2008) Contemporary perspectives on the niche that can improve models of species range shifts under climate change. *Biology Letters*, **4**, 573-576.

Morris, R.V., Shoemaker, C.B., David, J.R., Lanzaro, G.C. and Titus, R.G. (2001) Sandfly maxadilan exacerbates infection with *Leishmania major* and vaccinating against it protects against *L. major* infection. *Journal of Immunology*, **167**, 5226-5230.

Morton, B.R. (1993) Chloroplast DNA codon use: evidence for selection at the *psbA* locus based on tRNA availability. *The Journal of Molecular Evolution*, **37**, 273-280.

Naderi, S., Rezaei, H.-R., Taberlet, P., Zundel, S., Rafat, S.-A., Naghash, H.-R., El-Barody, M.A.A., Ertugrul, O. and Pompanon, F. (2007) Large-scale mitochondrial DNA analysis of the domestic goat reveals six haplogroups with high diversity. *PLoS ONE*, **2**: e1012.

Nagy, M., Entz, P., Otremba, P., Schoenemann, C., Murphy, N. and Dapprich, J. (2007) Haplotype-specific extraction: a universal method to resolve ambiguous genotypes and detect new alleles - demonstrated on HLA-B. *Tissue Antigens*, **69**, 176-180.

Naucke, T.J. and Schmitt, C. (2004) Is leishmaniasis becoming endemic in Germany? *International Journal of Medical Microbiology Supplements*, **293**, 179-181.

Nei, M. (1987) *Molecular Evolutionary Genetics*. New York: Columbia University Press.

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

Nelson, G., and Platnick, N.I. (1981) Systematics and biogeography: Cladistics and vicariance. New York: Columbia University Press.

Nichols, R.A. and Hewitt, G.M. (1994) The genetic consequences of long distance dispersal during colonization. *Heredity*, **72**, 312-317.

Nielsen, R. (2001) Statistical tests of selective neutrality in the age of genomics. *Heredity*, **86**, 641-647.

Nielsen, R. (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197-218.

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

Nogueira, F.S., Moreira, M.A., Borja-Cabrera, G.P., Santos, F.N., Menz, I., Parra, L.E., Xu, Z., Chu, H.J., Palatnik-de-Sousa, C.B. and Luvizotto, M.C. (2005) Leishmune blocks the transmission of canine visceral leishmaniasis: absence of Leishmania parasites in blood, skin and lymph nodes of vaccinated exposed dogs. *Vaccine*, **23**, 4805-4810.

Nylander, J.A.A. (2004) MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University. Available from: http://www.abc.se/~nylander/.

Ohno S. (1970) *Evolution by Gene Duplication*. Berlin: Springer-Verlag.

Oliveira, F., Kamhawi, S., Seitz, A.E., Pham, V.M., Guigal, P.M., Fischer, L., Ward, J. and Valenzuela, J.G. (2006) From transcriptome to immunome: identification of DTH inducing proteins from a *Phlebotomus ariasi* salivary gland cDNA library. *Vaccine*, **24**, 374-390.

Oliveira, F., Lawyer, P.G., Kamhawi, S. and Valenzuela, J.G. (2008) Immunity to distinct sand fly salivary proteins primes the anti-*Leishmania* immune response towards protection or exacerbation of disease. *PLoS Neglected Tropical Diseases*, **2**: e266.

O'Loughlin, S.M., Somboon, P., Walton, C. (2007) High levels of population structure caused by habitat islands in the malarial vector *Anopheles scanloni*. *Heredity*, **99**, 31-40.

Olson, S. (2002) Seeking the signs of selection. *Science*, **298**, 1324-1325.

Palatnik-de-Sousa, C.B. (2008) Vaccines for leishmaniasis in the fore coming 25 years. *Vaccine*, **26**, 1709-1724.

Palit, A., Bhattacharya, S.K. and Kundu, S.N. (2005) Host preference of *Phlebotomus argentipes* and *Phlebotomus papatasi* in different biotopes of West Bengal, India. *International Journal of Environmental Health Research*, **15**, 449-454.

Parvizi, P and Ready, P.D. (2006) Molecular investigation of the population differentiation of *Phlebotomus papatasi*, important vector of *L. major*, in different habitats and regions of Iran. *Iranian Biomedical Journal*, **10**, 69-77.

Parvizi, P. and Assmar, M. (2007) Nuclear elongation factor-1α a molecular marker for Iranian sandly identification. *Iranian Journal of Public Health*, **36**, 25-37.

Parvizi, P., Benlarbi, M. and Ready, P.D. (2003) Mitochondrial and *Wolbachia* markers for the sandfly *Phlebotomus papatasi*: little population differentiation between peridomestic sites and gerbil burrows in Isfahan province, Iran. *Medical and Veterinary Entomology*, **17**, 351–362.

Patz, J.A., Graczyk, T.K., Geller, N. and Vittor, A.Y. (2000) Effects of environmental change on emerging parasitic diseases. *International Journal for Parasitology*, **30**, 1395-1405.

Peakall, R., Ruibal, M. and Lindenmayer, D.B. (2003) Spatial autocorrelation analysis offers new insights into gene flow in the Australian bush rate, *Rattus fuscipes*. *Evolution*, **57**, 1182-1195.

Peakall, R. and Smouse, P. (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* (*electronic*), **6**, 288-295.

Penny, D. (2005) Evolutionary biology: relativity for molecular clocks. *Nature*, **436**, 183-184.

Perrotey, S., Mahamdallie, S.S., Pesson, B., Richardson, K.J., Gállego, M. and Ready, P.D. (2005) Postglacial dispersal of *Phlebotmus perniciosus* into France. *Parasite*, **12**, 283-291.

Pesson, B., Ready, J.S., Benabdennbi, I., Martín-Sánchez, J., Esseghir, S., Cadi-Soussi, M., Morillas-Márquez, F. and Ready, P.D. (2004) Sandflies of the *Phlebotomus perniciosus* complex: mitochondrial introgression and a new sibling species of *P. longicuspis* in the Moroccan Rif. *Medical and Veterinary Entomology*, **18**, 25-37.

Peterson, T.A. and Shaw, J. (2006) *Lutzomyia* vectors for cutaneous leishmaniasis in Southern Brazil: ecological niche models, predicted geographic distributions, and climate change effects. *International Journal for Parasitology*, **33**, 919-931.

Petit, R.J., Aguinagalde, I., de Beaulieu, J.-L., Bittkau, C., Brewer, S., Cheddadi, R., Ennos, R., Fineschi, S., Grivet, D., Lascoux, M., Mohanty, A., Müller-Starck, G., Demesure-Mulsch, B., Palmé, A., Martin, J.P., Rendell, S. and Vendramin, G.G. (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*, **300**, 1563-1565.

Pinho, C., Harris, D.J. and Ferrand, N. (2007) Contrasting patterns of population subdivision and historical demography in three western Mediterranean lizard

species inferred from mitochondrial DNA variation. *Molecular Ecology*, **16**, 1191-1205.

Pons, J., Barraclough, T.G., Gomez-Zurita, J., Cardoso, A., Duran, D.P., Hazell S., Kamoun, S., Sumlin, W.D. and Vogler, A.P. (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, **55**, 595-609.

Posada, D. and Crandall, K.A. (2001) Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology and Evolution*, **16**, 37-45.

Pratlong, F., Rioux, J.A., Marty, P., Faraut-Gambarelli, F.F., Dereure, J., Lanotte, G. and Dedet, J.P. (2004) Isoenzymatic analysis of 712 strains of *Leishmania infantum* in the south of France and relationship of enzymatic polymorphism to clinical and epidemiological features. *Journal of Clinical Microbiology*, **42**, 4077-4082.

Pritchard, J.K., Stephens, M. and Donnelly, P.J. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945-995.

Pritchard, J.K., Wen, X. and Falush, D. (2009) Documentation for *structure* software: version 2.3. Available from: http://pritch.bsd.uchicago.edu/structure.html.

Prugh, L.R., Hodges, K.E., Sinclair, A.R.E. and Brashares, J.S. (2008) Effect of habitat area and isolation on fragmented animal populations. *Proceedings of the National Academy of Sciences*, **105**, 20770-20775.

Queller, D.C. and Goodnight, K.F. (1989) Estimating relatedness using genetic markers. *Evolution*, **43**, 258-275.

Rambaut, A. (2009) FigTree v1.2.2. Available from: http://tree.bio.ed.ac.uk /software/figtree.

Rambaut, A. and Drummond, A.J. (2007) Tracer v1.4. Available from: http://beast.bio.ac.uk/Tracer.

Ramírez-Soriano, A., Ramos-Onsins, S., Rozas, J., Calafell, F. and 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. and Rozas, J. (2002) Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution*, **19**, 2092-2100.

Rand, D.M. and Kann, L.M. (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among gene from *Drosophila,* mice, and humans. *Molecular Biology and Evolution*, **13**, 735-748.

Rannala, B. and Mountain, J.L. (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of the USA*, **94**, 9197-9201.

Raymond, M. and Rousset, F. (1995) GENEPOP (Version1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248-249.

Ready, P.D. (2000) Sand fly evolution and its relationship to *Leishmania* transmission. *Memorias do Instituto Oswaldo Cruz*, **95**, 589-590.

Ready, P.D. (2008) Leishmaniasis emergence and climate change. *Revue Scientifique et Technique*, **27**, 399-412.

Ready, P.D. and Croset, H. (1980) Diapause and laboratory breeding of *Phlebotomus perniciosus* Newstead and *Phlebotomus ariasi* Tonnoir (Diptera: Psychodidae) from southern France. *Bulletin of Entomological Research*, **70**, 511-523.

Ready, P.D., Lainson, R., Shaw, J.J. and Souza, A.A. (1991) DNA probes for distinguishing *Psychodopygus wellcomei* from *Psychodoygus complexus* (Diptera: Psychodidae). *Memorias do Instituto Oswaldo Cruz*, **86**, 41-49.

Ribeiro, J.M.C. (1987a) Role of saliva in blood-feeding by arthropods. *Annual Review of Entomology*, **32**, 463-478.

Ribeiro, J.M.C. (1987b) Vector salivation and parasite transmission. *Memorias do Instituto Oswaldo Cruz*, **82** (suppl. III), 1-3.

Ribeiro, J.M.C. (1995) Blood-feeding arthropods: live syringes or invertebrate pharmacologists? *Infectious Agents and Disease*, **4**, 143-152.

Ribeiro, J.M.C. and Franscischetti, I.M.B. (2003) Role of arthropod saliva in blood feeding: sialome and post-sialome perspectives. *Annual Review of Entomology*, **48**, 73-88.

Ribeiro, J.M.C., Modi, G.B. and Tesh, R.B. (1989) Salivary apyrase activity of some Old World phlebotomine sand flies. *Insect Biochemistry*, **19**, 409-412.

Ribeiro, J.M.C., Rossignol, P.A. and Spielman A. (1986) Blood-finding strategy of a capillary-feeding sandfly, *Lutzomyia longipalpis*. *Comparative Biochemistry and Physiology - Part A: Molecular & Integrative Physiology*, **83**, 683-686.

Ribeiro J.M.C., Sarkis, J.J.F., Rossignol, P.A. and Spielman, A. (1984) Salivary of *Aedes aegypti*: characterization and secretory fate. *Comparative Biochemistry and Physiology - Part B: Biochemistry & Molecular Biology*, **79**, 81-86.

Ribera, I. and Vogler, A.P. (2004) Speciation of Iberian diving beetles in Pleistocene refugia (Coleoptera, Dytiscidae). *Molecular Ecology*, **13**, 179-193.

Rice, W.R. (1989) Analyzing tables of statistical tests. *Evolution*, **43**, 223-225.

Riou, S. (2004) The distribution and abundance of *Phlebotomus ariasi* (Diptera: Psychodidae) in southern France: a bioclimatic and molecular approach. Thesis (M.Sc.). Imperial College, London.

Rioux, J.-A. and Golvan, Y.J. (1969) Epidémiologie des leishmanioses dans le sud de la France. *Monographie de l'institut national de la santé et de la recherche médicale*, Paris, n° 37.

Rioux, J.-A., Golvan, Y.J., Croset, H., Houin, R., Juminer, B., Bain, O. and Tour, S. (1967) Écologie des leishmanioses dans le sud de la France 1. Les Phlébotomes. Échantillonnage-éthologie. *Annales de Parasitologie Humaine et Comparée*, **42**, 561-603.

Rioux, J.-A., Killick-Kendrick, R., Perieres, J., Turner, D.P. and Lanotte, G. (1980) Ecology of leishmaniasis in the south of France. 13. Middle slopes of hillsides as sites at maximum of transmission of visceral leishmaniasis in the Cévennes. *Annales de Parasitologie Humaine et Comparée*, **55**, 445-453.

Rioux, J.-A., Rispail. P., Lanotte, G. and Lepart, J. (1984) Relations phlébotomes-bioclimats en écologie des leishmanioses. Corollaires épidémiologiques. L'exemple du Maroc. *Bulletin de la Société Botanique de France*, **131**, 549-557.

Rispail, P., Dereure, J. and Jarry, D. (2002) Risk zones of human leishmaniasis in the western Mediterranean Basin. Correlations between vector sand flies, bioclimatology and phytosociology. *Memorias do Instituto Oswaldo Cruz*, **97**, 477-483.

Rodríguez, F., Oliver, J.L., Marín, A. and Medina, J.R. (1990) The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, **142**, 485-501.

Rogers, A. R. (1995) Genetic evidence on modern human origins. *Human Biology*, **67**, 1-36.

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

Rogers, D.J. and Randolph, S.E. (2000) The Global Spread of Malaria in a Future, Warmer World. *Science*, **289**, 1763-1766.

Ronquist, F. and Huelsenbeck, J.P. (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572-1574.

Rousset, F. (1997) Genetic differentiation and estimation of gene flow from *F*- statistics under isolation by distance. *Genetics*, **145**, 1219-1228.

Rozas, J., Sánchez-Delbarrio, J.C., Messeguer, X. and Rozas, R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496-2497.

Sarkis, J.J.F., Guimarães, J.A. and Ribeiro, J.M.C. (1986) Salivary apyrase of *Rhodnius prolixus*. *Biochemical Journal*, **233**, 885-891.

Sasai, N., Kato, Y., Kimura, G., Takeuchi, T. and Yamaguchi, M. (2007) The *Drosophila jumonji* gene encodes a JmjC-containing nuclear protein that is required for metamorphosis. *FEBS Journal*, **274**, 6139-6151.

Saunders, D.A., Hobbs, R.J. and Margules, C.R. (1991) Biological consequences of ecosystem fragmentation - a review. *Conservation Biology*, **5**, 18-32.

Sauvage, C. and Brignon, C. (1963) Etages bioclimatiques. *Atlas du Maroc, section II, carte 6b. Notice explicative par C. Sauvage*, Comité National de Géographie du Maroc, Rabat. pp. 1-44.

Schlein, Y., Jacobson, R.L. and Messer, G. (1992) *Leishmania* infections damage the feeding mechanism of the sandfly vector and implement parasite transmission by bite. *Proceedings of the National Academy of Sciences*, **89**, 9944-9948.

Schmidt, D. and Pool, J. (2002) The effect of population history on the distribution of the Tajima's D statistic. *In press*.

Schmitt, T. (2007) Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in Zoology*, **4**: 11.

Schmitt, T., Röber, S. and Seitz, A. (2005) Is the last glaciation the only relevant event for the present genetic population structure of the meadow brown butterfly *Maniola jurtina* (Lepidoptera: Nymphalidae)? *Biological Journal of the Linnean Society*, **85**, 419-431.

Schwalie, P.C. and Schultz, J. (2009) Positive selection in the tick saliva proteins of the Salp15 family. *Journal of Molecular Evolution*, **68**, 186-191.

Segarra-Moragues, J.G., Palop-Esteban, M., González-Candelas, F. and Catalán, P. (2007) *Nunatak* survival vs. *tabula rasa* in the Central Pyrenees: a study on the endemic plant species *Borderea pyrenacia* (Dioscoreaceae). *Journal of Biogeography*, **34**, 1893-1906.

Shortt, H. and Swaminath, C. (1928) The method of feeding of *Phlebotomus argentipes* with relation to its bearing on the transmission of kala azar. *Indian Journal of Medical Research*, **15**, 827-836.

Shriver, M.D. and Kittles, R.A. (2004) Genetic ancestry and the search for personalized genetic histories. *Nature Reviews Genetics*, **5**, 611-618.

Simmons, R.B., and S.J. Weller. (2001) Utility and evolution of cytochrome b in insects. *Molecular Phylogenetics and Evolution*, **20**, 196-210.

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

Slatkin, M. (1973) Gene flow and selection in a cline. *Genetics*, **75**, 733-756.

Slatkin, M. (1995) Gene flow and population structure. **In**: Real, L.A. (ed.) *Ecological genetics*. Princeton, NJ: Princeton University Press. pp3-17.

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

Smith, C.I. and Farrell, B.D. (2005) Range expansions in the flightless longhorn cactus beetles, *Moneilelma gigas* and *Moneilema armatum*, in response to Pleistocene climate changes. *Molecular Ecology*, **14**, 1025-1044.

Smouse, P.E. and Peakall, R. (1999) Spatial autocorrelation analysis of multi-allele and multi-locus genetic microstructure. *Heredity*, **82**, 561-573.

Sommer S.S., Groszbach, A.R. and Bottema, C.D. (1992) PCR amplification of specific alleles (PASA) is a general method for rapidly detecting known single-base changes. *Biotechniques*, **12**, 82-87.

Spielman, D., Brook, B.W. and Frankham, R. (2004) Most species are not driven to extinction before genetic factors impact them. *Proceedings of the National Academy of Sciences USA*, **101**, 15261-15264.

Spurgin, L.G. and Richardson, D.S. (2010) How pathogens drive genetic diversity:MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society of London B: Biological Sciences*, **277**, 979-988.

Stamatakis, A., Hoover, P. and Rougemont, J. (2008) "A fast bootstrapping Algorithm for the RaxML web-servers". *Systematic Biology*, **57**, 758-771.

Steininger F.F. and Rogl, F. (1984) Palaeogeography and palinspastic reconstruction of the Neogene of the Mediterranean and Paratethys. **In**: Dixon, J.E. and Robertson A.F.H. (eds.) *The Geological Evolution of the Eastern Mediterranean*. Oxford: Blackwell Scientific Publications.

Stephens, J.C., Briscoe, D. and O'Brien, S.J. (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *American Journal of Human Genetics*, **55**, 809-824.

Stephens, M., Smith, N., and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978-989.

Sudia, W.D. and Chamberland, R.W. (1962) Battery operated light trap, an improved model. *Mosquito News*, **22**, 126-129.

Sunnucks, P. (2000) Efficient genetic markers for population biology. *Trends in Ecology and Evolution*, **15**, 199-203.

Svobodová, M., Sádlová, J., Chang K.-P. and Volf, P. (2003) Short report: distribution and feeding preference of the sand flies *Phlebotomus sergenti* and *P. papatasi* in a cutaneous leishmaniasis focus in Sanliurfa, Turkey. *American Journal of Tropical Medicine and Hygiene*, **68**, 6-9.

Swofford, D.L. (2002) *PAUP. Phylogenetic Analysis Using Parsimony,* version 4.0b10. Washington DC: Smithsonian Institution Press.

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

Tajima, F. (1989) The effect of change in population size on DNA polymorphism. *Genetics*, **123**, 597-601.

Tamura, K., Dudley, J., Nei, M and Kumar, S. (2007) MEGA4:Molecular Evolutionary Genetic Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**, 1596-1599.

Testa, J.M., Montoya-Lerma, J., Cadena, H., Oviedo, M. and Ready, P.D. (2002) Molecular identification of vectors of *Leishmania* in Colombia: mitochondrial introgression in the *Lutzomyia townsendi* series. *Acta Tropical*, **84**, 205-218.

Tetteh, K.K.A., Stewart, L.B., Ochola, L.I.O., Amambua-Ngwa, A., Thomas, A.W., Marsh, K., Weedall, J.D. and Conway, D.J. (2009) Prospective identification of malaria parasite genes under balancing selection. *PLoS ONE*, **4**: e5568.

Titus, R.G., Bishop, J.V. and Mejia, J.S. (2006) The immunomodulatory factors of arthropod saliva and the potential for these factors to serve as vaccine targets to prevent pathogen transmission. *Parasite Immunology*, **28**, 131-141.

Trotz-Williams, L.A. and Trees, A.J. (2003) Systematic review of the distribution of the major vector-borne parasitic infections in dogs and cats in Europe. *The Veterinary Record*, **152**, 97-105.

Valenzuela, J.G., Belkaid, Y., Garfield, M.K., Mendez, S., Kamhawi, S., Rowton, E.D., Sacks, D.L. and Ribeiro, J.M. (2001a) Toward a defined anti-*Leishmania* vaccine targeting vector antigens: characterization of a protective salivary protein. *Journal of Experimental Medicine*, **194**, 331-342.

Valenzuela, J.G., Belkaid, Y., Rowton, E. and Ribeiro, J.M.C. (2001b) The salivary apyrases of the blood-sucking sand fly *Phlebotomus papatasi* belongs to the novel *Cimex* family of apyrases. *The Journal of Experimental Biology*, **204**, 229-237.

Valenzuela, J.G., Charlab, R., Galperin, M.Y. and Ribeiro, J.M.C. (1998) Purification, cloning, and expression of an apyrase from the bed bug *Cimex lectularius*. A new type of nucleotide-binding enzyme. *Journal of Biological Chemistry*, **273**, 30583-30590.

