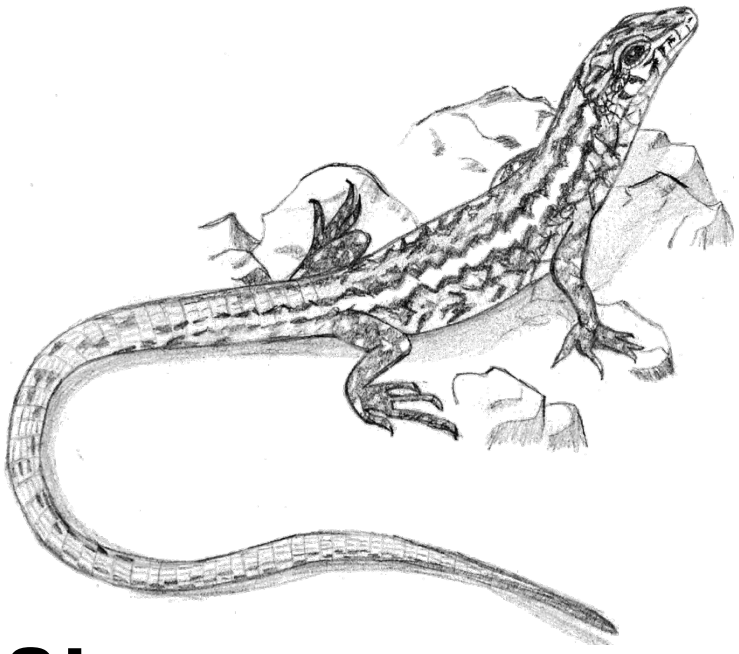


Multilocus Phylogenetics: Inferring the Species-Tree of the Iberian and North African *Podarcis* Wall Lizards

Alvarina Santos Couto

Mestrado em Biodiversidade, Genética e Evolução

Departamento Biologia

Centro de Investigação em Biodiversidade e Recursos Genéticos (CIBIO)

2014

Orientadora

Catarina Pinho, Post-Doc, Centro de Investigação em Biodiversidade e Recursos Genéticos (CIBIO), Universidade do Porto.

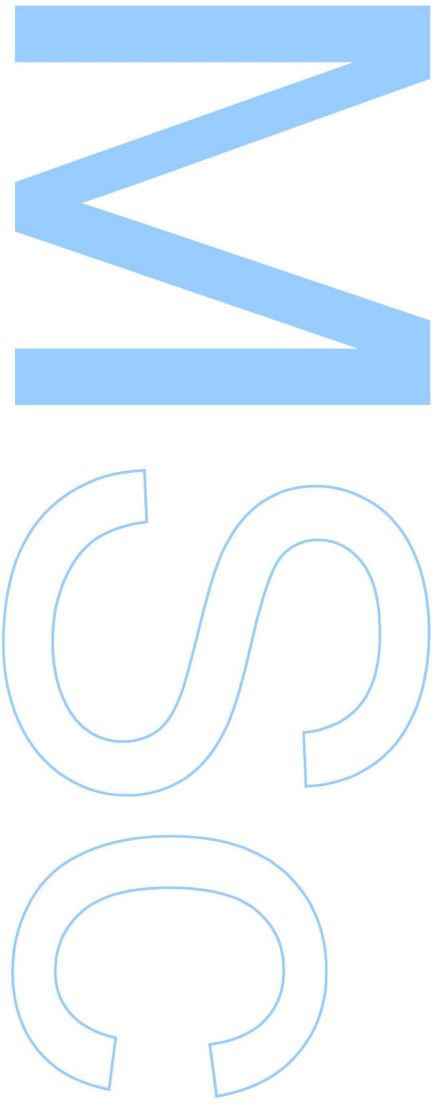
Coorientadora

Sara Rocha, Post-Doc, Centro de Investigação em Biodiversidade e Recursos Genéticos (CIBIO), Universidade do Porto e Departamento de Xenética, Bioquímica e Inmunoloxía, Facultad de Bioloxía, Universidade de Vigo.

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____



I think

Charles Darwin

Acknowledgements

A realização desta dissertação não teria sido possível sem a ajuda e apoio de várias pessoas.

Em primeiro lugar gostaria de agradecer de forma especial à minha orientadora, Catarina Pinho, e co-orientadora, Sara Rocha, que despertaram o meu interesse pela genética e pelas *Podarcis* e me acolheram no projeto que deu origem a esta tese. Por todo o entusiasmo contagiante e ensinamentos, todo o interesse e disponibilidade e por terem sido dos melhores orientadores que poderia ter tido.

Também gostaria de expressar um agradecimento especial à "equipa *Podarcis*" do CIBIO pelo input. Em especial à Antigoni Kaliontzopoulou, ao Miguel A. Carretero e ao James Harris pela partilha dos seus conhecimentos, discussões e comentários que ajudaram no meu trabalho. Agradeço a todos os que colaboraram nesta tese cedendo amostras e à Carolina Pereira e à Carla Luís pela contribuição na correção manual das sequências. Ao Pedro Tarroso pelos scripts e disponibilidade. Aos restantes colegas do CIBIO agradeço a boa-disposição, amizade e entreajuda.

Agradeço a todos os meus colegas de mestrado, em especial ao Pedro Cardoso e à Mariana Ribeiro, por todas as conversas e apoio.

Ao Bruno, por todo carinho e apoio, que sempre aturou as minhas mudanças de humor e me ouviu a falar de lagartixas mesmo sem entender o que eu dizia. Obrigada também por todos os scripts e ajudas em programas que ateimavam em não funcionar.

Finalmente, à minha família por todo o apoio. Aos meus pais, irmã e às minhas tias que sempre me apoiaram e sem os quais aqui não tinha chegado.

I would also like to express my gratitude to the Group of Phylogenomics at the University of Vigo that provided a computational cluster of extreme importance to perform the large-scale analyses of this thesis. Thank you David, for hosting me for two weeks and for helping me with guidance and comments regarding my work. I must also acknowledge Diego Mallo, Leonardo Martins and Ramon for their help in the important analyses of this thesis.

This work fits in the framework of the research project "On the road to speciation: an integrated