Valenzuela, J.G., Chuffe, O.M. and Ribeiro, J.M.C. (1996) Apyrase and anti-platelet activities from the salivary glands of the bed bug *Cimex lectularius*. *Insect Biochemistry and Molecular Biology*, **26**, 557-562.

Varga, Z.S. and Schmitt, T. (2008) Types of oreal and oreotundral disjunctions in the western Palearctic. *Biological Journal of the Linnean Society*, **93**, 415-430.

Volf, P. and Rohoušová, I. (2001) Species-specific antigens in salivary glands of phlebotomine sandflies. *Parasitology*, **122**, 37-41.

Vorou, R.M., Papavassiliou, V.G. and Tsiodras, S. (2007) Emerging zoonoses and vector-borne infections affecting humans in Europe. *Epidemiology and Infection*, **135**, 1231-1247.

Wade, T.G., Riitters, K.H., Wickham, J.D. and Jones, K.B. (2003) Distribution and causes of global forest fragmentation. *Conservation Ecology*, **7**: 7.

Wang, J. and Caballero, A. (1999) Developments in predicting the effective size of subdivided populations. *Heredity*, **82**, 212-226.

Warburg, A., Saraiva, E., Lanzaro, G.C., Titus, R.G. and Neva, F. (1994) Saliva of *Lutzomyia longipalpis* sibling species differs in its composition and capacity to enhance leishmaniasis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **345**, 223-230.

Wayne, M.L. and Simonsen, K.L. (1998) Statistical tests of neutrality in the age of weak selection. *Trends in Ecology and Evolution*, **13**, 236-240.

Weir, B.S. and Cockerham, C.C. (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **36**, 1358-1370.

Weir, J.T. and Schulter, D. (2008) Calibrating the avian molecular clock. *Molecular Ecology*, **17**, 2321-2328.

Wheeler, Q.D. and Platnick, N.I. (2000) The phylogenetic species concept (*sensu* Wheeler and Platnick). **In:** Wheeler, Q.D. and Meier, R. (eds.) *Species concepts and phylogenetic theory: A debate*. New York: Columbia University Press.

Willmann, R. and Meier, R. (2000) A critique from the Hennigian Species Concept Perspective. **In:** Wheeler, Q.D. and Meier, R. (eds.) *Species Concepts and Phylogenetic Theory*. New York: Columbia University Press. pp. 101-108.

World Health Organisation. (2002) The world health report, Geneva. pp. 192-197.

World Health Organisation. (2010a) - Burden of disease. Available from: http://www.who.leishmaniasis/burden/en/.

World Health Organisation. (2010b) - Leishmania and HIV co-infection. Available from: http://www. who.leishmaniasis/burden/hivcoinfection/burdenhivcoinfection/en/.

Wright, F. (1990) The 'effective number of codons' used in a gene. *Gene*, **87**, 23-29.

Wright, S. (1921) Systems of mating. *Genetics*, **6**, 111-178.

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

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

Wright, S. (1946) Isolation by distance under diverse systems of mating. *Genetics*, **31**, 39-59.

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

Wright, S. (1978) *Evolution and the Genetics of Populations*, Vol. IV. Variability within and among natural populations. Chicago: University of Chicago Press.

Yang, M. and Kirley, T.L. (2004) Site-directed mutagenesis of human soluble calcium-activated nucleotidase 1 (hSCAN-1): Identification of residues essential for enzyme activity and the Ca(2+)-induced conformational change. *Biochemistry*, **43**, 9185-9194.

Yang, Z. (2002) Inference of selection from multiple species alignments. *Current Opinion in Genetics and Development*, **12**, 688-694.

Yang, Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586-1591.

Yang, Z., and Swanson, W.J. (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Molecular Biology and Evolution*, **19**, 49-57.

Yang, Z., Wong, W.S.W. and Nielsen, R. (2005) Bayes Empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*, **22**, 1107-1118.

Young, R., Taylor, J.E., Kurioka, A., Becker, M., Louis, E.J. and Rudenko, G. (2008) Isolation and analysis of the genetic diversity of repertoires of VSG expression site containing telomeres from *Trypanosoma brucei gambiense*, *T. b. brucei* and *T. equiperdum*. *BMC Genomics*, **9**: 385.

Zhang, P., Gu, Z. and Li, W-H. (2003) Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biology*, **4**: R56.

Zhao, Z., Fu, Y.X., Hewett-Emmett, D. and Boerwinkle, E. (2003) Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. *Gene*, **312**, 207-213.

APPENDICES

Appendix 2.1 DNA extraction protocol

Sandfly tissue was ground using the tip of a sterile plastic pipette in 10 µl of 10x Tris grinding buffer (0.1M Tris-HCl pH 7.5; 0.6M NaCl; 0.1M EDTA; 0.15mM spermine; 0.15mM spermidine; 5% (w/v) sucrose). A further 90 µl of 10x Tris grinding buffer was added together with 10 µl of 2x sodium SDS buffer (0.3M Tris-HCl pH9.0; 0.1M EDTA; 5% (w/v) sucrose; 1.22% (w/v) SDS; 0.34% (v/v) diethylpyrocarbonate). This solution was mixed by gentle tapping, followed by a pulse vortex and incubated at 65°C for 45 min to lyse cells and denature proteins. After cooling, 30 µl of ice-cold 8M KOAc was added, the solution pulse vortexed and left on ice for 45 min to remove the SDS from the solution. Proteins etc were pelletted at 14 Kr.p.m for two min from the remaining DNA supernatant. DNA was precipitated overnight from the supernatant at -20°C by the addition of 350 µl 96-100% ethanol. The following morning DNA was pelleted by centrifugation for 30 min at 14 Kr.p.m and washed three times in 500 µl 70% ethanol: vortexing and centrifuging for 5 min and decanting off the ethanol between each wash. The final cleaned pellet was dried under vacuum for 10 min and the DNA re-suspended in 15 to 25 µl of 1 x TE. DNA extract was placed at 4°C and -20°C, for short- and long-term storage, respectively.

Appendix 2.2 PCR product purification protocol

(i) GENECLEAN® II: submerged agarose gel horizontal electrophoresis was used to purify the amplified PCR product by size separation from primer dimers and/or secondary non-specific products. The entire PCR reaction volume was loaded with 5 µl of 6x Orange G DNA loading buffer, into wells on either a 1.5% (Cyt b and EF-1α) or 2% (AAm20 and AAm24) agarose gel (agarose electrophoresis grade of Invitrogen™ Corporation dissolved in 1x TBE) and stained with 0.01% ethidium bromide (Fisher Scientific Inc.) for DNA visualization under ultra-violet (UV) light. Electrophoresis was conducted at 80 V for 1 hour. An image of the gel exposed to UV light using a GeneGenius documentation system allowed the identification of the correct DNA band (by size) and an approximate estimation of DNA concentration per specimen, both

calibrated against Promega PCR Markers (6 bands between 50 to 1000 bp) or Bioline Hyperladder™ IV (10 bands between 100 to 1000 bp).

Excised DNA bands (where the razor blade was cleaned between cutting each band to avoid carry-over) where purified by binding to GLASSMILK® (GENECLEAN® II Spin Kit, BIOL 101 Qbiogene, Inc). DNA band volume was calculated and 0.5x band volume of TBE modifier and 4.5x band volume of NaI where added to a 1.5 ml Eppendorf tube containing the band, and incubated at 55°C for 5 min or until all the agarose band had dissolved. When DNA band plus NaI total volume was < 500 µl or between 500 to 1000 ul, 5 µl or 7.5 µl of GLASSMILK® (from kit), respectively, was added to the Eppendorf tube, where the solution was then rotated at room temperature for 10 min: DNA is drawn out of the solution and binds to the silica matrix of the GLASSMILK®. Samples were then centrifuged at 13.2 Kr.p.m. for 20 sec to pellet the GLASSMILK® and bound DNA. Supernatant was discarded. The pellet was re-suspended and cleaned by washing three times in 500 µl of NEW Wash solution (from kit), between each wash DNA was re-pelletted by centrifugation at 13.2 Kr.p.m. for 20 sec. After the final aspiration of supernatant, the pellet was allowed to dry at room temperature for approximately 10 min to evaporate all ethanol: ethanol can interfere during downstream stages, i.e. sequencing. Finally, the pellet was re-suspended by gentle tapping in PCR grade water to give a final DNA concentration of 1-2 ng/100 bp of target product in 5 µl (accounting for a 20% loss). The solution was incubated at 55°C for 10 min. The supernatant containing DNA was aspirated from the GLASSMILK® pellet following centrifugation at 13.2 Kr.p.m. for 1 min. Purified DNA was stored at -20°C for sequencing.

(ii) Millipore MultiScreen® PCR_{96} Filter Plates: 4 µl of PCR product was loaded onto an agarose gel where horizontal electrophoresis was run at 80 V for 1 hour to assess the success of the PCR reaction per specimen. The remaining product was purified by the Millipore filter plate method when a single DNA band of the correct fragment size and having product yield of at least 2 ng per 100 bp in the remaining volume was observed on the gel post-electrophoresis; calibrated against the a PCR marker (Promega Corporation PCR Markers or Bioline Hyperladder™ IV).

The manufacturer's protocol of purification by Millipore MultiScreen® PCR_{96} Filter Plates was optimized by the author, where adjusted methods for fragment size were followed. The following details the protocol for 'long' fragments (> 300 bp), numbers in square brackets indicate how the protocol was adjusted for 'short' fragments

(< 300 bp). PCR reaction product was made-up to 100µl [200 µl] with PCR grade water (Sigma), and mixed by pipetting up and down before loading into a well of a Millipore MultiScreen® PCR_{96} Filter Plate whose membrane had been previously wetted using 40 µl of PCR grade water. The filter plate was then placed on a manifold under vacuum at 500mBar (14.8 inches Hg) [250mBar (7.4 inches Hg). The low pressure ensured that small fragment product loss was minimal] until all the solution had filtered through. Each well was washed with 200 µl [100 µl] PCR grade water and again place under vacuum at 550mBar (16.2 inches Hg) [200mBar (6 inches Hg)] until all the solution had filtered through. Samples were reconstituted into 50 µl of PCR grade water and placed on to an automatic shaker for 10 min [15 min] which aided to lift the DNA from the membrane. For short fragments a 15 min "preincubation" reconstitution step was applied where the filter plate was left on the bench at room temperature before placing on to an automatic shaker. The DNA solution was recovered by aspiration into clean tubes or plate.

Appendix 2.3 Table Sequences of *Phlebotomus* species used to reconstruct phylogenies and determine outgroups to *P. ariasi* for population genetic analyses. AF = Africa; ER = Europe; MD = Mediterranean; ME = Middle East.

Subgenus	Species	Locus	Species code	Area	Country	* GenBank Accession number; # haplotype code in phylogenies
Paraphlebotomus	*Phlebotomus caucasicus*	Cyt b	cauc	ME	Iran	FJ217389*
Phlebotomus	*Phlebotomus papatasi*	Cyt b	papa	MD	Tunisia	AF161214*
Adlerius	*Phlebotomus halepensis*	Cyt b	hale	ME	Iran	CB01# (HQ023283); CB02# (HQ023284); CB03# (HQ023285)
		EF-1α	hale	ME	Iran	EF416842*
	Phlebotomus brevis	Cyt b	brev	ME	Iran	CB01# (HQ023282)
Transphlebotomus	*Phlebotomus mascittii*	Cyt b	masc	MD	France	CB01# (HQ023281)
		EF-1α	masc	MD	France	EF01# (HQ025987), EF02# (HQ025988)
		AAm20	masc	MD	France	20m01# (HQ026018)
		AAm24	masc	MD	France	24m01# (HQ026000)
Larroussius	*Phlebotomus ariasi*	Cyt b	aria	MD	France	CBx#
		EF-1α	aria	MD	France	AF160803*; EFx#
		AAm20	aria	MD	France	20mx#
		AAm24	aria	MD	France	24mx#
	near *Phlebotomus ariasi*	Cyt b	nraria	MD	Tunisia	AF161196*
		EF-1α	nraria	MD	Tunisia	AF160804*
	Phlebotomus neglectus	Cyt b	negl	MD; ER	Greece; Hungary	AF161188*; AF161189*; AF161190*; AF161191*; CB01# (HQ023286)
		EF-1α	negl	MD	Greece	AF160801*
	near *Phlebotomus neglectus*	EF-1α	nrnegl	MD	Greece	AF160802*
	Phlebotomus major	Cyt b	majo	ME	Iran	CB01# (HQ023286)
		EF-1α	majo	ME	Iran	EF416834a,b*
	Phlebotomus perfiliewi	Cyt b	perf	MD	Tunisia	AF161197*; AF161198*; AF161199*; AF161200*
		EF-1α	perf	MD	Tunisia	AF160805a,b,c,d*
	Phlebotomus perniciosus	Cyt b	pern	MD	Spain; France	AF161205*; 01_PDR_CB# (HQ023288)
		EF-1α	pern	MD	Spain	AF160807*
		AAm20	pern	MD	Spain	AJ303377*
		AAm24	pern	MD	Spain	AJ303378*
	Phlebotomus tobbi	Cyt b	tobb	MD	Greece	AF161209* AF161210*; AF161211*; AF161212*
		EF-1α	tobb	ME	Iran	EF416833*
	Phlebotomus orientalis	Cyt b	orie	AF	Ethiopia	AF161202*; AF161203*
		EF-1α	orie	AF	Ethiopia	AF160806*
	Phlebotomus langeroni	Cyt b	lang	MD	Tunisia, Egypt	AF161207*; AF161208*
		EF-1α	lang	MD	Tunisia	AF160809*
	Phlebotomus longicuspis	Cyt b	long	MD	Tunisia	AF161206*
		EF-1α	long	MD	Tunisia	AF160808*

Appendix 2.4 Table Parameters of models used for Bayesian estimation to reconstruct phylogenies using locus cyt b. Bayesian analysis no. is referred to in the main text. Outgroup *Phlebotomus* species codes: papa = *P. papatasi*; cauc = *P. caucasicus*.

Bayesian analysis no.	Out-group	Codon partition	Substitution model (no. variable sites)	Bayes factor	Hypothesis tested
Cyt b_bayes1	papa & cauc	1 ≠ 2 ≠ 3	HKY + I (65) ≠ HKY + I (20) ≠ GTR + I + G (190)	-4262.10	Testing the effect of partitioning.
Cyt b_bayes2	papa & cauc	1 + 2 ≠ 3	GTR + I + G (85) ≠ GTR + I + G (190)	-4349.48	Models from MRMODELTEST.
Cyt b_bayes3	papa & cauc	None	GTR + I + G (275)	-4553.92	
Cyt b_bayes4	papa	1 ≠ 2 ≠ 3	HKY + I (64) ≠ HKY + I (19) ≠ GTR + I + G (182)	NA	Testing the effect of outgroup choice.
Cyt b_bayes5	cauc	1 ≠ 2 ≠ 3	HKY + I (64) ≠ HKY + I (17) ≠ GTR + I + G (186)	NA	Models from MRMODELTEST.
Cyt b_bayes6	papa & cauc	1 ≠ 2 ≠ 3	GTR + I + G (65) ≠ GTR + I + G (20) ≠ GTR + I + G (190)	-4272.99	Testing the effect of substitution model.
Cyt b_bayes11	papa & cauc	1 ≠ 2 ≠ 3	GTR + I + G (64) ≠ HKY + I (17) ≠ GTR + I + G (189)	NA	Testing the effect of 'ingroup' data level. Models from MRMODELTEST.

Appendix 2.5 Figure A multi-species alignment of AAm20 alleles, showing variable base positions and indels (-).

```
                                                                                                                    1
                       1 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990
              1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
masc_20m01    TGAAGGAGTC TCCGGACGAG GAGGAAGAGG AGGAGGTGGA TGAGGCACCA GAGCTGCCGA CGCATTTCAC C--------- TATCGCCCTC CGTCCACCAC
pern_AJ303377 .......... AGTC------ .....G.... ......---. ......C..T ...T...T.. .A.GC..... .CAACTCAAC .....G..C. .A..GGTTC.
20m01         .......... CG.C..G... .......... ......---. ......T..T ........C. ...GC..... TCAGCTCAAC .......... .A...GTTC.
20m02         .......... CG.C..G... .......... ......---. ......T..T .....A..C. ...GC..... TCAGCTCAAC .......... .A...GTTC.
20m03         .......... CG.C..G... .......... ......---. ......T..T ........C. ...GC..... TCAGCTCAAC .......... .A..TGTTC.
20m04         .......... CG.C..G... .......... ......---. ......T..T ........C. ...GC..... TCAGCTCAAC .......... .A...GTTC.
20m05         .......... CG.C..G... ..A....... ......---. ......T..T ........C. ...GC..... TCAGCTCAAC .......... .A...GTTC.
20m06         .......... CG.C..G... .......... ......---. ......T..T ........C. ....C..... TCAGCTCAAC .......... .A...GTTC.
20m07         .......... CG.C..G... .......... ......---. ......T..T .....A..C. ...GC..... TCAGCTCAAC .......... .A...GTTC.
20m08         .......... CG.C..G... .......... ......---. ......T..T .....A..C. ...GC..... TCAGCTCAAC .......... .A...GTTC.
20m09         .......A.. CG.C..G... ..A....... ......---. ......T..T .....A..C. ...GC..... TCAGCTCAAC .......... .A...GTTC.
20m10         .......... CG.C..G... .......... ......---. ......T..T .....A..C. ...GC..... TCAGCTCAAC .......... .A...GTTC.
20m11         .......... CG.C..G... .......... ......---. ......T..T .....A..C. ...GC..... TCAGCTCAAC .......... .A.G.GTTC.
20m12         .......... CG.C..G... .......... ......---. ......T..T .....A..C. ...GC..... TCAGCTCAAC ........C. .A...GTTC.
20m13         .......... CG.C..G... ..A....... ......---. ......T..T .....A..C. ...GC..... TCAGCTCAAC .......... .A...GTTC.

              1111111111 1111111111 1111111111 1111111111 111111111
              0000000001 1111111112 2222222223 3333333334 444444444
              1234567890 1234567890 1234567890 1234567890 123456789
masc_20m01    AAGGCCCCCG AGCAGCCTGA GTCGTCGCTC TCTGGGCGGA TACTCAA--
pern_AJ303377 C..A..A..C .....T.... .......... CT........ .......GT
20m01         C.....A..C .......... .......... .......... ...G...GC
20m02         C.....A..C .......... .......... .......... ...G...GC
20m03         C.....A..C .......... .......... .......... ...G...GC
20m04         C.....G..C ........A. .......... .T....T... ...G...GC
20m05         C.....A..C .......... .......... .......... ...G...GC
20m06         C.....A..C .......... .......... .......... ...G...GC
20m07         C.....G..C ........A. .......... ......T... ...G...GC
20m08         C.....A..C .......... .......... ......T... ...G...GC
20m09         C.....A..C .......... .......... .......... ...G...GC
20m10         C.....A..C .......... .......... .T........ ...G...GC
20m11         C.....A..C .......... .......... .......... ...G...GC
20m12         C.....A..C .......... .......... .......... ...G...GC
20m13         C.....A..C .......... .......... .......... ...G...GC
```

P. mascittii (masc; GenBank accession HQ026018)
P. perniciosus (pern_GenBank accession)
P. ariasi (20mNN; GenBank accessions HQ026001-HQ026013)

Appendix 2.6 Figure A multi-species alignment of AAm24 alleles, showing variable base positions and indels (-); *P. mascittii* (masc; Genbank accession HQ026000), *P. perniciosus* (pern_GenBank accession), *P. ariasi* (24mNN; Genbank accessions HQ025989-HQ025999).

```
                          1 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889
                 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
masc_24m01       GGTGTCGGTG ACGAAGCAA- -----CCGCC GTCGCCGCAA ---GTGCAAC AAGTTCAGCA GCAACCGTCG CAGCAGCAG- ----------
pern_AJ303378    .......... ........GC AACAA..A-- -A....A... CAA....... ........-- ----T..... .........C AGCAGCAACA
24m01            ......A... .........- -----..A-- -A....A... ---....... ........-- ----...... ........AC AG--------
24m02            ......A... .........- -----..A-- -A....A... ---....... ........-- ----...... .....A..AC AG--------
24m03            ......A... .........- -----..A-- -A....A... ---....... ........-- ----...... .....A..AC AG--------
24m04            ......A... .........- -----..A-- -A....A... ---....... ........-- ----...... .....A..AC AG--------
24m05            ......A... .........- -----..A-- -A....A... ---....... ........-- ----...... ........AC AG--------
24m06            ......A... .........- -----..A-- -A....A... ---....... ........-- ----...... ........AC AG--------
24m07            ......A..A .........- -----..A-- -A....A... ---....... ........-- ----...... ........AC AG--------
24m08            ......A... .........- -----..A-- -A....A... ---....... ........-- ----...... .........C AG--------
24m09            ......A... .........- -----G.A-- -A....A... ---....... ........-- ----...... ........AC AG--------
24m10            ......A..A .........- -----..A-- -A....A... ---....... ........-- ----...... ........AC AG--------
24m11            ......A... .........- -----..A-- -A....A... ---A...... ........-- ----...... ........AC AG--------

                          1 1111111111 1111111111 1111111111 1111111111 1111111111 1111111111 1111111111 11111
                 9999999990 0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 77777
                 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 12345
masc_24m01       ---------- ---------C ACTATGTTCA CGTGCAGAGC AAATCCAAAG TCTCCTCACC GGCGCCATCG CACATCTACG GGAAG
pern_AJ303378    GCAACAACAA CAGCAACAG. .T........ T......... ..G..T.... .G..G..G.. ......G..A ..T..A..T. .....
24m01            ---------- ---------. .T........ T......... ..G..T.... .A..G..G.. .......... ..T..A..T. .....
24m02            ---------- ---------. .T........ T......... ..G..T.... .A..G..G.. .......... ..T..T..T. .A...
24m03            ---------- ---------. .T........ T......... ..G..T.... .A..G..G.. .......... ..T..A..T. .A...
24m04            ---------- ---------. .T........ T......... ..G..T.... .A..G..G.. .......... ..T..A..T. .....
24m05            ---------- ---------. .T........ T.....A... ..G..T.... .A..G..G.. .......... ..T..A..T. .....
24m06            ---------- ---------. .T........ T......... ..G..T.... .A..G..G.. .......... ..T..A..T. ....A
24m07            ---------- ---------. .T........ T......... ..G..T.... .A..G..G.. .......... ..T..A..T. .....
24m08            ---------- ---------. .T........ T......... ..G..T.... .A..G..G.. .......... ..T..A..T. ....A
24m09            ---------- ---------. .T........ T......... ..G..T.... .A..G..G.. .......... ..T..A..T. .....
24m10            ---------- ---------. .T........ T......... ..G..T.... .A..G..G.. .......... ..T..A..T. ....A
24m11            ---------- ---------. .T........ T......... ..G..T.... .A..G..G.. .......... ..T..A..T. ....A
```

Appendix 2.7 Figure Alignment of the 96 variable base positions from 119 haplotypes amplified for all populations of *P. ariasi* at locus cyt b (including the 3' IgS and tRNA). The most frequent haplotype CB25 is used as the reference sequence and given in full. For population genetic analyses a 738 bp fragment was analysed as missing data was removed; the 5' 7 bp. GenBank AF161194 (*P. ariasi*) begins on base position 2 of CB25.

```
                1 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990
     1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
CB25 CTTCTGAGGA GCAACAGTAA TTACAAATTT ACTATCTGCT ATCCCTTATT TAGGAAATAT ATTAGTTCAA TGAATCTGAG GAGGATTTGC TGTTGATAAT

     1111111111 1111111111 1111111111 1111111111 1111111111 1111111111 1111111111 1111111111 1111111111 1111111112
     0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990
     1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
CB25 GCAACTCTAA CACGATTTTT CACCTTTCAT TTTTTATTTC CATTTATTAT TGCTGCTATA ACTATAATCC ATTTATTATT TTTACATCAA ACAGGATCAA

     2222222222 2222222222 2222222222 2222222222 2222222222 2222222222 2222222222 2222222222 2222222222 2222222223
     0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990
     1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
CB25 ATAATCCTTT AGGATTAAAT AGTAATTCAG ACAAAATTCC ATTTCATCCT TACTTTTCAT TCAAAGATAT TATTGGATTT ATTATTATAA TTATAATTCT

     3333333333 3333333333 3333333333 3333333333 3333333333 3333333333 3333333333 3333333333 3333333333 3333333334
     0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990
     1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
CB25 CTCTATTTTA TCTATTATAG CCCCATATTA TTTAGGAGAT CCAGATAATT TTATTCCAGC AAATCCATTA GTAACCCCCC CTCATATTCA ACCTGAATGA

     4444444444 4444444444 4444444444 4444444444 4444444444 4444444444 4444444444 4444444444 4444444444 4444444445
     0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990
     1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
CB25 TATTTTTTAT TTGCCTATGC TATTTTACGT TCTATCCCCA ATAAATTAGG AGGTGTAATT GCACTTGTTA TATCAATTGC AATCCTATTT ATTCTTCCTA

     5555555555 5555555555 5555555555 5555555555 5555555555 5555555555 5555555555 5555555555 5555555555 5555555556
     0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990
     1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
CB25 TTTTACATGT AAATAAATCT CAAGGATTAC AATTTTATCC TCTTAATCAA ATCTTATTTT GATATATAGT TATTATTATT ATTTTATTAA CATGAATTGG

     6666666666 6666666666 6666666666 6666666666 6666666666 6666666666 6666666666 6666666666 6666666666 6666666667
     0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990
     1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
CB25 AGCACGACCA GTTGAATCCC CTTTTATTTT AACAGGACAA ATTCTTACAG TACTCTACTT CTCATACTAT ATTTTAAACC CTATAATTTC TAAATTTTGA

     7777777777 7777777777 7777777777 7777777777 77777
     0000000001 1111111112 2222222223 3333333334 44444
     1234567890 1234567890 1234567890 1234567890 12345
CB25 GATAAATTTC TAAATTAACC TATTTAGTTA ATGAGCTTGA TTTAA
```

```
          111 1111112222 2222222223 3333333333 3333344444 4444444444 4555555555 5666666666 6666666666 777777
   4555888000 2335690022 2334678990 1222344466 7778901133 3445556888 9000234446 8001111123 5556677889 011224
   3569235369 4342962506 8287270061 5025503947 1698105806 9581457457 0259921279 4170367923 0256709261 309021
CB25 CAAAAGAATA CTTGCATTTT CCTTCATAAC TGCAGTATTA GCCTAACTTC CAAATGGCCA TTAGAATCTG TAAATATCTC GACACTCTAT TCCCAT
CB01 .......... .......... .......... .......... .......... .......... .......... .......... .......... ......
CB02 .G........ .C..T..... .T........ .......... .......... T......... .......... .G........ .......... ......
CB03 .G........ .C.....C.. .T......G. ........C. .......... T...A..... C......T.. .......... .......... ...T.G
CB04 .G........ .C.....C.. .TC.....G. ........C. .......... T...A..... C......... .......... .......... ...T.G
CB05 .G........ .C.....C.. .T......G. ........C. .......... T...A..... C......... .......... .......... ...T.G
CB06 .GG....... .C.....C.. .T......G. ........C. .......... T...A..... C......... .......... .......... ...T.G
CB07 .......... .C.....C.. .TC.....G. ........C. .......... T...A..... C......... .......... .......... ...T.G
CB10 .......... .......... .......... .A........ .......... .......... .......... .......... .......... ......
CB11 .......... .......... .......... ....A..... .......... .......... .......... .......... .......... ......
CB12 .......... .......... .......... ..T....... .......... .......... .......... .......... .......... ......
CB14 .......... .......... .......... .....C.... .......... ...G...... .......... .......... .......... ......
CB15 .......... .......... .......... .........G .......... .......... .......... .......... .......... ......
CB17 .G........ .C.....C.. .T......G. .......... .......... T...A..... C......... .......... .......... ...T.G
CB18 .......... .......... .......... .......... .......... .......... ........C. .......... .......... ......
CB19 .......... .......... .......... .......... .........T .......... .......... .......... .......... ......
CB20 .......... .......... .......... .......... .......... ...G...... .......... .......... .......... ......
CB21 .......... .......... .......... .......... .......... ........T. .......... .......... .......... ......
CB22 .......... .......... .......... .......... .......... .......... ......C... .......... .......... ......
CB23 .......... .......... .......... .......... .......... .......... .......... ...C...... .......... ......
CB24 .......... .......... .......... .......... .......... .......... .......... .......... .......... G.....
CB26 .G........ .C..T..... .T........ .......... .......... T......... .......... .......... .......... ......
CB28 .......... .......... .......... .......... ........C. .......... .......... .......... .......... ......
CB31 .......... .......... .......... .......... .......... ..G....... .......... .......... .......... ......
CB32 .G........ .C.....C.. .T......G. ........C. .......... T...A..... C......... .......... .......... ...T.A
CB35 .......... .......... .......... .......... .......... .......... .......... .......... ......T... ......
CB36 .G........ .C.....C.. .TC.....G. ........C. .......... T...AA.... C......... .......... .......... ...T.G
CB37 .G........ .C.AT..... .T........ .......... .......... T......... .......... .......... .......... ......
CB38 .......... .......... .......... .......... .......... .......... ....G..... .......... .......... ......
CB39 .......... .C..T..... .T........ .......... .......... T......... .......... .......... .......... ..T.G.
CB40 .......... .........C .......... .......... .......... .......... .......... .....G.... .......... ......
CB41 ....G..... .......... .......... .......... .......... .......... .......... .......... .......... ......
CB44 .G........ .C.....C.. .TC.....G. ........C. .......... T...A..... C..A...... .......... .......... ...T.G
```

```
          111 1111112222 2222222223 3333333333 3333344444 4444444444 4555555555 5666666666 6666666666 777777
   4555888000 2335690022 2334678990 1222344466 7778901133 3445556888 9000234446 8001111123 5556677889 011224
   3569235369 4342962506 8287270061 5025503947 1698105806 9581457457 0259921279 4170367923 0256709261 309021
CB45 .......... .......... .......... .......... .......... .......... .......... .......... ...T...... ......
CB47 .......... .......... .......... .......... .......... .......... .......... .G........ .......... ......
CB50 .......... ......C... .......... .......... .......... .......... .......... .......... .......... ......
CB52 .G........ .C.....C.. .T......G. ........C. .......... T...A..... C......... .......... ......T... ...T.G
CB53 .......... .C.....C.. .T......G. ........C. .......... T...A..... C......... .......... .......... ...T.G
CB55 .......... .......... ......C... .......... .......... .......... .......... .......... .......... ......
CB60 .G........ .C.....C.. .T......G. ........C. .......... T...A..... C......... .......... .G........ ...T.G
CB61 .......... .......... T......... .......... .......... .......... .......... .......... .......... ......
CB62 .G........ .C..T..... .T........ .......... .......... T......T.. .......... .......... .......... ......
CB63 .......... .......... .......... .......... ....G..... .......... .......... .......... .......... ......
CB64 .......... ........C. .......... .......... .....G.... .......... .......... .......... .......... ......
CB65 .G........ .C..T..... .T........ .......... .......... TG........ .......... .......... .......... ......
CB66 .......... .......... .......... .....C.... .T........ .......... .......... .......... .......... ......
CB67 ....G..... .......... .......... .......... .......... ...G...... .......... .......... .......... ......
CB68 .......... .......... .......... .......... ...C...... .......... .......... .......... .......... ......
CB69 .......... .......... .......... .....C.... .......... .......... .......... .......... .........C ......
CB73 .......... .......... .......... .......... ........C. .......... ..G....... .......... .......... ......
CB74 .......... .......... .......... .......... .......... .......... .......... .......... ....T..... ......
CB75 .......... .......... .......... .....C.... .......... .......... .......... .......... .......... ......
CB76 .......... T......... .......... C......... .......... .......... .......... .......... .......... ......
CB79 .......... ...A...... .......... .......... .......... .......... .......... .......... .......... ......
CB80 .G........ .C..T..... .T........ .......... .......... T......... ....G..... .......... ..T....C.. ......
CB81 .......... .......... .......... .......... .......... ......A... .......... .......... .......... ......
CB82 .......... .......... .......... .......... .......... .........G .......... .......... .......... ......
CB83 .G........ .C..T..... .T........ .......... .......... T......... .C........ .......... .......... ......
CB84 .......... .......... .......... .......... .......... .......... .......... .......... .......... ...T..
CB85 .G........ .C..T..... .T........ .......... .......... T.......T. ....G..... .......... .......C.. ......
CB86 ........C. .......... .......... .......... .......... .......... .......... .......... .......... ......
CB87 .......... .......... .......... .......... .......... .......... ......C... ...C...... .......... ......
CB88 .G........ .C.....C.. .T......G. ........C. .......... T...A..... C......... .......... ..T....... ...T.G
CB89 .......... .......... .........A .......... .......... .......... .......... .......... .......... ......
CB90 .......... .....G.... .......... .......... .......... .......... .......... .......... .......... ......
CB91 .......... .......... .....G.... .......... .......... .......... .......... .......... .......... ......
```

```
          111 1111112222 2222222223 3333333333 3333344444 4444444444 4555555555 5666666666 6666666666 777777
   4555888000 2335690022 2334678990 1222344466 7778901133 3445556888 9000234446 8001111123 5556677889 011224
   3569235369 4342962506 8287270061 5025503947 1698105806 9581457457 0259921279 4170367923 0256709261 309021
CB92  .......... .......... ..C....G.. .......... .......... .......... .......... .......... .......... ......
CB93  .......... .C......CC .......... .........C .TT....... T......... .......... ........A. .......... ......
CB94  .......... T......... .......... .......... .......... .......... .......... .......... .......... ......
CB95  .G...A.... .C.....C.. .T......G. ........C. .......... T...A..... C......... .......... .......... ...T.G
CB96  .G........ .C.....C.. .T......G. .A......C. .......... T...A..... C......... .......... .......... ...T.G
CB97  .G........ .C.....C.. TTC.....G. ........C. .......... T...A..... C......... .......... .......... ...T.G
CB98  .G........ .CC....C.. .T......G. ........C. .......... T...A..... C......... .......... .......... ...T.G
CB99  .......... .......... .......... .......... .......... .......... .......... C......... .......... ......
CB100 .......... .........C .......... .......... .......... .......... .......... .........T .......... ......
CB101 .......... .......... .......... .......... ......A... .......... .......... .......... .......... ......
CB102 .......... .C......CC .......... .......... .TT....... T......... .......... ........A. .....C.... ......
CB103 .......... ....T..... .......... .......... .......... .......... .......... .......... .......... ......
CB104 .......... .C......CC .......... .......... .TT....... T......... .......... ........A. .......... ......
CB105 .G........ .C.....C.. .T......G. ...C....C. .......... T...A..... C......... .......... .......... ...T.G
CB106 .......... .......... .......... .......... .......... .......... .......... .......A.. .......... ......
CB107 .......... .......... .......... .......... .......... .......... .........A .......... .......... ......
CB108 .......... .......... .......... .......... .......C.. .......... .......... .......... .......... ......
CB109 .......... .......... .......... .......... .......... .......... .......... .......... .......... ......
CB110 .G........ .C...G.... .......... .......... .T........ T......... ....G..... .......... .......... ...T..
CB111 .G........ .C...G.... .......... .......... .......... T......... ....G..... .......... .......... ...T..
CB112 .G.......G .C...G.... .......... .......... .T........ T......... ....G.C... .......... .......... ...T..
CB113 .G.....G.. .C...G.... .......... .......... .T........ T......... ....G..... .......... .......... ...T..
CB114 .G........ .C...G.... .......... .......... .T........ T......... ....G..... .......... ..T....... ...T..
CB115 .......... .C...G.... ..C....... .......... .T........ T......... ....G..... .......... .......... ...T..
CB116 .G........ .C...G.... .T........ .......... .T........ T......... ....G..... .......... ........G. ...T..
CB117 TG.......G .C...G.... .......... .......... .T........ T......... ....G.C... .......... .......... ...T..
CB118 .G........ .C...G..C. ..C....... .......... .T........ T......... ....GG.... .......... ........G. ...T..
CB119 .G........ .C...G.... .......... .......... .T........ T......... ....G..... .......... .T....... ...T..
CB08  .......... .C.....C.. .........A .......... .......... T..G...... ......A... ..C....... ........GC .T....
CB09  .......... ...A...... .......... .......... .......... .......... .........A .......... .......... ......
CB13  .......... .......... .......... ...C.....G .......... .......... .......... .......... .......... ......
CB16  .......... .......... .......... .......... .T........ .......... .......... .......... .......... ......
CB27  ..G....... .......... .......... .......... .......... .......... .......... .......... .......... ......
```

```
          111 1111112222 2222222223 3333333333 3333344444 4444444444 4555555555 5666666666 6666666666 777777
     4555888000 2335690022 2334678990 1222344466 7778901133 3445556888 9000234446 8001111123 5556677889 011224
     3569235369 4342962506 8287270061 5025503947 1698105806 9581457457 0259921279 4170367923 0256709261 309021
CB29 .......... .......... .......... .......... .......... .......... .......... .......... A......... ......
CB30 .......... .....G.... ....T..... .......... .......... .......... .......... .......... .......... ......
CB33 .......... .......... .......... .......... .......... .......... .......... .........T .......... ......
CB34 .G........ .C..T..... .T........ .......... .......... T...C..... .......... .......... .......... ......
CB42 .......... ........C. .......... .......... .......... .......... .......... .......... .......... ......
CB43 .G........ .C.....C.. .TC.....G. .A......C. .......... T...A..... C......... .......... .......... ...T.G
CB46 .......... .......... ...C...... .......... .......... .......... .......... .......... .......... ......
CB48 .......... .......... .......... .......... .......... .......... .......... .......... ...G...... ......
CB49 .T........ .C.....C.. .TC.....G. ......C.C. .......... T...A..... C......... .......... .......... ...T.G
CB51 .......... .......... .......... .......... .......... .......... .......... ....C..... .......... ......
CB54 .......... .......... .......... .......... .......... .......... .......... ......C... .......... ......
CB56 ...C...... .......... .......... .......... .......... .......... .......... .......... .......... ...T..
CB57 ......G... .......... .......... .......... .......... .......... .......... .......... .......... ......
CB58 .G........ .C.....C.. .T......G. ........C. .......... T...A..... C......... .......... .......... ...T..
CB59 .G........ .C.....C.. .TC.....G. ........C. .......... T...A..... C......... .......... .......... ...T..
CB70 .......... .......... .......... .......C.. .......... .......... .......... .......... .......... ......
CB71 .G........ .C.....C.. .T......G. ........C. .....G.... T...A..... C......... .......... .......... ...T.A
CB72 .G........ .......... .......... .......... .......... .......... .......... .......... A......... ......
CB77 .......... .......... .......... .......... A......... .......... .......... .......... .......... ......
CB78 .......... .......... .......... .......... .......... .......... .......... .......... .......... ....T.
```

Appendix 2.8 Figure Alignment of 31 variable base positions from 51 alleles amplified for all populations of *P. ariasi* at locus EF-1α. The most widespread allele EF03 is used as the reference sequence and given in full. For population genetic analyses a 777 bp fragment was analysed as missing data was removed; the 5' 22 bp and 3' 18 bp. *P. ariasi* GenBank AF160803 begins on base position 49 of EF03.

```
                1 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990
     1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
EF03 ACCATTGATA TTGCTCTGTG GAAATTCGAG ACGGCCAAGT ACTACGTCAC CATCATCGAT GCTCCCGGAC ATCGTGATTT CATCAAGAAC ATGATCACAG

     1111111111 1111111111 1111111111 1111111111 1111111111 1111111111 1111111111 1111111111 1111111111 1111111112
     0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990
     1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
EF03 GGACATCTCA GGCTGACTGT GCTGTGCTGA TTGTGGCTGC TGGCACTGGT GAATTCGAGG CTGGCATCTC CAAGAATGGT CAGACCCGTG AACATGCCCT

     2222222222 2222222222 2222222222 2222222222 2222222222 2222222222 2222222222 2222222222 2222222222 2222222223
     0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990
     1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
EF03 GCTCGCCTTC ACGCTGGGCG TGAAGCAGCT GATTGTGGGT GTTAACAAGA TGGACTCCAC TGAGCCACCA TTCAGTGAGC CGAGGTTTGA GGAGATCAAG

     3333333333 3333333333 3333333333 3333333333 3333333333 3333333333 3333333333 3333333333 3333333333 3333333334
     0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990
     1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
EF03 AAGGAGGTGT CATCGTACAT CAAGAAGATC GGTTACAATC CAGCTGCTGT GGCTTTCGTG CCCATCTCCG GATGGCATGG AGACAACATG CTGGAGGCTT

     4444444444 4444444444 4444444444 4444444444 4444444444 4444444444 4444444444 4444444444 4444444444 4444444445
     0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990
     1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
EF03 CCAGTAATAT GGGATGGTTC AAGGGGTGGG CCATTGAGCG CAAGGAGGGT AAGGCTGATG GTAAGACTCT GATTGAGGCT CTGGACGCCA TTCTGCCACC

     5555555555 5555555555 5555555555 5555555555 5555555555 5555555555 5555555555 5555555555 5555555555 5555555556
     0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990
     1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
EF03 GGCTCGTCCC ACCGATAAGC CCCTCCGTCT GCCACTGCAG GATGTCTACA AGATTGGTGG AATTGGAACT GTGCCAGTGG GTCGTGTGGA GACTGGTGTG

     6666666666 6666666666 6666666666 6666666666 6666666666 6666666666 6666666666 6666666666 6666666666 6666666667 7777777777
     0000000001 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990 0000000001
     1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
EF03 CTGAAGCCAG GAACTGTCGT GACTTTCGCC CCGGCCAATC TCACCACTGA GGTGAAATCC GTGGAGATGC ACCACGAGGC TCTGCAGGAG GCCGTTCCCG GTGACAATGT

     7777777777 7777777777 7777777777 7777777777 7777777777 7777777777 7777777777 7777777777 7777777778 8888888888 8888888
     1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889 9999999990 0000000001 1111111
     1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567
EF03 GGGCTTCAAC GTGAAGAACG TGTCCGTGAA GGAGTTGCGT CGTGGCTATG TGGCTGGTGA CTCGAAGAAT AATCCACCCA AGGCGGCGTC TGACTTTACC GCTCAGG
```

```
          1111222 2233334444 5555666667 7
       5790379037 8913395788 0133012493 9
       4562818479 8180660706 1347024505 8
EF03   CTCGTCCCGG TGCCCGTGTC GCAGGATCGC G
EF01   .......... .......... .......... C
EF02   .......... .......... ........A. .
EF04   .......... .A........ .......... C
EF05   ....C..... .......... .......... .
EF06   .......... C......... ........A. C
EF07   ....C..... .......... .......... C
EF08   ..A....... .......... .......... C
EF09   .......... .......... .....CA... C
EF10   .......... .......... ........A. C
EF11   ........T. .......A.. .......... C
EF12   .......... C......... .......... .
EF13   .......... .......... .....CA.A. C
EF14   .......... .......A.. .......... .
EF15   ..A....... .......... .......... .
EF16   ....C..... .......... ........A. .
EF17   .......... ....T..... .......... C
EF18   .C........ .......... .......... C
EF19   .......... .......... .....CA... .
EF20   .......... C......... .....C.... .
EF21   ........T. .......A.. .......... .
EF22   .......... .......... .....C.... .
EF23   .......... .....A.A.. .......... .
EF24   .......A.. .......... .......... .
EF25   .......... C......... .....CA... .
EF26   .......... C......A.. .......... .
EF27   ...A...... .....A.... .......... C
EF28   .......... ........C. T......... .
EF29   .......... C......... ....A..... .
EF30   .......... C......... T......... .
EF31   .......... C........T ....A..... .
EF32   .....T.... C......... ....A..... .
EF33   .......... C..T...... ....A..... .
EF34   .......... .........T .......... .
EF35   .......... C.....C... ....A..... .
EF36   ......T... C......... .......... .
EF37   .......... ...T...... .......... .
EF38   .......... ..T....... .......... C
EF39   .......... .......... ...A...... .
EF40   .......... .......... ........AT C
EF41   .........A .......... ........A. .
EF42   ....C..... .......... ........A. C
EF43   .......... C......... .T......A. C
EF44   .......... C..T...... .......... .
EF45   T......... .......... T......... .
EF46   ....C..... ......A... .......... .
EF47   .......... .......... ..C....... C
EF48   .......... C......... .......... C
EF49   .......... .......... .......AA. .
EF50   .......... .......... .....CA.A. .
EF51   ....C..... .......... .....T..A. .
```

Appendix 2.9 Table Genotype frequencies of locus EF-1α in 18 populations of *P. ariasi*. Table annotations given in legend Appendix 2.11.

Region	Morocco	Portugal	NW Spain	C Pyrenees, France		Eastern Pyrenees, France						NE Spain	Massif Central (MC) & Rhone, France				Lot, France	
													MC(S)	MC(S)	MC	Rhone		
EF-α genotype code	AGH	CHR	CSP	HP1	HP2	PAS	IRL07	TUL	ARQ06	ARQ08	CAT	TRJ	CTU	SPV	SAM13	DRAz4	LNP	RME
0303, Hm		0.043	0.238		0.111	0 021	0.182	0.174	0 286	0 043		0.130	0 083	0 087	0 542	0.227	0.750	0.769
0103; Ht			0.143	0.360	0.278	0.250	0.227	0.217	0.214	0.174	0 063		0.125	0.391	0.208	0.364	0.208	0.231
0101, Hm			0.095	0 400	0 556	0 521	0.318	0.348	0.286	0.304	0.313	0.478	0 375	0.217	0 042	0 318		
0102, P			0.048		0.056	0 042	0.045			0.130	0.250	0.043	0.208	0.087		0.045		
0306; P			0.095										0 042	0 043				
0206, P			0 048			0.021						0.043	0 042					
0112, P		0 043				0.021	0.045											
0114, P		0 043										0.043						
1414, Hm		0.130	0.095				0.045	0 043		0 043	0 063							
0305; Ht			0 048			0.021												
0113; P			0.048				0.045						0.125	0 043	0.167		0.042	
0205, P			0.048				0.091		0 071		0.125				0.042			
0203, Ht										0.130				0.043				
0202; Hm								0 087	0 071	0.130								
0105, P				0.040		0 021						0 043						
0110; Ht				0 040		0 021												
0108, Ht				0.040								0.043						
0111, P				0 040								0.043						
0608, P						0.042		0.087			0.063							
0106, P																		
2929; Hm	0.176																	
0332, P	0 059																	
1229; Ht	0 059																	
1231, P	0.059																	
2829, P	0 059																	
2837, I	0 059																	
2844, I	0.059																	
2930; P	0 059																	
2933, Ht	0.059																	
2936, P	0.059																	
2937, P	0 059																	
2938, Ht	0 059																	
3044, I	0.059																	
3045, I	0.059																	
3435, P	0.059																	
1114, P		0 087																
1214, P		0.087																
1219; P		0.087																
0319, Ht		0 043																
0320, P		0 043																
0323; P		0 043																
1221, P		0 043																
1424, P		0.043																
1427; P		0 043																
1920; P		0 043																
2020; Hm		0 043																
2026, P		0 043																
2122; P		0.043																
2225, P		0.043																
0314, Ht			0.048															
0342, P			0 048															
0109, P				0.040														
0143; I				0 040		0.021												
0215, P							0 043											
0512, I								0.071										
0139; P									0 043									
0138; Ht										0 063								
0104, Ht										0.063								
0140; I												0 043						
0618, P												0.043						
1010; Hm												0 043						
1017; Ht														0 043				
0116; Ht														0 043				
0141, Ht																0 045		
0307; Ht																0.045		
N	17	23	21	25	18	48	22	23	14	23	16	23	24	23	24	22	24	13

Appendix 2.10 Figure Alignment of 13 variable base positions from 14 alleles (without size variation; GenBank accessions included in HQ026000-HQ026017) amplified for all populations of *P. ariasi* at locus AAm20. The most widespread allele 20m02 is used as the reference sequence and given in full. For population genetic analyses a 90 bp fragment was analysed as missing data was removed; the 3' 12 bp. Sequence begins on base position 170 of *P. perniciosus* GenBank AJ303377.

```
                  1 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889
         1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
20m02    CCTGAGCTAC CCACGCGCTT CACTCAGCTC AACTATCGCC CTCCATCCGT TCCCAGGCCA CCCAGCAGCC TGAGTCGTCG CTCTCTGGGC

                  1 11
         9999999990 00
         1234567890 12
20m02    GGATACGCAA GC


           13444467 889
         4975247802 250
20m02    GAGATCCCAG TCC
20m01    .G........ ...
20m03    .G.....T.. ...
20m04    .G......GA .TT
20m06    .GA....... ...
20m07    ........GA ..T
20m08    .......... ..T
20m10    .......... .T.
20m11    ......G... ...
20m12    ....C..... ...
20m14    ...G...... ...
20m15    .......... A..
20m16    A......... ...
20m17    .....AG... ...
```

Appendix 2.11 Table Genotype frequencies of locus AAm20 in 18 populations of *P. ariasi*.

Region	Morocco	Portugal	NW Spain	C Pyrenees, France		Eastern Pyrenees, France						NE Spain	Massif Central (MC) and Rhone, France				Lot, France	
													MC(S)	MC(S)	MC	Rhone		
Genotype / Population code	AGH	CHR	CSP	HP1	HP2	PAS	IRL07	TUL	ARQ06	ARQ08	CAT	TRJ	CTU	SPV	SAM13	DRAz4	LNP	RME
0102; Ht			0.333	0.407	0.500	0.375	0.455	0.417	0.542	0.478	0.563	0.435	0.500	0.417	0.583	0.591	0.478	0.385
0202; Hm	1		0.375	0.037		0.167		0.083	0.042	0.087	0.063	0.043	0.167	0.250	0.208	0.318	0.043	0.308
0203; I			0.042	0.037	0.111			0.042				0.043						
0103; Ht			0.042	0.037	0.167					0.043		0.043						
0101; Hm				0.481	0.222	0.458	0.455	0.458	0.417	0.391	0.375	0.391	0.333	0.292	0.208	0.091	0.478	0.308
0107; P		0.083																
0204; P		0.042																
0207; P		0.167																
0208; Ht		0.042																
0707; Hm		0.542																
0708; P		0.083																
0710; P		0.042																
0211; Ht			0.125															
0212; Ht			0.042															
0217; I			0.042															
0114; I							0.045											
0115; I							0.045											
0106; Ht												0.043						
0116; I														0.042				
N	17	24	24	27	18	24	22	24	24	23	16	23	24	24	24	22	23	13

Legend Method of allele/genotype scoring for nucleotide sequence: Hm = directly read from conserved primer sequence of a homozygous fly; Ht = deduced from conserved primer sequence of a fly with a single heterozygous base; P = directly read from PASA primer sequence; I = inferred based on a define algorithm (section 2.2.4).

Appendix 2.12 Figure Alignment of 10 variable base positions from 13 alleles amplified for all populations of *P. ariasi* at locus AAm24 (GenBank accessions HQ025989-HQ02599). The most widespread allele 24m01 is used as the reference sequence and given in full. For population genetic analyses a 121 bp fragment was analysed as missing data was removed; the 5' 9 bp. Sequence begins on base position 81 of *P. perniciosus* GenBank AJ303378.

```
                  1 1111111112 2222222223 3333333334 4444444445 5555555556 6666666667 7777777778 8888888889
         1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
24m01    GGTGTCAGTG ACGAAGCAAC CAACGCCACA AGTGCAACAA GTTCAGCCGT CGCAGCAGCA ACAGCATTAT GTTCATGTGC AGAGCAAGTC

                  1 1111111111 1111111111 1111111111
         9999999990 0000000001 1111111112 2222222223
         1234567890 1234567890 1234567890 1234567890
24m01    TAAAGTATCG TCGCCGGCGC CATCGCATAT ATATGGGAAG


               111
         1235689223
         0028127170
24m01    GCGGAGAAGG
24m02    ...A...TA.
24m03    ...A....A.
24m04    ...A......
24m05    .....A....
24m06    .........A
24m07    A.........
24m08    ....G....A
24m09    .G........
24m10    A........A
24m11    ..A......A
24m12    .G.......A
24m13    ......G..A
```

Appendix 2.13 Table Genotype frequencies per of locus AAm24 alleles in 18 populations of *P. ariasi*.

Region	Morocco	Portugal	NW Spain	C Pyrenees, France		Eastern Pyrenees, France						NE Spain	Massif Central (MC) and Rhone valley, France				Lot, France	
													MC(S)	MC(S)	MC	Rhone		
Genotype / Population code	AGH	CHR	CSP	HP1	HP2	PAS	IRL07	TUL	ARQ06	ARQ08	CAT	TRJ	CTU	SPV	SAM13	DRAz4	LNP	RME
0101; Hm	0.059	0.625	0.417	0.037				0.042			0.063				0.042	0.130		
0606; Hm		0.042	0.208	0.148	0.556	0.542	0.273	0.458	0.625	0.391	0.375	0.348	0.583	0.542	0.667	0.435	1	1
0106; Ht		0.208	0.250	0.148	0.111	0.167	0.227	0.292	0.125	0.304	0.125	0.261	0.167	0.208	0.208	0.435		
0607; Ht		0.042	0.042	0.296	0.167	0.083	0.182	0.042	0.208	0.087	0.313	0.217	0.167	0.167				
0107; Ht		0.083	0.042	0.111	0.056		0.045	0.042		0.043	0.063	0.087	0.042	0.042	0.083			
0109; Ht			0.042			0.042				0.043								
0707; Hm				0.185	0.056	0.042	0.136	0.042		0.043			0.042	0.042				
0609; P				0.074		0.083					0.063							
0610; Ht					0.056							0.043						
0709; I							0.091	0.042				0.043						
0608; Ht								0.042	0.04167	0.087								
0102; P	0.235																	
0103; P	0.176																	
0203; Ht	0.176																	
0202; Hm	0.118																	
0104; Ht	0.059																	
0205; P	0.059																	
0303; Hm	0.059																	
0304; Ht	0.059																	
0611; Ht						0.042												
0108; P							0.045											
N	17	24	24	27	18	24	22	24	24	23	16	23	24	24	24	23	24	13

Legend Method of allele/genotype scoring for nucleotide sequence: Hm = directly read from conserved primer sequence of a homozygous fly; Ht = deduced from conserved primer sequence of a fly with a single heterozygous base; P = directly read from PASA primer sequence; I = inferred based on a define algorithm (section 2.2.4).

Appendix 2.14 Table Models used to estimate pairwise values of d_N and d_S for protein coding loci cyt b and EF-1α, where $d_S < 0.5$ indicates non-saturation of synonymous substitutions and an appropriate outgroup of the MK population test for selection. d_S estimated under the approximate Nei and Gojobori method[¥] (1986) (with Jukes-Cantor correction), and in PAML CODEML runmode -2 according to the maximum likelihood method of Goldman and Yang[§] (1994).

Locus	Outgroup species	Species 2	d_N[¥]	d_S[¥]	d_N[§]	d_S[§]
Cyt b	papa_MD_AF161214	CB05_aria	0.0652	0.8193	0.0385	2.5211
	papa_MD_AF161214	CB25_aria	0.0642	0.8319	0.0381	2.7954
	cauc_ME_FJ217389	CB05_aria	0.0685	1.2092	0.0441	4.0645
	cauc_ME_FJ217389	CB25_aria	0.0675	1.1998	0.0435	4.7398
	masc_MD_CB01	CB05_aria	0.0556	0.6776	0.0398	3.4383
	masc_MD_CB01	CB25_aria	0.0537	0.68	0.0371	2.9970
	hale_MD_CB03	CB05_aria	0.0443	0.9939	0.0240	6.9903
	hale_MD_CB03	CB25_aria	0.0452	1.0357	0.0239	8.5217
	nraria_MD_AF161196	CB05_aria	0.0054	0.1712	0.0052	0.2702
	nraria_MD_AF161196	CB25_aria	0.0036	0.1557	0.0034	0.2686
	negl_ER_CB01	CB05_aria	0.0435	0.7097	0.0281	2.7026
	negl_ER_CB01	CB25_aria	0.0416	0.7122	0.0261	3.0793
	pern_MD_AF161205	CB05_aria	0.0203	0.6183	0.0133	2.5724
	pern_MD_AF161205	CB25_aria	0.0184	0.5776	0.0111	2.2427
	CB05_aria	CB25_aria	0.0018	0.053	0.0018	0.0683
EF-1α	masc_MD_EF01	EF03_aria	0.0327	0.6189	0.0268	0.7726
	masc_MD_EF01	EF13_aria	0.0327	0.6189	0.0268	0.7726
	hale_ME_EF416842	EF03_aria	0.0282	0.5063	0.0238	0.5891
	hale_ME_EF416842	EF13_aria	0.0282	0.5063	0.0238	0.5891
	nraria_MD_AF160804	EF03_aria	0.0029	0.0767	0.0031	0.0721
	nraria_MD_AF160804	EF13_aria	0.0029	0.0767	0.0031	0.0721
	negl_MD_AF160801	EF03_aria	0.0088	0.4191	0.0030	0.4723
	negl_MD_AF160801	EF13_aria	0.0088	0.4191	0.0030	0.4723
	pern_MD_AF160807	EF03_aria	0.0088	0.3725	0.0030	0.4092
	pern_MD_AF160807	EF13_aria	0.0088	0.3725	0.0030	0.4092
	EF03_aria	EF13_aria	0	0	0.0000	0.0000

Legend Species codes denote the first four letters of the formal species name (see Appendix 2.3). Origins of species indicated by: ER = Europe; MD = Mediterranean; ME = Middle East

Appendix 2.15 Table Population pairwise F_{ST} estimates (using haplotype/allele frequencies) (below the diagonal) and significance level (above the diagonal) at all loci of those *P. ariasi* associated with cyt b haplogroup A: (a) cyt b, (b) EF-1α, (c) AAm20, (d) AAm24. Levels of significance, for nuclear loci after Bonferroni correction in FSTAT (v2.9.3.2), and cyt b in ARLEQUIN (v3.11): * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; NS not significant.

(a)

	CHR	CSP	HP1	HP2	PAS	IRL07	TUL	ARQ06	ARQ08	CAT	TRJ	CTU	SPV	SAM13	DRAz4	ROQ	LNP	RME
CHR	-	**	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	NS	NS	NS	NS	NS
CSP	0.3256	-	**	*	**	*	**	*	**	NS	NS	*	**	**	*	NS	**	*
HP1	-0.0005	0.3281	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
HP2	0.0720	0.5014	0.0492	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
PAS	0.0094	0.3588	-0.0048	0.0175	-	NS	NS	NS	NS	NS	NS	NS	*	*	NS	NS	NS	NS
IRL07	0.0650	0.4918	0.0428	-0.0332	0.0127	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
TUL	0.0157	0.4112	-0.0142	-0.0058	-0.0146	-0.0102	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
ARQ06	-0.0039	0.3343	-0.0245	0.0357	-0.0156	0.0298	-0.0165	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
ARQ08	0.0304	0.4817	0.0361	-0.0340	0.0076	-0.0356	-0.0148	0.0235	-	NS	NS	NS	NS	NS	NS	NS	NS	NS
CAT	0.0136	0.2429	0.0101	0.1541	0.0440	0.1436	0.0647	0.0118	0.1324	-	NS	NS	NS	NS	NS	NS	NS	NS
TRJ	0.0219	0.2067	0.0187	0.1422	0.0471	0.1334	0.0678	0.0204	0.1241	-0.0316	-	NS	NS	NS	NS	NS	NS	NS
CTU	0.0509	0.2346	0.0493	0.1744	0.0851	0.1659	0.1027	0.0522	0.1569	-0.0011	0.0069	-	NS	*	NS	NS	*	NS
SPV	0.1090	0.2615	0.1090	0.2460	0.1508	0.2369	0.1702	0.1131	0.2273	0.0455	0.0519	-0.0184	-	**	NS	NS	*	NS
SAM13	0.0936	0.2787	0.0919	0.2192	0.1279	0.2106	0.1461	0.0946	0.2015	0.0434	0.0498	0.0636	0.1047	-	NS	NS	*	NS
DRAz4	0.0132	0.3907	-0.0018	0.0167	-0.0091	0.0116	-0.0157	-0.0093	0.0062	0.0375	0.0387	0.0777	0.1364	0.1336	-	NS	NS	NS
ROQ	0.0084	0.1843	0.0027	0.1218	0.0181	0.1126	0.0437	0.0025	0.1028	-0.0175	-0.0187	-0.0016	0.0391	0.0580	0.0276	-	NS	NS
LNP	0.1180	0.5642	0.0908	-0.0221	0.0485	-0.0213	0.0242	0.0742	-0.0201	0.2250	0.2003	0.2303	0.3063	0.2765	0.0510	0.1836	-	NS
RME	0.1153	0.5442	0.0900	-0.0136	0 0509	-0 0100	0.0333	0.0747	-0.0055	0 2083	0.1842	0.2118	0 2833	0.2574	0 0567	0.1726	-0.0284	-

(b)

	CHR	CSP	HP1	HP2	PAS	IRL07	TUL	ARQ08	CAT	TRJ	CTU	SPV	SAM13	DRAz4	LNP	RME
CHR	-	***	***	***	***	***	***	***	***	***	***	***	***	***	***	**
CSP	0.0831	-	*	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	NS
HP1	0.2219	0.1441	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	***	NS	***	***
HP2	0.2653	0.1613	-0.0114	-	NS	NS	NS	NS	NS	NS	NS	NS	**	NS	**	**
PAS	0.3559	0.3071	0.0459	0.0336	-	NS	NS	NS	NS	NS	NS	NS	***	NS	***	**
IRL07	0.1524	0.0026	0.0759	0.0688	0.2409	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
TUL	0.1978	0.1078	-0.0111	-0.0110	0.0657	0.0470	-	NS	NS	NS	NS	NS	***	NS	***	**
ARQ08	0.2157	0.1605	0.0152	0.0243	0.0610	0.1037	-0.0025	-	NS	NS	NS	NS	***	NS	***	***
CAT	0.1610	0.0844	0.0200	0.0272	0.1200	0.0372	0.0062	-0.0063	-	NS	NS	NS	**	NS	***	***
TRJ	0.1867	0.1245	-0.0171	-0.0024	0.0468	0.0659	-0.0173	-0.0085	-0.0044	-	NS	NS	***	NS	***	**
CTU	0.1733	0.0671	0.0291	0.0284	0.1329	0.0133	0.0176	0.0269	-0.0247	0.0169	-	NS	**	NS	***	**
SPV	0.1713	0.0475	0.0190	0.0139	0.1382	-0.0098	0.0018	0.0460	0.0066	0.0165	-0.0037	-	*	NS	***	*
SAM13	0.2520	0.0712	0.3627	0.3798	0.5567	0.1190	0.3315	0.4065	0.3066	0.3487	0.2504	0.2154	-	NS	NS	NS
DRAz4	0.1898	0.0343	0.0619	0.0473	0.2144	-0.0270	0.0400	0.1152	0.0586	0.0648	0.0321	-0.0082	0.1584	-	*	NS
LNP	0.3692	0.1910	0.5097	0.5472	0.7247	0.2734	0.4853	0.5726	0.4863	0.4965	0.4063	0.3597	0.0276	0.2981	-	NS
RME	0.3286	0.1634	0.4796	0.5234	0.7300	0.2415	0.4510	0.5414	0.4503	0.4592	0.3749	0.3283	0.0229	0.2661	-0.0260	-

(c)

	CHR	CSP	HP1	HP2	PAS	IRL07	TUL	ARQ06	ARQ08	CAT	TRJ	CTU	SPV	SAM13	DRAz4	LNP	RME
CHR	-	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
CSP	0.457	-	**	*	**	***	NS	**	**	NS	***	**	NS	NS	NS	***	NS
HP1	0.4922	0.2624	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
HP2	0.4393	0.2026	-0.0014	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	NS	NS
PAS	0.5573	0.3754	-0.0059	0.0487	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
IRL07	0.5268	0.3297	-0.0126	0.0315	-0.0296	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
TUL	0.476	0.1986	-0.0130	0.0005	0.0302	0.0121	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
ARQ06	0.5208	0.2944	-0.0236	0.018	-0.0283	-0.0249	-0.0095	-	NS	NS	NS	NS	NS	NS	NS	NS	NS
ARQ08	0.5401	0.3361	-0.0134	0.0354	-0.0333	-0.023	0.0115	-0.028	-	NS	NS	NS	NS	NS	NS	NS	NS
CAT	0.4851	0.1929	-0.0113	0.0089	0.0345	0.0158	-0.0248	-0.0104	0.0148	-	NS	NS	NS	NS	NS	NS	NS
TRJ	0.5092	0.3211	-0.0135	0.0139	-0.022	-0.0177	0.0128	-0.0181	-0.0172	0.0192	-	NS	NS	NS	**	NS	NS
CTU	0.484	0.1918	-0.0037	0.014	0.042	0.0213	-0.0202	-0.0031	0.0196	-0.0264	0.0269	-	NS	NS	NS	NS	NS
SPV	0.454	0.0994	0.0414	0.0343	0.1164	0.0842	0.0021	0.0513	0.0843	-0.0078	0.0876	-0.0027	-	NS	NS	NS	NS
SAM13	0.4662	0.0984	0.0538	0.0469	0.1354	0.1005	0.0116	0.0661	0.1006	0.0024	0.1033	0.0048	-0.0203	-	NS	NS	NA
DRAz4	0.4811	0.0217	0.167	0.1339	0.2819	0.2343	0.1038	0.1936	0.2383	0.0949	0.2307	0.0914	0.0183	0.0165	-	NS	NS
LNP	0.5294	0.309	-0.0148	0.028	-0.0171	-0.0176	-0.0002	-0.0264	-0.0205	-0.0004	-0.0104	0.0069	0.0649	0.0777	0.2053	-	NS
RME	0.463	0.087	0.0416	0.0331	0.1243	0 0906	-0 0014	0.0522	0.0913	-0.0128	0.0929	-0 0082	-0.0353	-0.0306	0.0051	0.0688	-

(d)

	CHR	CSP	HP1	HP2	PAS	IRL07	TUL	ARQ06	ARQ08	CAT	TRJ	CTU	SPV	SAM13	DRAz4	LNP	RME
CHR	-	NS	***	***	**	**	*	***	**	*	**	***	***	**	**	***	***
CSP	-0.0029	-	*	**	NS	NS	NS	**	NS	NS	*	**	**	*	NS	***	***
HP1	0.2985	0.2182	-	NS	NS	NS	NS	*	NS	NS	NS	NS	NS	*	***	***	***
HP2	0.4878	0.363	0.1272	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	NS	NS
PAS	0.498	0.3621	0.1744	-0.0292	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
IRL07	0.2934	0.2032	-0.0286	0.1036	0.1461	-	NS	NS	NS	NS	NS	NS	NS	NS	*	***	**
TUL	0.2952	0.1781	0.0611	0.0228	0.026	0.0361	-	NS	NS	NS	NS	NS	NS	NS	NS	***	*
ARQ06	0.5743	0.4439	0.243	0.001	-0.0302	0.2232	0.0864	-	NS	NS	NS	NS	NS	NS	NS	NS	NS
ARQ08	0.4325	0.2939	0.1731	0.0079	-0.0282	0.1448	0.0079	0.0178	-	NS	NS	NS	NS	NS	NS	NS	NS
CAT	0.3761	0.2476	0.0739	-0.0182	-0.0148	0.0467	-0.0244	0.0427	-0.0121	-	NS	NS	NS	NS	NS	**	NS
TRJ	0.3622	0.2444	0.0653	-0.0048	0.0108	0.0428	-0.016	0.0617	0.0107	-0.0273	-	NS	NS	NS	NS	***	NS
CTU	0.4584	0.3361	0.1364	-0.0271	-0.0256	0.1128	0.0125	0.0066	-0.0067	-0.0207	-0.0072	-	NS	NS	NS	*	NS
SPV	0.4316	0.3095	0.1217	-0.0224	-0.0198	0.0977	0.0018	0.0176	-0.0088	-0.0255	-0.0132	-0.0226	-	NS	NS	**	NS
SAM13	0.4522	0.3206	0.2022	0.0151	-0.0205	0.1773	0.0225	0.0157	-0.0305	-0.0005	0.0205	-0.0005	-0.0021	-	NS	NS	NS
DRAz4	0.2954	0.165	0.1765	0.1027	0.0741	0.1507	0.0135	0.1413	0.018	0.0325	0.0456	0.0711	0.056	0.0257	-	***	*
LNP	0.7678	0.6543	0.4722	0.2166	0.1646	0.4889	0.309	0.1058	0.2253	0.2939	0.2752	0.182	0.1982	0.1739	0.3584	-	NA
RME	0.7094	0.5829	0.3967	0.1507	0 0984	0.3972	0.2376	0.0627	0.1578	0.213	0.211	0.1333	0.1476	0.1258	0.2868	NA	-

Appendix 3.1 Cloning of *Phlebotomus* apyrase

In general good microbiological practices were carried out: all vessels were kept closed, opening only for the minimum time required to introduce or remove materials, in order to prevent contamination; to minimize the possibility of producing contaminated aerosols, solutions were mixed by gentle rolling and swirling rather than vigorous shaking (to avoid frothing); and during pipetting tips were placed into the liquid or onto a surface prior to gently ejecting the contents. PCR product (563 bp) using conserved primers APY-1F with APY-3R were amplified no more than 24 hours in advance of cloning. Amplification used *Taq* polymerase, causing 3′ adenylation the PCR product. PCR product was purified using GENECLEAN® II as described in Appendix 2.2.

Ligation reagents were defrosted at room temperature. Vector was mixed by flicking then centrifuged, all other reagents were flicked and shaken down, especially insert to prevent loss of PCR product 'A' tails. Each TOPO® ligation reaction was made individually (no master mix). Ligation reaction reagents were added in order of: 1 µl salt solution (from kit); 1 µl of pCR®4-TOPO® vector (10 ng/µl); 4 µl of apyrase insert (0.715 ng/µl). The ligation reaction was mixed gently by flicking and tapping down (not vortexed or centrifuged), and incubated at room temperature for 5 min. Completed ligation was place on ice, and transformation of the TOPO® Cloning reaction (vector construct) into chemically competent *E. coli* (Mach1™-T1^R) cells was carried out immediately to ensure the highest cloning and transformation efficiencies.

Chemically competent *E. coli* cells were thawed on ice shortly before use (approximately 15 min). Cells were mixed by gentle flicking. 2µl of the TOPO® Cloning reaction was pipetted into to one vial (50 µl) of chemically competent *E. coli* cells and mixed gently by shaking the tube. The reaction was incubated on ice for 30 min. Cells were heat shocked in a water bath set at 42°C for 30 secs without shaking, and then immediately placed on ice for 2 min. After removal from ice, 250 µl of SOC medium (from kit, at room temperature) was added to the tube of transformed cells, and shaken horizontally at 200 r.p.m., 37°C, for one hour.

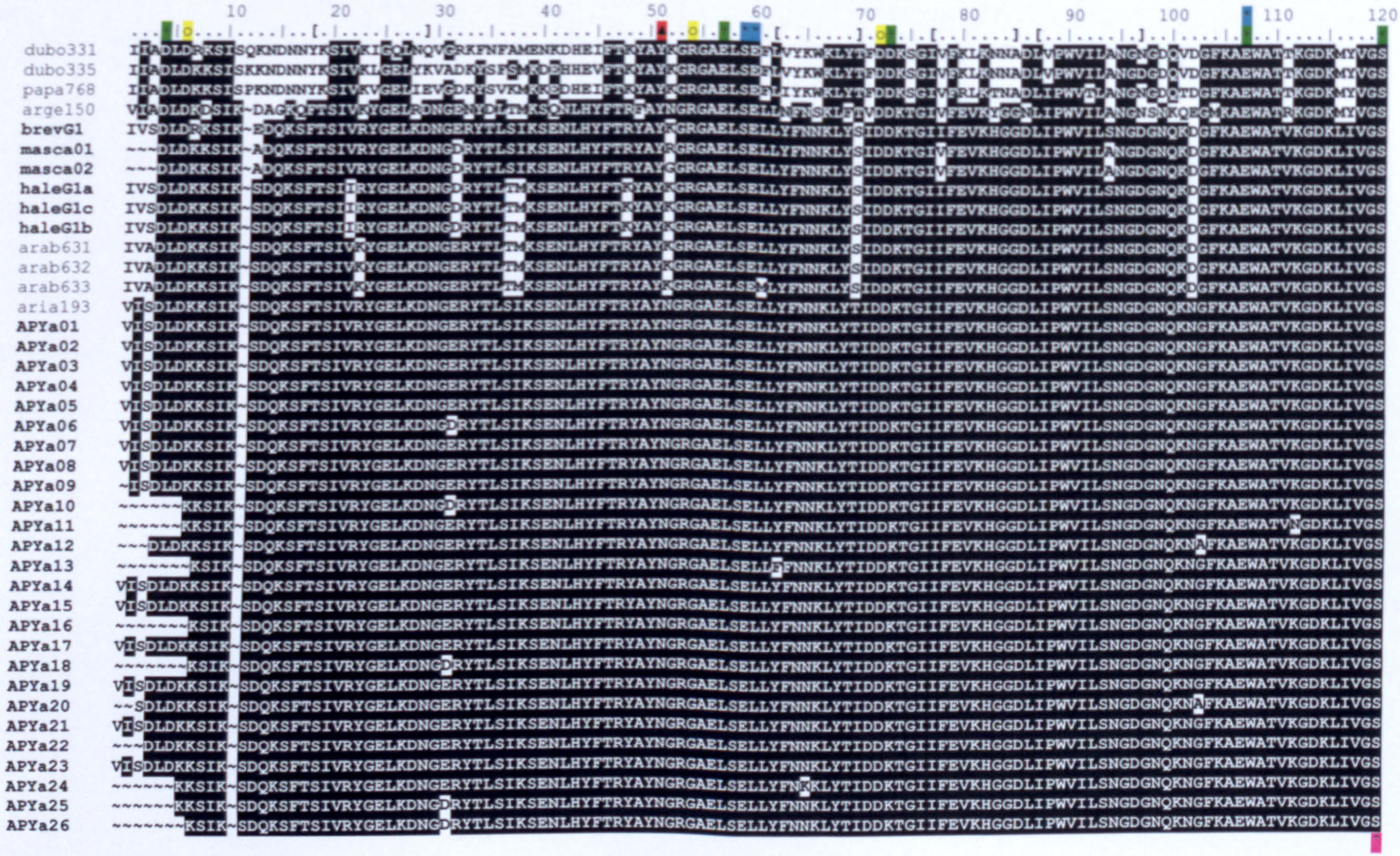
Transformed cells were plated on LB agar selective plates (0.3l LB Agar: 25 g/l of Lauria Broth (LB) medium (Merck); 15 g/l of Agar; 50 µg/µl final concentration of selective agent kanamycin). LB agar plates (25 ml) were pre-warmed at 37°C 30mins before use. Transformed cells were mixed by pipetting, and in a fume-hood 10 µl and 50 µl of each transformation were spread onto separate agar plates. For 10 µl plates, to

ensure surface of plate was uniformly covered, 20 µl of SOC medium was pipetted into the middle of the plate, to which the 10 µl of transformed cells were added then spread. The remainder of the vial of transformed cells was spread onto a third plate. With replaced lids, the plates were left to stand for 10 min at room temperature then incubated inverted overnight (16-24 hrs) at 37°C.

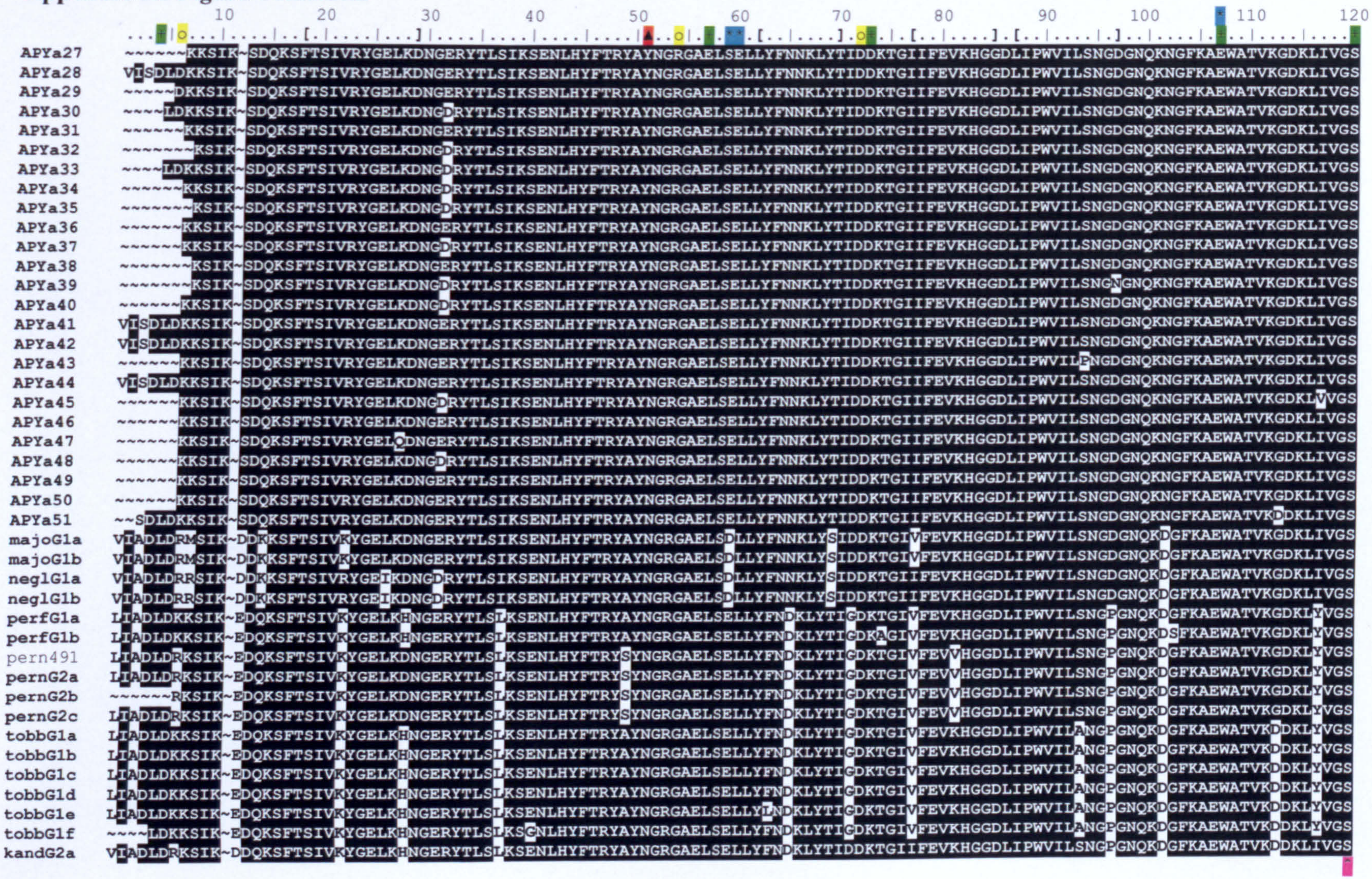
Colonies were picked, and placed directly into a sterlin tube containing 5 ml of LB medium with kanamycin (concentrations as above): one colony per tube. Colonies were grown overnight in a shaking incubator set at 125 r.p.m., 37°C. Following colony growth DNA plasmid was isolated from bacterial colonies by alkaline lysis in miniprep purification. Part of the LB medium with grown colonies was decanted into a 2 ml tube and centrifuged at 14 Kr.p.m. for one min, after which the supernatant was discarded. The pellet was then re-suspended vortexing in 300 µl of Buffer P1 (15 mM Tris pH8, 10 mM EDTA; 10 µg/ml RNAse). 300 µl of lysis Buffer P2 (0.2 M NaOH; 1% SDS) was added and immediately mixed by moderate inversion (5 times) until sample became clear. 300 µl of neutralizing Buffer P3 (3 M KOAc), mixed by four inversions then four vertical rapid/vigorous shakes. The sample was then left on ice for 30 min (minimum), then centrifuged at 14 Kr.p.m. for five min. The supernatant (containing small bacterial DNA plasmids) was transferred to a sterile 1.5 ml Eppendorf tube, to which 700 µl of isopropanol was added, mixed well, and left at room temperature for up to 30 min. A 30 min centrifugation step (14 Kr.p.m.) followed, after which the supernatant was pipetted off. The plasmid pellet was then washed with 500 µl of 70% ethanol, respun for five min to re-pellet. All ethanol was then removed by pipetting and air drying the sample. The plasmid pellet was re-suspended in 100 µl 1x TE to which in a fume-hood 100 µl of phenol chloroform (to denature and dissolve proteins) was added and vortexed until a milky solution was produced. The sample was centrifuged at 14 Kr.p.m. for five min. The upper liquid layer was then pipetted off into a sterile tube. 2.5x volume of ethanol was added and the sample centrifuged at 14 Kr.p.m. for 15 min. Again all ethanol was removed and the dried plasmid pellet re-dissolved in 20 µl of 1x TE.

Appendix 3.2 Figure Alignment of the 174-translated amino acid apyrase fragment from all unique nucleotide alleles amplified by conserved primer pair APY-1F with APY-3R. Alignment starts on nucleotide 110 of GenBank accession AY845193 (*P. ariasi*). Code names of alleles amplified in this thesis for *P. ariasi* (APYa_NN) and other *Phlebotomus* are denoted in bold. *Phlebotomus* GenBank sequences published (up to 01/09/2009) are identified by the first for letters of the formal species name (see Appendix 2.3) followed by the last three digits of their accession number: *P. duboscqi* (DQ834331, DQ834335); *P. papatasi* (AF261768); *P. argentipes* (DQ136150); *P. arabicus* (EZ000631, EZ000632, EZ000633); *Phlebotomus perniciosus* (DQ192490, DQ192491), *Phlebotomus ariasi* (AY845193). Amino acid changes at functional sites are highlighted including: binding (–) and calcium binding (*) in the human homologues (Dai *et al*., 2004); after *in vitro* mutagenesis of the human homologue, O are essential to APDase activity, ▲ single residue mutation from Glu to Tyr with high associated ADPase nucleotidase activity; and ^ (carets under sequence alignment) point mutations that convert the wild-type human CAN into 100-fold more potent ADPase that abolishes platelet aggregation (Yang and Kirely, 2004). [] Brackets enclose putative MHC epitope sites in the sandfly *P. dubosqci* (Kato *et al*., 2006).

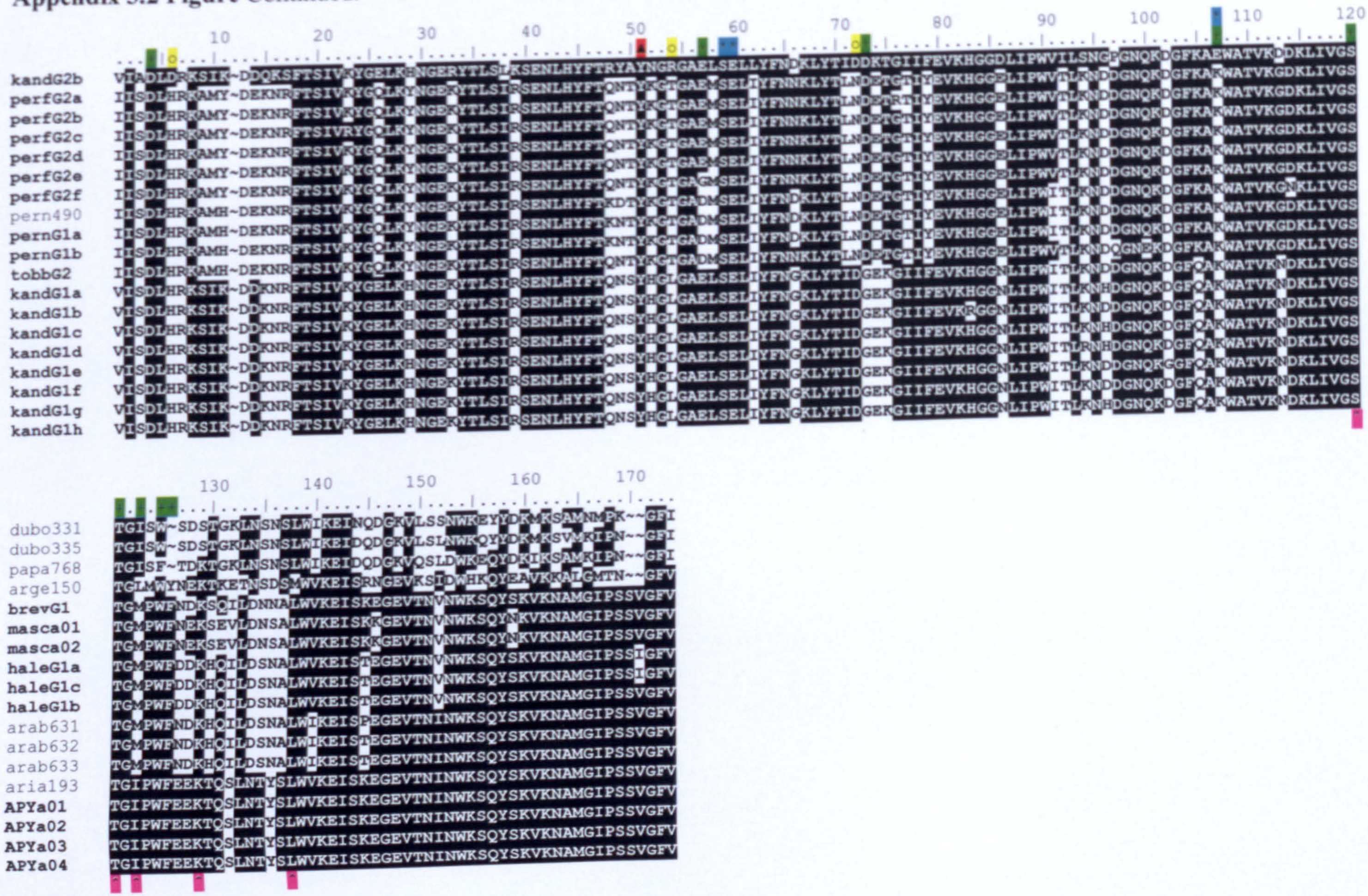
Appendix 3.2 Figure Continued.



Appendix 3.2 Figure Continued.

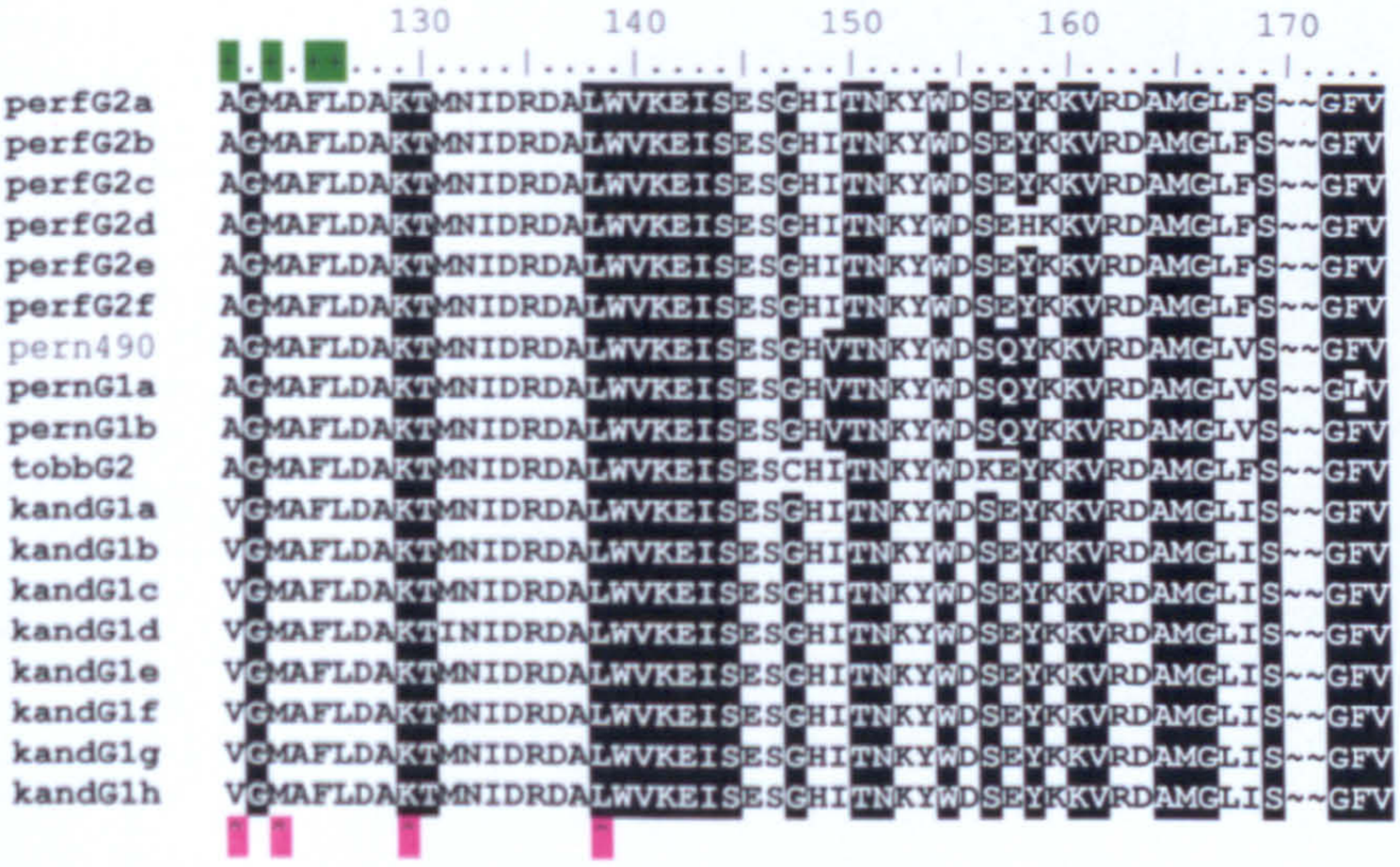


Appendix 3.2 Figure Continued.



Appendix 3.2 Figure Continued.

Appendix 3.2 Figure Continued.



Appendix 3.3 Figure Alignment of the 284 variable base positions from 92 unique apyrase nucleotide alleles amplified in this thesis by conserved primer pair APY-1F with APY-3R. Alignment starts on nucleotide 110 of *P. ariasi* GenBank accession AY845193 (aria193), which is given as the reference sequence. Code names: *P. ariasi* (APYa_NN); and other *Phlebotomus* species are identified by the first for letters of the formal species name (see Appendix 2.3). Missing data = ?; gaps = -.

```
                                                      111 1111111111 1111111111 1111111111 1111111111 1111122222 2222222222
              1111 2222233333 3444555666 6677778888 8999999000 0011111222 2334444445 5555566667 7777888888 9999900011 1111112222
        1346792368 0345701378 9238147034 6825691245 7036789256 8912467236 9251234580 3456901280 1247013468 0567817801 3456890234
aria193 GTACTCCTGC AAGTCTAATC CTCGTCCCCG GGTCGTGAGG CTAGAGATGT ATCACAGAAT TCCGCGATGT CAATACGCCA ATATATGTGA TCAACACACA TGATACAAAC
brevG1  A.GA.G...T G.A.AC.GGA G...C..... ..C....... .C..C..CTC G.......G. C.....G..C ...GCA.GG. .CCCG..CC. .......T.. C..C...G..
masca01 ??????...T ....A..GG. GC..C..... ...G...... T..T..GCTC G.......G. C...A.G..C ..GGGA.GT. .CCCG..CC. .......T.. ...C...G..
masca02 ??????...T ....A..GG. GC..C..... ...G...... T..T..GCTC G.......G. C...A.G..C .GGGGA.GT. .CCCG..CC. .......T.. ...C...G..
haleG1a A.GA.G...T ....A..... T...C....A C.CT...... .CCT...CTC GA..G...G. C...AAG..C ...GCA.AG. ....GC.... .......T.. ...C.T.G..
haleG1c A.GA.G...T ....A..... T...C....A C.CT...... .CCT...CTC GA..G...G. C...AAG..C ...GCA.AG. ....GC.... .......T.. ...C.T.G..
haleG1b A.GA.G...T ....A..... T...C....A C.CT...... .CCT...CTC GA..G...G. C...AAG..C ...GCA.AG. ....GC.... .......T.. ...C.T.G..
APYa01  .......... .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa02  .......... .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa03  .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa04  .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa05  .......... .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa06  .......... ..A....... .......... .......... ...T...... .......... .......... .......... .......... .......... ..........
APYa07  .......... .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa08  .......... .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa09  ??........ .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa10  ?????????? .......... .......... .......... ...T...... .......... .......... .......... .......... .......... ..........
APYa11  ?????????? .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa12  ??????.... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa18  ?????????? ?......... .......... .......... ...T...... .......... .......... .......... .......... .......... ..........
APYa13  ?????????? ?......... .......... .......... .C.....C.. .......... .......... .......... .........T .......... ..........
APYa14  .......... .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa15  .......... .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa16  ?????????? ?......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa17  .......... .......... .......... .......... .......... .......... .......... .......... .......... ......T... ..........
APYa19  .......... .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa20  ????...... .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa21  .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa22  ??????.... .......... .......... .......... .......... .......... .......... .......... .......... ......T... ..........
APYa23  .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
```

```
                                                        111 1111111111 1111111111 1111111111 1111111111 1111122222 2222222222
                1111 2222233333 3444555666 6677778888 8999999000 0011111222 2334444445 5555566667 7777888888 9999900011 1111112222
          1346792368 0345701378 9238147034 6825691245 7036789256 8912467236 9251234580 3456901280 1247013468 0567817801 3456890234
APYa24    ?????????? .......... .......... .......... .......... .......... .......... .......... .......... ....A..... ..........
APYa25    ?????????? .......... .......... .......... ...T...... .......... ...C...... .......... .......... .......... ..........
APYa26    ?????????? ?......... .......... .......... ...T...... .......... ...C...... .......... .......... .......... ..........
APYa27    ?????????? .......... ......T... .......... .......... .......... .......... .......... .......... .......... ..........
APYa28    .......... .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa29    ????????.. .......... .......... .......... .C.....C.. .......... .......... .......... .......... ......T... ..........
APYa30    ???????... .......... .......... .......... ...T...... .......... ...C...... .......... .......... .......... ..........
APYa31    ?????????? .......... .......... .......... .C.....C.. .......... ...C...... .......... .......... .......... ..........
APYa32    ?????????? ?......... .......... .......... ...T...... .......... ...C...... .......... .......... .......... ..........
APYa33    ???????... .......... .......... .......... ...T...... .......... .......... .......... .......... .......... ..........
APYa34    ?????????? .......... .......... .......... ...T...... .......... .......... .......... .......... .......... ..........
APYa35    ?????????? ?......... .......... .......... ...T...... .......... ...C...... .......... .......... .......... ..........
APYa36    ?????????? .......... .......... .......... .C.....C.. .......... ...C...... .......... .......... .......... ..........
APYa37    ?????????? .......... .......... .......... ...T...... .......... ...C...... .......... .......... .......... ..........
APYa38    ?????????? ?......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa39    ?????????? ?......... .......... .......... ...T...... .......... ...C...... .......... .......... .......... ..........
APYa40    ?????????? .......... .......... .......... ...T...... .......... .......... .......... .......... .......... ..........
APYa41    .......... .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa42    .......... .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa43    ?????????? .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa44    .......... .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa45    ?????????? .......... .......... .......... ...T...... .......... .......... .......... .......... .......... ..........
APYa46    ?????????? .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa47    ?????????? .......... .......... .......C.. .C.....C.. .......... .......... .......... .......... .......... ..........
APYa48    ?????????? .......... .......... .......... ...T...... .......... ...C...... .......... .......... .......... ..........
APYa49    ?????????? .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... .....T....
APYa50    ?????????? .......... .......... .......... .C.....C.. .......... .......... .......... .......... .......... ..........
APYa51    ????...... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
majoG1a   .C..GA.C.. GT......GA .CA.C..... .A...C.... .CTA...... G.......G. C.T......C .......... ....TC.... .......T.. ..........
majoG1b   .C..GA.C.. GT......GA .CA.C..... .A...C.... .CTA...... G.......G. C.T......C .......... ....TC.... .......T.. ..........
neglG1a   .C..GATC.. GG.....GGA .CA.C..... .....AC... .C.T....TC G.......G. C.T......C .......... ...CT..... .......T.. C.........
neglG1b   .C..GATC.. GG.....GGA .CA.C..... .....AC... .C.T....TC G.......G. C.T......C .......... ...CT..... .......T.. C.........
perfG1a   CC..GG.C.T ........GA A...C..TT. .A......AC .C.A.....C ...C....G. CT....G..C .......T.. ..GCG..... .TG.T...T. ..GA......
perfG1b   CC..GG.C.T ........GA A...C..TT. .A......AC .C.A.....C ...C....G. CT....G..C .......T.. ..GCG..... .TG.T...T. ..GA....G.
pernG2a   CC..GA.C.T G.A.T...GA A...C..TT. .A......A. .C.......C ..AC....GC CT....G.TC .......... ..GCG..... .TG.....T. ..GA......
pernG2b   ?????????? G.A.T...GA A...C..TT. .A......A. .C.......C ..AC....GC CT....G.TC T......... ..GCG..... .TG.....T. ..GA......
pernG2c   CC..GA.C.T G.A.T...GA A...C..TT. .A......A. .C.......C ..AC....GC CT....G.TC T......... ..GCG..... .TG.....T. ..GA......
```

```
                                                     111 1111111111 1111111111 1111111111 1111111111 1111122222 2222222222
              1111 2222233333 3444555666 6677778888 8999999000 0011111222 2334444445 5555566667 7777888888 9999900011 1111112222
        1346792368 0345701378 9238147034 6825691245 7036789256 8912467236 9251234580 3456901280 1247013468 0567817801 3456890234
tobbG1a CC..GGTC.T ........GA A...C.TT.. .A......AC .........C ...C...... CTT...G..C .......T.. ...CG.A... .TG.T...T. ..G.......
tobbG1b CC..GGTC.T ........GA A...C.TT.. .A......AC .........C ...C...... CTT...G..C .......T.. ...CG.A... .TG.T...T. ..G.......
tobbG1c CC..GGTC.T ........GA A...C.TT.. .A......AC .........C ...C...... CTT...G..C .......T.. ...CG.A... .TG.T...T. ..G.......
tobbG1d CC..GGTC.T ........GA A...C.TT.. .A......AC .........C ...C...... CTT...G..C .......T.. ...CG.A... .TG.T...T. ..G.......
tobbG1e CC..GGTC.T ........GA A...C.TT.. .A......AC .........C ...C...... CTT...G..C .......T.. ...CG.A... CTG.T...T. ..G.......
tobbG1f ???????C.T ........GA A...C.TT.. .A......AC .........C ...C...G.. CTT...G..C .......T.. ...CG.A... .TG.T...T. ..G.......
kandG2a ....GA.C.T G.......GA ....C..TT. .A.T.....C ..G....C.C G..C....G. .T....G..C .......T.. .CGCG..... .TG.....T. ..........
kandG2b ....GA.C.T G.......GA ....C..TT. .A.T....AC ..G....C.C G..C....G. .T....G..C .......T.. .CGCG..... .TG.....T. ..........
perfG2a AC...ATCCT G..G.GTCGA TAACAT.TT. .A.TC...AT .....A...C G....GA.G. .TTC.AGAAC ...A.AC... .AGCG..AC. .T...G..TC .A.C..G...
perfG2b AC...ATCCT G..G.GTCGA TAACAT.TT. .A.TC...AT .....A...C G....GA.G. .TTC.AGAAC ...A.AC... .AGCG..AC. .T...G..TC .A.C..G...
perfG2c AC...ATCCT G..G.GTCGA TAACAT.TT. ...TC...AT .....A...C G....GA.G. .TTC.AGAAC ...A.AC... .AGCG..AC. .T...G..TC .A.C..G...
perfG2d AC...ATCCT G..G.GTCGA TAACAT.TT. .A.TC...AT .....A...C G....GA.G. .TTC.AGAAC ...A.AC... .AGCG..AC. .T...G..TC .A.C..G...
perfG2e AC...ATCCT G..G.GTCGA TAACAT.TT. .A.TC...AT .....A...C G....GA.G. .TTC.AGAAC ...A.AC... .AGCG..AC. .T...G..TC .A.C..G...
perfG2f AC...ATCCT G..G.GTCGA TAACAT.TT. .A.TC...AT .....A...C G....GA.G. .TTC.AGAAC ...A.AC..G .AGCG..AC. .T...G..TC .A.C..G...
pernG1a AC...ATCC. G..G.GCCGA TAACAT.TA. .A.TC....T .....A..TC G....GA.G. CTTAAAGAAC ...G.AC... TAGCG..AC. .TG..G..TC .A.C..G...
pernG1b AC...ATCC. G..G.GCCGA TAACAT.TA. .A.TC....T .....A..TC G....GA.G. CTTAAAGAAC ...G.AC... TAGCG..AC. .TG..G..TC .A.C..G...
tobbG2  AC...A..C. G..G.GCCGA .AACA..TT. .A.TC...AT .C...A...C G....GA.G. CT...AGAAC ...A.AC... TAGCG..AC. .T...G..TC .A.C..G...
kandG1a .C.T.ATCC. G.......GA ..ATGT.TT. .A.T....AC .....A...C G....G.... .T...AGAT. .C.C..T... ..GCG..AC. .TGGT...T. ....GTG..A
kandG1b .C.T.ATCC. G.......GA ..ATGT.TT. .A.T....AC .....A...C G....G.... .T...AGAT. .C.C..T... ..GCG..AC. .TGGT...T. ....GTG..A
kandG1c .C.T.ATCC. G.......GA ..ATGT.TT. .A.T....AC .....A...C G....G.... .T...AGAT. .C.C..T... ..GCG..AC. .TGGT...T. ....GTG..A
kandG1d .C.T.ATCC. G.......GA ..ATGT.TT. .A.T....AC .....A...C G....G.... .T...AGAT. .C.C..T... ..GCG..AC. .TGGT...T. ....GTG..A
kandG1e .C.T.ATCC. G....C..GA ..ATGT.TT. .A.T....AC .....A...C G....G.... .T...AGAT. .C.C..T... ..GCG..AC. .TGGT...T. ....GTG..A
kandG1f .C.T.ATCC. G.......GA ..ATGT.TT. .A.T....AC .....A...C G....G..G. .T...AGAT. .C.C..T... ..GCG..AC. .TGGT...T. ....GTG..A
kandG1g .C.T.ATCC. G.......GA ..ATGT.TT. .A.T....AC .....A...C G....G.... .T.A.AGAT. .C.C..T... ..GCG..AC. .TGGT...T. ....GTG..A
kandG1h .C.T.ATCC. G.......GA ..ATGT.TT. .A.T....AC .....A...C G....G.... .T...AGAT. .C.C..T... ..GCG..AC. .TGGT...T. ....GTG..A

        2222222222 2222222222 2222222222 2222222222 2233333333 3333333333 3333333333 3333333333 3333333333 3333333334 4444444444
        2223333334 4444445555 5666777778 8888888999 9900000001 1111233333 3445555666 6667777777 7888888888 9999999990 0000000011
        5680124670 3456892568 9147135690 1256789014 7834567892 3589703678 9290147123 4690245678 9012346789 0123456790 1245678901
aria193 AGATTACTTG GAAAATTGGT CCTAGGTCGT CGTGGCGATA TCGAATGGCT AACGGATGGG TCATTCAACA GACCCGGTTT GAGGAGAAAC CCAGTCTCTA ACCCTACAGC
brevG1  G...C..... ...G.CCA.. ..C.....C. ..C.....C. ...G.C...C ......C... ....CG.... ..G......C A.T..C..T. ....ATC.GG ..A.A..GC.
masca01 C.G..G.... .....CCA.. .T.....T.G ........CG C..G.....C .GG.AGG... G....GC... ..G......C A.T.....T. .G..GT..GG ..A.AG.GC.
masca02 C.G..G.... .....CCA.. .T.....T.G ........CG C..G.....C .GG.AGG... G....GC... ..G......C A.T.....T. .G..GT..GG ..A.AG.GC.
haleG1a ..C.C..... .....CAT.. .T.T...TC. .........C ...G.C...C ....A.G... .....G.... ..G......C ..T..C.GCA ....AT..GG ..G.A..GCT
haleG1c ..C.C..... A....CAT.. .T.T...TC. .........C ...G.C...C ....A.G... .....G.... ..G......C ..T..C.GCA ....AT..GG ..G.A..GCT
haleG1b ..C.C..... .....CAT.. .T.T...TC. .........C ...G.C...C ....A.G... .....G.... ..G......C ..T..C.GCA ....AT..GG ..G.A..GCT
APYa01  .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
```

```
          2222222222 2222222222 2222222222 2222222222 2233333333 3333333333 3333333333 3333333333 3333333333 3333333334 4444444444
          2223333334 4444445555 5666777778 8888888999 9900000001 1111233333 3445555666 6667777777 7888888888 9999999990 0000000011
          5680124670 3456892568 9147135690 1256789014 7834567892 3589703678 9290147123 4690245678 9012346789 0123456790 1245678901
APYa02    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa03    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa04    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa05    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa06    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ....C.....
APYa07    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa08    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa09    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa10    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ....C.....
APYa11    .......... .......... .......... .......... .......... .......T.. .......... .......... .......... .......... ..........
APYa12    .......... .......... .......... .......... .......C.. .......... .......... .......... .......... .......... ..........
APYa18    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ...A......
APYa13    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa14    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa15    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa16    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa17    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa19    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ...A......
APYa20    .......... .......... .......... .......... .......C.. .......... .......... .......... .......... .......... ..........
APYa21    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa22    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa23    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa24    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa25    ..........A......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa26    .......... .......... .......... .......... .......... .......... .....T.... .......... .......... .......... ..........
APYa27    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa28    .......... .......... .......... .......... .......... .......... .........C .......... .......... .......... ..........
APYa29    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa30    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa31    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa32    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa33    ........C. .......... .......... .......... .......... .......... .....T.... .......... .......... .......... ..........
APYa34    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa35    .......... .......... .......... .......... ..A....... .......... .......... .......... .......... .......... ..........
APYa36    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa37    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa38    .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
```

```
                2222222222 2222222222 2222222222 2222222222 2233333333 3333333333 3333333333 3333333333 3333333333 3333333334 4444444444
                2223333334 4444445555 5666777778 8888888999 9900000001 1111233333 3445555666 6667777777 7888888888 9999999990 0000000011
                5680124670 3456892568 9147135690 1256789014 7834567892 3589703678 9290147123 4690245678 9012346789 0123456790 1245678901
APYa39          .......... .......... .......... ......A... .......... .......... .......... .......... .......... .......... ..........
APYa40          .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa41          .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa42          .......... .......... .......... .......... .......... .......... .......... .......... .......... .....T.... ..........
APYa43          .......... .......... .........C .......... .......... .......... .......... .......... .......... .......... ..........
APYa44          .......... .......... .......... .......... ........G. .......... .......... .......... .......... .......... ..........
APYa45          .......... .......... .......... .......... .......... .......... ..G....... .......... .......... .......... ..........
APYa46          .......... .......... .......... .......... .......... .......... .....T.... .......... .......... .......... ..........
APYa47          .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa48          .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa49          .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
APYa50          .......... .......... .......... .......... .......... .......... .......... .......... .......... .....T.... ..........
APYa51          .......... .......... .......... .......... .......... .........A .......... .......... .......... .......... ..........
majoG1a         .....G.... .....CCA.. ..CG...TC. .......... ...G.C.... .......... .....T.... ..G......C AGC.....GA ....AAA.C. ..A.A..GC.
majoG1b         .....G.... .....CCA.. ..CG...TC. .......... ...G.C.... .......... .......... ..G......C AGC.....GA ....AAA.C. ..A.A..GC.
neglG1a         .......... .....CCA.. ...G...TC. .......... ...G.C...C .G........ ......C... ..G......C ..C.....G. .GT.AAA.C. ....A..G..
neglG1b         .......... .....CCA.. ...G...TC. .......... ...G.C...C .G........ ......C... ..G......C ..C.....G. .GCAAAA.C. ....A..G..
perfG1a         .....G.... .....CCA.C TA...T.TT. ......CCA. ...G.C...C ....C.G... ..TA...... .TGGTTTC.G ..CA..G... .GGAGA.AC. G.AAA.TGC.
perfG1b         .....G.... .....CCA.C TA...T.TT. ......CCA. ...G.CA..C ....C.G... ..TA...... .TGGTTTC.G ..CA..G... .GGAGA.AC. G.AAA.TGC.
pernG2a         .....G.... .GT...CA.C TA...A.TC. ......CCA. ...G.C...C ....A..... ..TA.A.... .TGATTCC.G ..CA..G... .GGAA..AC. G..AA.TGC.
pernG2b         .....G.... .GT...CA.C TA...A.TC. ......CCA. ...G.C...C ....A..... ..TA.A.... .TGATTCC.G ..CA..G... .GGAA..AC. G..AA.TGC.
pernG2c         .....G.... .GT...CA.C TA...A.TC. ......CCA. ...G.C...C ....A..... ..TA.A.... ATGATTCC.G ..CA..G... .GGAA..AC. G..AA.TGC.
tobbG1a         .....G.... .....CCT.C TA...T.TCG ......CCA. ...G.C...C ....A....A ..TA...... ..GATTTC.G ..CA..G... .GGAAA.AC. G.AAA..GCT
tobbG1b         .....G.... .....CCT.C TAC..T.TCG ......CCA. ...G.C...C ....A....A ..TA...... ..GATTTC.G ..CA..G... .GGAAA.AC. G.AAA..GCT
tobbG1c         .....G.... .....CCT.C TA...T.TCG ......CCA. ...G.C...C .G..A....A ..TA...G.. ..GATTTC.G ..CA..G... .GGAAA.AC. G.AAA..GCT
tobbG1d         .....G.... .....CCT.C TA...T.TCG ......CCA. ...G.C...C ....A....A ..TA...... ..GATTTC.G ..CA..G... .GGAAA.AC. G.AAA..GCT
tobbG1e         .....G.... .....CCT.C TA...T.TCG ......CCA. ...G.C...C ....A....A ..TA...... ..GATTTC.G ..CA..G... .GGAAA.AC. G.AAA..GCT
tobbG1f         .....G.... .....CCT.C TA...T.TCG ......CCA. ...G.C...C ....A....A ..TA...... ..GATTTC.G ..CA..G... .GGAAA.AC. G.AAA..GCT
kandG2a         ......T..A .....CCA.. TA.....TC. .C....CCA. ...G.....C .........A .......... ..AG.TTCAG ..C.C..... .TC.A.AAC. ..AAA.TGC.
kandG2b         ......T..A .....CCA.. TA.....TC. .C....CCA. ...G.....C .........A .......... ..AG.TTCAG ..C.C..... .TC.A.AAC. ..AAA.TGC.
perfG2a         ...C..AA.. ...G.CCA.G ......CTCA A...A..... ...G.C...C ...A...... .T.....G.. ..GG.TTC.G ..C.CA.... .AT.AA.A.G .TGAG..GC.
perfG2b         .A.C..AA.. ...G.CCA.G ......CTCA A...A..... ...G.C...C ...A...... .T.....G.. ..GG.TTC.G ..C.CA.... .AT.AA.A.G .TGAG..GC.
perfG2c         ...C..AA.. ...G.CCA.G ......CTCA A...A..... ...G.C...C ...A...... .T.....G.. ..GG.TTC.G ..C.CA.... .AT.AA.A.G .TGAG..GC.
perfG2d         ...C..AA.. ...G.CCA.G ......CTCA A...A..... ...G.C...C ...A...... .T.....G.. ..GG.TTC.G ..C.CA.... .AT.AA.A.G .TGAG..GC.
perfG2e         ...C..AA.. ...G.CCA.G ......CTCA A...A..... ...G.C...C ...A...... .T.....G.. ..GG.TTC.G ..C.CA.... .AT.AA.A.G .TGAG..GC.
perfG2f         ...C..AA.. ...G.CCA.G ......CTCA A...A..... ...G.C...C ...A...... .T.....G.. ..GG.TTC.G ..C.CA.... .AT.AA.A.G .TGAG..GC.
pernG1a         ...C...A.. .....CCA.G ....AACACA A...A..... ...G.C...C ..TAA..... .......G.. ..GG.TTC.G ..C.C..... .AT.AA.A.G ..GAG..GC.
```

```
          2222222222 2222222222 2222222222 2222222222 2233333333 3333333333 3333333333 3333333333 3333333333 3333333334 4444444444
          2223333334 4444445555 5666777778 8888888999 9900000001 1111233333 3445555666 6667777777 7888888888 9999999990 0000000011
          5680124670 3456892568 9147135690 1256789014 7834567892 3589703678 9290147123 4690245678 9012346789 0123456790 1245678901
 pernG1b  ...C...A.. .....CCA.G ....AACACA A...A..... ...G.C...C ..TAA..... .......G.. ..GG.TTC.G ..C.C..... .AT.AA.A.G ..GAG..GC.
 tobbG2   ...CA..A.. .....CCA.G ......CTCA A...A.C.A. .G.G.C...C ...A..... .......G.. ..GG.TCC.G ..C.C..... .AT.AA.ACG ..GAG.TGC.
 kandG1a  ........CA .....CCAAC TA..ATCTCA AA..AT.... ...G.C...C C..AT...AA .T.....GT. ..GGATTC.G ..C.C..... AAT.AA.ACG ..GAG.TGCT
 kandG1b  ........CA .....CCAAC TA..ATCTCA AA.CAT.... ...G.C...C C..AT...AA .T.....GT. ..GGATTC.G ..C.C..... AAT.AA.ACG ..GAG..GCT
 kandG1c  ........CA ....GCCAAC TA..ATCTCA AA..AT.... ...G.C...C C..AT...AA .T.....GT. ..GGATTC.G ..C.C..... AAT.AA.ACG ..GAG.TGCT
 kandG1d  ........CA .....CCAAC TA..ATCTCA AA.CAT.... ...G.C...C C..AT...AA .T.....GT. ..GGATTC.G ..C.C..... AATAAA.ACG ..GAG..GCT
 kandG1e  ........CA .....CCAAC TA..ATCTCA GA.CAT.... ...G.C...C C..AT...AA .T.....GT. ..GGATTC.G ..C.C..... AAT.AA.ACG ..GAG..GCT
 kandG1f  ........CA .....CCAAC TA..ATCTCA AA..AT.... ...GGC...C C..AT...AA .T.....GT. ..GGATTC.G ..C.C..... AAT.AA.ACG ..GAG.TGCT
 kandG1g  ........CA .....CCAAC TA..ATCTCA AA..AT.... ...G.C...C C..AT...AA .T.....GT. ..GGATTC.G ..C.C..... AAT.AA.ACG ..GAG.TGC.
 kandG1h  ........C. .....CCAAC TA..ATCTCA AA.CAT.... ...G.C...C C..AT...AA .T.....GT. ..GGATTC.G ..C.C..... AAT.AA.ACG ..GAG..GCT

          4444444444 4444444444 4444444444 4444444444 4444444444 5555555555 5555
          1112222233 3333333344 4444555555 5666666777 7788888999 0000001111 1122
          2480356901 2345678912 4579234567 9035789124 6735679289 1234570136 7902
 aria193  CTGGAAGCAG CAAGGAAGCG AGCCACATCA TTAGGTCATC GCGAGATAAA TCCTTCTGGA TCGC
 brevG1   .C..G..... ......G... ..G...G... C....C.G.. ......CCC. ...C..C..C ....
 masca01  .C..G..... ....A.G... ......G... C......G.. A.....CCG. ...C..C.A. ...G
 masca02  .C..G..... ....A.G... ......G... C......G.. A.....CCG. ...C..C.A. ...G
 haleG1a  TG........ ..CA..G.T. ..G...G... C......G.. .......GC. ...A..CAC. .T.G
 haleG1c  TG........ ..CA..G.T. ..G...G... C......G.. .......GC. ...A..CAC. .T.G
 haleG1b  TG........ ..CA..G.T. ..G...G... C......G.. .......GC. ...A..C.C. .T.G
 APYa01   .......... .......... .......... .......... .......... .......... ....
 APYa02   .......... T......... .......... .......... .......... .......... ....
 APYa03   .......... T..A...... .....T.... .......... .......... .......... ....
 APYa04   .......... T......... .....T.... .......... .......... .......... ....
 APYa05   .......... T......... .....TG... .......... .......... .......... ....
 APYa06   .......... T......... .....T.... .......... .......... .......... ....
 APYa07   .......... T......... .....T.... .......... .......... .......... ....
 APYa08   .......... T..A...... .....T.... .......... .......... .......... ....
 APYa09   .......... T......... .....T.... .......... A......... .......... ....
 APYa10   .......... T......... .....T.... .......... .......... .......... ....
 APYa11   .......... T......... .....TG... .......... .......... .......... ....
 APYa12   .......... T......... .......... .......... .......... .......... ....
 APYa18   .......... .......... .....TG... .......... .......... .......... ....
 APYa13   .......... .......... .......... .......... .......... .......... ....
 APYa14   .......... .......... ....T..... .......... .......... .......... ....
 APYa15   .......... T......... .......... .......... .......... .....A.... ....
```

```
          4444444444 4444444444 4444444444 4444444444 4444444444 5555555555 5555
          1112222233 3333333344 4444555555 5666666777 7788888999 0000001111 1122
          2480356901 2345678912 4579234567 9035789124 6735679289 1234570136 7902
APYa16    .......... T..A...... .......... .......... .......... .......... ....
APYa17    .......... T......... .......... .......... .......... .......... ....
APYa19    .......... .......... .....TG... .......... .......... .......... ....
APYa20    .......... T......... .......... .......... .......... .......... ....
APYa21    .......... .......... .......... .......... .......... .......... ....
APYa22    .......... T......... .......... ...A...... .......... .......... ....
APYa23    .......... T......... .......... .......... .......... .......... ....
APYa24    .......... T.........  .....T.... .......... .......... .......... ....
APYa25    .......... T.........  .....T.... .......... .......... .......... ....
APYa26    .......... T.........  .....T.... .......... .......... .......... ....
APYa27    .......... T..A......  .....T.... .......... .......... .......... ....
APYa28    .......... .......... .......... .......... .......... .......... ....
APYa29    .......... T......... .......... .......... .......... .......... ....
APYa30    .......... T.........  .....T.... .......... .......... .......... ....
APYa31    .......... T.........  .....T.... .......... .......... .......... ....
APYa32    .......... .......... .....T.... .......... .......... .......... ....
APYa33    .......... T.........  .....T.... .......... .......... .......... ....
APYa34    .......... T.........  .....T.... .......... .TC....... ........T. ....
APYa35    .......... .......... .....T.... .......... .......... .......... ....
APYa36    .......... .......... .....T.... .......... .......... .......... ....
APYa37    .......... T.........  .....T.... .......... ..A....... .......... ....
APYa38    .......... T.........  .....T.... .......... .TC....... .......... ....
APYa39    .......... T.........  .....T.... .......... .......... .......... ....
APYa40    .......... .......... .....T.... .......... .......... .......... ....
APYa41    .......... T.........  ......G... .......... .......... .......... ....
APYa42    .......... T.........  .......... .......... .......... .......... ....
APYa43    .......... T..A......  .....T.... .......... .......... .......... ....
APYa44    .......... .......... .......... .......... .......... .......... ....
APYa45    .......... T.........  .......... .......... .T........ .......... ....
APYa46    .......... T.........  .......... .......... .......... .....A.... ....
APYa47    .......... T.........  .......... .......... .......... .....A.... ....
APYa48    .......... T.........  .....T.... .......... .TA....... .......... ....
APYa49    .......... T.........  .......... .......... .......... .......... ....
APYa50    .......... T.........  .....TG... .......... .......... .......... ....
APYa51    .......... T..A......  .....T.... .......... .......... .......... ....
majoG1a   .C..G.A..C .GCC....G.  ......G... C......G.. ......CC.. C..CCAA.C. .T..
majoG1b   .C..G.A..C .GCC....G.  ......G... C......G.. ......CC.. C..CCAA.C. .T..
neg1G1a   .C..G.A..C .GCC....G.  ........T. C...A..... ......CC.. C...C.A.C. .T..
neg1G1b   .C..G.A..C .GCC....G.  .......... C......G.. ......CC.. C...C.A.T. .T..
perfG1a   .CAAG.ATGA ...AA.T...  ...TG.....  ..G.A.....  AG..AGCC..  .AGC.T---.  ..A.
perfG1b   .CAAG.ATGA ...AA.T...  ...TG.....  ..G.A.....  AG..AGCC..  .AGC.T---.  ..A.
pernG2a   .C..GTA.GA .C.CA.T...  ..A.G.....  ..G.A.....  AG..AGCC..  GAGC.T---.  ....
pernG2b   .C..GTA.GA .C.CA.T...  ..A.G.....  ..G.A.....  AG..AGCC..  GAGC.T---.  ....
pernG2c   .C..GTA.GA .C.CA.T...  ..A.G.....  ..G.A.....  AG..AGCC..  GAGC.T---.  ...A
tobbG1a   .C..G.A.GA ...AA.T...  ...TG...T.  ..G.A....T  AG..AG.C..  .AGC.T---.  ....
tobbG1b   .C..G.A.GA ...AA.T...  ...TG...T.  ..G.A....T  AG..AG.C..  .AGC.T---.  ....
tobbG1c   .C..G.A.GA ...AA.T...  ...TG...T.  ..G.A....T  AG..AG.C..  .AGC.T---.  ....
tobbG1d   .C..G.A.GA ...AA.T...  ...TG...T.  .CG.A....T  AG..AG.C..  .AGC.T---.  ....
tobbG1e   .C..G.A.GA ...AA.T...  ...TG...T.  ..G.A....T  AG..AG.C..  .AGC.T---.  ....
tobbG1f   .C..G.A.GA ...AA.T...  ...TG...T.  ..G.A....T  AG..AG.C..  .AGC.T---.  ....
kandG2a   .C..G.A.GA ...AA.....  .........T  ..GAA.....  AG....CC..  CA.C.T---.  ....
kandG2b   .C..G.A.GA ...AA.....  .........T  ..GAA.....  AG....CC..  CA.C.T---.  ....
perfG2a   .C..G.A...  .G.ATC..TC CA...T.AGT  ..GT..G...  AA.G.GCC.C  CTTC..---.  ....
perfG2b   .C..G.A...  .G.ATC..TC CA...T.AGT  ..GT..G...  AA.G.GCC.C  CTTC..---.  ....
perfG2c   .C..G.A...  .G.ATC..TC CA...T.AGT  ..GT..G...  AA.G.GCC.C  CTTC..---.  ....
perfG2d   .C..G.A...  .G.ATC..TC CA...T.AGT  ..GT..G.C.  AA.G.GCC.C  CTTC..---.  ....
perfG2e   .C..G.....  .G.ATC..TC CA...T.AGT  ..GT..G...  AA.G.GCC.C  CTTC..---.  ....
perfG2f   .C..G.A...  .G.ATC..TC CA...T.AGT  ..GT..G...  AA.G.GCC.C  CTTC..---.  ....
pernG1a   .C..G.A...  .G.ATCT..C C....T.AAT  ..GT......  AG.G.GCC.C  CGTC..---.  C...
pernG1b   .C..G.A...  .G.ATCT..C C....T.AAT  ..GT......  AG.G.GCC.C  CGTC..---.  ....
tobbG2    .C..G.A...  .G.ATC.TTC CA.....AAT  ..GTAAG...  AA.G.GCC.C  CTTC..---.  ....
kandG1a   .C..G.A...  .G.ATC...C CA.....AGT  ..GT..G...  AG.G.GCC.C  GATC..---.  .T..
kandG1b   .C..G.A...  .G.ATC...C CA.....AGT  ..GT..G...  AG.G.GCC.C  GATC..---.  .T..
kandG1c   .C..G.A...  .G.ATC...C CA.....AGT  ..GT..G...  AG.G.GCC.C  GATC..---.  .T..
kandG1d   .C..G.A...  .G.ATC...C CA.....AGT  ..GT..G...  AG.G.GCC.C  GATC..---.  .T..
kandG1e   .C..G.A...  .G.ATC...C CA.....AGT  ..GT..G...  AG.G.GCC.C  GATC..---.  .T..
```

Appendix 3.3 Figure Continued.

```
          4444444444 4444444444 4444444444 4444444444 4444444444 5555555555 5555
          1112222233 3333333344 4444555555 5666666777 7788888999 0000001111 1122
          2480356901 2345678912 4579234567 9035789124 6735679289 1234570136 7902
kandG1f   .C..G.A... .G.ATC...C CA.....AGT ..GT..G... AG.G.GCC.C GATC..---. .T..
kandG1g   .C..G.A... .G.ATC...C CA.....AGT ..GT..G... AG.G.GCC.C GATC..---. .T..
kandG1h   .C..G.A... .G.ATC...C CA.....AGT ..GT..G... AG.G.GCC.C GATC..---. .T..
```

Appendix 3.4 Table Models used to estimate pairwise values of d_N and d_S for protein coding locus APY, where $d_S < 0.5$ indicates non-saturation of synonymous substitutions and an appropriate outgroup of the MK population test for selection. d_S estimated under the approximate Nei and Gojobori method[¥] (1986) (with Jukes-Cantor correction), and in PAML CODEML runmode -2 according to the maximum likelihood method of Goldman and Yang[§] (1994).

Outgroup species	*P. ariasi* allele	d_N[¥]	d_S[¥]	d_N[§]	d_S[§]
masca01	APYa01	0.0783	0.9415	0.0781	0.9563
masca01	APYa30	0.0752	1.0563	0.0751	1.0755
haleG1a	APYa01	0.0719	0.8634	0.0718	0.8769
haleG1a	APYa30	0.0689	1.0462	0.0688	1.066
arab632	APYa01	0.0702	0.8331	0.0701	0.8425
arab632	APYa30	0.0732	1.0099	0.0731	1.0239
majoG1b	APYa01	0.0716	0.4114	0.0715	0.4145
majoG1b	APYa30	0.0746	0.4475	0.0745	0.4511
neglG1b	APYa01	0.0632	0.4592	0.0631	0.4636
neglG1b	APYa30	0.0602	0.5173	0.0601	0.5227
perfG1a	APYa01	0.1405	0.7646	0.1402	0.7736
perfG1a	APYa30	0.1438	0.8227	0.1435	0.8335
pernG2a	APYa01	0.1461	0.6582	0.1458	0.6648
pernG2a	APYa30	0.1495	0.7338	0.1492	0.742
tobbG1e	APYa01	0.1441	0.6794	0.1437	0.6874
tobbG1e	APYa30	0.1474	0.7047	0.1471	0.7136
kandG2b	APYa01	0.1049	0.6338	0.1047	0.6406
kandG2b	APYa30	0.1081	0.7076	0.1078	0.7162
perfG2f	APYa01	0.2114	0.6809	0.2106	0.6946
perfG2f	APYa30	0.2151	0.6088	0.2142	0.6208
pernG1a	APYa01	0.2224	0.7065	0.2215	0.7225
pernG1a	APYa30	0.2261	0.6561	0.2252	0.6707
tobbG2	APYa01	0.2353	0.5558	0.2344	0.5662
tobbG2	APYa30	0.2391	0.6228	0.2381	0.6356
kandG1f	APYa01	0.2218	0.8012	0.2215	0.8093
kandG1f	APYa30	0.2255	0.8317	0.2251	0.8409
APYa01	APYa30	0.0028	0.0524	0.0028	0.0528

Legend Species codes denote the first four letters of the formal species name (see Appendix 2.3).

Appendix 3.5 Table Population pairwise F_{ST} estimates (using allele frequencies) (below the diagonal) and significance level (above the diagonal) at locus apyrase, in 20 natural populations of *P. ariasi.* Levels of significance, after Bonferroni correction in FSTAT (v2.9.3.2): * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; NS not significant.

	AGH	ARQ06	ARQ08	HP1	CAT	CHR	CSP	CTU	DRAz4	IRL07	LNP	HP2	MLQ	PAS	PLB	RME	SAM13	SPV	TRJ	TUL
AGH	-	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**
ARQ06	0.2942	-	NS	NS	NS	**	**	NS	NS	NS	NS	NS	NS	*	NS	NS	NS	NS	*	NS
ARQ08	0.2213	0.0308	-	NS	NS	**	**	NS	NS	NS	**	NS	NS	NS	NS	*	NS	NS	NS	NS
HP1	0.1875	0.1238	0.0145	-	NS	**	**	NS	**	NS	**	NS	NS	NS	NS	**	NS	NS	NS	NS
CAT	0.2421	0.002	-0.0158	0.0369	-	**	**	NS	NS	NS	**	NS	NS	NS	NS	*	NS	NS	NS	NS
CHR	0.1446	0.2704	0.198	0.1611	0.2141	-	**	**	**	**	**	**	**	**	**	**	**	**	**	**
CSP	0.2492	0.3345	0.2425	0.1917	0.2653	0.1843	-	**	**	**	**	**	**	**	**	**	**	**	**	**
CTU	0.2458	0.0115	-0.0152	0.0381	-0.0216	0.2211	0.267	-	NS	NS	**	NS	NS	NS	NS	*	NS	NS	**	NS
DRAz4	0.2978	0.0053	0.0526	0.1331	0.012	0.2678	0.3153	0.0278	-	NS	NS	NS	NS	**	*	NS	NS	NS	**	NS
IRL07	0.2596	-0.0093	0.0033	0.0758	-0.0046	0.2376	0.2989	-0.0031	0.0391	-	*	NS	NS	NS	NS	NS	NS	NS	NS	NS
LNP	0.5394	0.1276	0.288	0.3982	0.2719	0.4956	0.5822	0.2508	0.1911	0.1961	-	**	**	**	**	NS	**	**	**	**
HP2	0.2562	0.0485	-0.0104	0.0292	0.0051	0.2289	0.2837	-0.0021	0.0963	0.0076	0.3314	-	NS	NS	NS	*	NS	NS	NS	NS
MLQ	0.2071	0.0688	-0.0107	0.0003	0.0042	0.1766	0.1955	0.0021	0.0856	0.0297	0.3366	-0.0035	-	NS	NS	**	NS	NS	NS	NS
PAS	0.2071	0.1517	0.0425	0.0021	0.0566	0.1756	0.174	0.0617	0.1343	0.1135	0.399	0.0741	0.0251	-	NS	**	NS	NS	NS	NS
PLB	0.2313	0.1465	0.0259	-0.0074	0.0436	0.1999	0.2086	0.047	0.1368	0.1023	0.4459	0.0505	0.01	-0.0081	-	**	NS	NS	NS	NS
RME	0.5197	0.1359	0.2884	0.3913	0.274	0.4771	0.5693	0.2543	0.1896	0.2058	-0.0208	0.3371	0.3342	0.3962	0.439	-	*	NS	**	*
SAM13	0.2895	0.0725	0.0469	0.0851	0.017	0.2562	0.2854	0.0337	0.0169	0.083	0.3379	0.0999	0.065	0.0649	0.0644	0.332	-	NS	**	NS
SPV	0.2544	0.0403	0.007	0.0445	-0.0143	0.2237	0.2484	-0.0002	0.011	0.039	0.3053	0.0455	0.0214	0.0398	0.0314	0.3	-0.0169	-	NS	NS
TRJ	0.2033	0.2044	0.0659	0.0001	0.1065	0.1754	0.1988	0.1036	0.2273	0.1385	0.4952	0.0724	0.0373	0.0273	0.0218	0.4863	0.1744	0.1213	-	**
TUL	0.2704	-0.0023	-0.0001	0.0692	-0.0172	0.2428	0.2856	-0.0131	0.009	-0.0023	0.2161	0.0181	0.0234	0.0879	0.0773	0.223	0.0311	0.002	0.1451	-

Appendix 4.1 Table Genotype frequencies of locus AAm20 in a fine-scale geographical analysis of populations of *P. ariasi* from southwest France.

Region	Eastern Pyrenees															S Massif Central	
Sub-Region	West Other								Foret de la Malepere				East Aude				
Population	PAS	PLB	MZE	MTD	IRL07	IRL08	TUL	MLQ	MQT	RUL06	RUL08	SJL	SHL	ARQ06	ARQ08	CTU	SPV
0101	0.4583	0.3478	0.7647	0.6250	0.4545	0.5217	0.4583	0.5417	0.6957	0.4375	0.4167	0.5217	0.5	0.4167	0.3913	0.3333	0.2917
0102	0.3750	0.5217	0.2353	0.2917	0.4545	0.4348	0.4167	0.3750	0.3043	0.4375	0.5	0.3913	0.3333	0.5417	0.4783	0.5000	0.4167
0202	0.1667	0.0870		0.0833			0.0833	0.0833		0.0625	0.0833	0.0435	0.1667	0.0417	0.0870	0.1667	0.2500
0103		0.0435				0.0435						0.0435			0.0435		
0114					0.0455					0.0625							
0203							0.0417										
0115					0.0455												
0116																	0.0417

Appendix 4.2 Table Genotype frequencies of locus AAm24 in a fine-scale geographical analysis of populations of *P. ariasi* from southwest France.

Region	Eastern Pyrenees															S Massif Central	
Sub-Region	West Other								Foret de la Malepere				East Aude				
Population	PAS	PLB	MZE	MTD	IRL07	IRL08	TUL	MLQ	MQT	RUL06	RUL08	SJL	SHL	ARQ06	ARQ08	CTU	SPV
Genotype																	
0606	0.5417	0.5	0.2941	0.4167	0.2727	0.4167	0.4583	0.5217	0.5652	0.4118	0.1667	0.5417	0.5833	0.6250	0.3913	0.5833	0.5417
0106	0.1667	0.1250	0.1176	0.2500	0.2273	0.1250	0.2917	0.2174	0.0870	0.2353	0.1667		0.1667	0.1250	0.3043	0.1667	0.2083
0607	0.0833	0.1667	0.3529	0.1250	0.1818	0.2083	0.0417	0.0870	0.2174	0.1765	0.4167	0.3333	0.0417	0.2083	0.0870	0.1667	0.1667
0107		0.0417		0.0833	0.0455	0.0833	0.0417		0.0870		0.0833		0.0417		0.0435	0.0417	0.0417
0707	0.0417	0.0417	0.0588		0.1364		0.0417				0.0833				0.0435	0.0417	0.0417
0608							0.0417				0.0833			0.0417	0.0870		
0109	0.0417							0.0435					0.0417		0.0435		
0609	0.0833	0.0833	0.1176	0.0417		0.0417		0.1304	0.0435				0.1250				
0101				0.0417		0.1250	0.0417			0.1765							
0108					0.0455							0.0417					
0709			0.0588		0.0909		0.0417										
0611	0.0417																
0613		0.0417															
0712				0.0417													
0708												0.0833					

Appendix 4.3 Table Genotype frequencies of locus APY in a fine-scale geographical analysis of populations of *P. ariasi* from southwest France.

Region	Eastern Pyrenees																
Sub-Region	West Other								Foret de la Malepere				East Aude			S Massif Central	
Population	PAS	PLB	MZE	MTD	IRL07	IRL08	TUL	MLQ	MQT	RUL06	RUL08	SJL	SHL	ARQ06	ARQ08	CTU	SPV
Genotype																	
0101	0.0435	0.0417	0.1176	0.0435	0.1818	0.2174	0.2083	0.0833	0.4348	0.2353	0.1667	0.3478	0.1667	0.4545	0.1304	0.1667	0.3043
0103	0.1304	0.1667	0.2353	0.1304	0.3182	0.3043	0.2500	0.2500	0.2174	0.0588	0.0833	0.2174	0.1667	0.0909	0.2609	0.2917	0.0870
0102	0.0652	0.1667	0.3529	0.1304	0.1818	0.2609	0.3333	0.2083	0.1304	0.1765	0.4167	0.0435	0.3333	0.1364	0.1739	0.2500	0.1739
0203	0.1957	0.2083	0.2353	0.2174	0.0455	0.0870	0.1250	0.0833	0.0870	0.3529		0.0435	0.0417	0.0455	0.1304	0.0417	0.1304
0202	0.1522	0.2083	0.0588	0.0870		0.0435		0.0417	0.0435	0.1176	0.1667		0.0417	0.0455	0.0435	0.0833	0.1739
0303	0.0435	0.0417		0.1304	0.0455	0.0435		0.0833	0.0870			0.0870		0.0455	0.0870	0.0417	0.0435
0215	0.0435																0.0435
0208	0.0435							0.0417									0.0435
0321	0.0217															0.0417	
0115	0.0435				0.0909										0.0435		
0120					0.0455									0.0455			
0315		0.0833										0.0870		0.0909			
0108				0.0435								0.0435	0.0833				
0205								0.0417					0.0417				
0117										0.0588			0.0417				
0121							0.0417						0.0417				
0308	0.0217	0.0417		0.0870			0.0417	0.0833			0.0833	0.0435					
0220	0.0435																
0320	0.0217																
0221	0.0435																
0219	0.0435																
1523	0.0217																
0521	0.0217																
0351		0.0417															
0150				0.0435													
0342				0.0870													
0119					0.0455												
0329					0.0455												
0343						0.0435											
0344								0.0417									
2344								0.0417									
0104											0.0833						
0105												0.0435					
0141												0.0435					
0128													0.0417				
0129														0.0455			
0247															0.0435		
0127															0.0435		
2328															0.0435		
0112																0.0417	
0323																0.0417	

Appendix 4.4 Table Genotype frequencies of locus EF-1α in a fine-scale geographical analysis of populations of *P. ariasi* from southwest France

Region	Eastern Pyrenees															S Massif Central	
Sub-Region	West Other								Foret de la Malepere				East Aude				
Population	PAS	PLB	MZE	MTD	IRL07	IRL08	TUL	MLQ	MQT	RUL06	RUL08	SJL	SHL	ARQ06	ARQ08	CTU	SPV
Genotype																	
0103	0.2500	0.3810	0.2941	0.1429	0.2273	0.5000	0.2174	0.2083	0.1250	0.5333	0.4167	0.1304	0.5417	0.2143	0.1739	0.1250	0.3913
0101	0.5208	0.3333	0.4118	0.3333	0.3182	0.2083	0.3478	0.3333	0.3750	0.2667	0.3333	0.1304	0.2083	0.2857	0.3043	0.3750	0.2174
0102	0.0417	0.1429	0.0588	0.0952	0.0455	0.0417		0.1250	0.1250		0.0833	0.1739	0.0417		0.1304	0.2083	0.0870
0303	0.0208		0.0588		0.1818	0.0417	0.1739	0.0417				0.0870		0.2857	0.0870	0.0833	0.0870
0105		0.0476					0.0870	0.0833	0.0417	0.0667		0.2609	0.0417	0.0714	0.1304		0.0435
0203				0.0476	0.0909				0.1667		0.0833		0.0833	0.0714		0.1250	0.0435
0305				0.0476	0.0455	0.0417	0.0435		0.0417				0.0417		0.0435		0.0435
0306			0.0588			0.1250						0.0435				0.0417	
0148				0.0476					0.0833								
0205					0.0455								0.0417				
0138								0.0417							0.0435		
0106	0.0417						0.0870				0.0833	0.0870					
0110	0.0208					0.0417		0.0417									
0112	0.0208			0.0952				0.0833									
0114	0.0208				0.0455												
0113	0.0208	0.0476															
0108	0.0208																
0215	0.0208																
0313		0.0476															
0210			0.0588														
0349			0.0588														
0314				0.0476													
0312				0.0476													
0151				0.0476													
1050				0.0476													
0512							0.0435										
0208								0.0417									
0348									0.0417								
0346										0.0667							
0147										0.0667							
0506												0.0435					
0142												0.0435					
0139													0.0714				
0202														0.0870			
0229															0.0417		
0116																0.0435	
0141																0.0435	

Appendix 4.5 Table Population pairwise Φ_{ST} estimates (using allele frequency and divergence data) (below the diagonal) and significance level (above the diagonal) at three nuclear loci characterized from 17 populations of *P. ariasi* from southwest France. (a) AAm24, (b) APY, (c) EF-1α. Levels of significance calculated in ARLEQUIN (v3.11) after 1,000 permutations: * $P < 0.05$; ** $P < 0.01$; $P < 0.001$; NS not significant. None were significant after manual sequential Bonferroni correction (Holm, 1979).

(a)

	PAS	PLB	MZE	MTD	IRL07	IRL08	TUL	MLQ	MQT	RUL06	RUL08	SJL	SHL	ARQ06	ARQ08	CTU	SPV
PAS	-	NS	NS	NS	**	NS	NS	NS	NS	NS	*	NS	NS	NS	NS	NS	NS
PLB	-0.01339	-	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
MZE	0.04357	0.01536	-	NS	NS	NS	NS	*	NS	NS	NS	NS	*	*	NS	NS	NS
MTD	-0.00257	-0.01089	0.00286	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
IRL07	0.09263	0.05721	-0.01596	0.03267	-	NS	*	**	*	NS	NS	*	**	**	*	**	*
IRL08	0.01574	0.00024	-0.00412	-0.0178	0.01526	-	NS	NS	NS	NS	NS	NS	NS	*	NS	NS	NS
TUL	-0.0069	-0.01177	0.01713	-0.01887	0.04857	-0.01296	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
MLQ	-0.01751	-0.00297	0.05817	0.00327	0.10878	0.02437	-0.00125	-	NS	NS	**	NS	NS	NS	NS	NS	NS
MQT	-0.0115	-0.02039	0.02057	-0.00801	0.06284	0.00324	-0.01021	0.00218	-	NS	NS	NS	NS	NS	NS	NS	NS
RUL06	0.00698	-0.0005	0.01662	-0.02015	0.03906	-0.02013	-0.02047	0.01192	0.0028	-	NS	NS	**	**	NS	NS	NS
RUL08	0.09524	0.0509	-0.02056	0.03402	-0.02771	0.01553	0.04856	0.12039	0.05489	0.0446	-	NS	NS	NS	NS	NS	NS
SJL	0.01107	-0.00621	0.02656	0.01445	0.06394	0.02357	0.00996	0.03188	-0.01016	0.02974	0.04311	-	NS	NS	NS	NS	NS
SHL	-0.01718	-0.00164	0.0648	0.00716	0.11686	0.02954	0.0017	-0.02153	0.00313	0.01662	0.12865	0.03241	-	NS	NS	NS	NS
ARQ06	-0.00085	0.00259	0.09215	0.03642	0.14519	0.06009	0.02394	0.01411	-0.00137	0.05426	0.14305	0.00389	0.01112	-	NS	NS	NS
ARQ08	-0.00721	-0.01225	0.01643	-0.01726	0.04725	-0.01098	-0.02093	-0.00107	-0.01098	-0.01766	0.04454	0.00399	0.00163	0.01941	-	NS	NS
CTU	-0.00858	-0.01746	0.03282	-0.00236	0.07556	0.0096	-0.00689	0.00706	-0.02077	0.0072	0.06713	-0.00959	0.0074	-0.00505	-0.00783	-	NS
SPV	-0.00812	-0.01814	0.02481	-0.00815	0.06435	0.00143	-0.01157	0.00658	-0.02081	-0.00109	0.05639	-0.00795	0.00763	0.00135	-0.01193	-0.02058	-

Appendix 4.5 Table Continued: Population pairwise Φ_{ST} estimates for loci: (b) APY, (c) EF-1α.

(b)

	PAS	PLB	MZE	MTD	IRL07	IRL08	TUL	MLQ	MQT	RUL06	RUL08	SJL	SHL	ARQ06	ARQ08	CTU	SPV
PAS	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	**	**	NS	NS	*
PLB	-0.00895	-	NS	NS	NS	NS	*	NS	*	NS	*	NS	**	**	NS	NS	*
MZE	0.00338	0.01655	-	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
MTD	0.02227	-0.00484	0.05801	-	*	*	**	NS	**	*	**	NS	***	***	NS	*	**
IRL07	0.02854	0.04019	-0.01527	0.07145	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
IRL08	0.01889	0.02379	-0.01775	0.05065	-0.01818	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
TUL	0.03753	0.05423	-0.01673	0.09336	-0.01868	-0.01495	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
MLQ	0.00301	-0.00927	0.00574	-0.00325	0.01406	0.00024	0.02735	-	NS	NS	*	NS	*	*	NS	NS	*
MQT	0.0543	0.06994	-0.00597	0.10391	-0.01808	-0.01241	-0.01915	0.03485	-	NS	NS	NS	NS	NS	NS	NS	NS
RUL06	0.00293	0.01916	-0.02844	0.06579	-0.00907	-0.01077	-0.01123	0.01356	0.00202	-	NS	NS	NS	NS	NS	NS	NS
RUL08	0.0446	0.07935	-0.01132	0.13681	0.00201	0.0139	-0.00719	0.06443	0.00521	-0.01364	-	NS	NS	NS	NS	NS	NS
SJL	0.02831	0.02897	-0.00478	0.04536	-0.01595	-0.01671	-0.0062	-0.00089	-0.0076	0.0057	0.03092	-	NS	NS	NS	NS	NS
SHL	0.07232	0.10623	0.00414	0.15915	0.00215	0.01724	-0.00627	0.07927	-0.00219	0.00528	-0.02346	0.02995	-	NS	NS	NS	NS
ARQ06	0.08007	0.11215	0.00996	0.16242	-0.00154	0.01642	-0.00681	0.0818	-0.00691	0.0128	-0.01393	0.02567	-0.01823	-	NS	NS	NS
ARQ08	0.00129	-0.00083	-0.0128	0.02124	-0.0071	-0.01511	0.00134	-0.01236	0.00751	-0.00747	0.03137	-0.01105	0.04224	0.04209	-	NS	NS
CTU	0.01561	0.02246	-0.01719	0.05281	-0.01477	-0.01927	-0.01176	0.00318	-0.00783	-0.01266	0.01621	-0.01033	0.02111	0.02021	-0.01535	-	NS
SPV	0.03639	0.06595	-0.01331	0.12364	0.00267	0.0106	-0.0048	0.0549	0.00809	-0.01463	-0.02946	0.0273	-0.00999	-0.00228	0.02449	0.01261	-

(c)

	PAS	PLB	MZE	MTD	IRL07	IRL08	TUL	MLQ	MQT	RUL06	RUL08	SJL	SHL	ARQ06	ARQ08	CTU	SPV
PAS	-	NS	NS	*	***	*	***	NS	*	NS	NS	***	***	***	**	**	***
PLB	0.00707	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
MZE	0.02033	-0.01592	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
MTD	0.0449	-0.00375	-0.01436	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
IRL07	0.11706	0.03632	0.01744	0.00456	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
IRL08	0.04794	0.00095	-0.01558	-0.01641	-0.00201	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
TUL	0.05441	0.01043	0.00242	-0.00802	0.00245	-0.01201	-	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
MLQ	0.00455	-0.01378	-0.0172	-0.0062	0.03321	-0.00627	-0.00148	-	NS	NS	NS	NS	NS	NS	NS	NS	NS
MQT	0.03306	-0.00728	-0.02089	-0.01808	0.01148	-0.01673	-0.00581	-0.01334	-	NS	NS	NS	NS	NS	NS	NS	NS
RUL06	0.03558	0.00291	0.00645	0.00728	0.01727	-0.00038	-0.01439	-0.00378	0.0055	-	NS	NS	NS	NS	NS	NS	NS
RUL08	0.00295	-0.02453	-0.03437	-0.02038	0.01909	-0.02171	-0.0065	-0.02789	-0.02704	-0.00431	-	NS	NS	NS	NS	NS	NS
SJL	0.08453	0.02369	0.01126	-0.00529	0.02261	0.00931	0.00734	0.01482	-0.00128	0.0295	0.00773	-	NS	NS	NS	NS	NS
SHL	0.09083	0.0206	0.00645	-0.00069	-0.01867	-0.00869	-0.00826	0.01651	0.00137	0.00131	0.00535	0.01459	-	NS	NS	NS	NS
ARQ06	0.12966	0.04576	0.02959	0.01142	-0.02322	0.00185	-0.00022	0.04099	0.02146	0.01191	0.03117	0.03416	-0.01888	-	NS	NS	NS
ARQ08	0.06935	0.00489	-0.00769	-0.00908	-0.0064	-0.00884	-0.00325	0.00198	-0.01086	0.00744	-0.00881	-0.00578	-0.01257	0.00318	-	NS	NS
CTU	0.07215	0.01086	-0.01215	-0.00773	0.01736	-0.00238	0.0252	0.01109	-0.00933	0.04427	-0.01127	0.00896	0.01469	0.03522	-0.00453	-	NS
SPV	0.08847	0.017	0.00178	-0.00414	-0.01822	-0.00959	-0.00606	0.01404	-0.00256	0.00517	0.00122	0.00766	-0.02034	-0.01527	-0.01638	0.00595	-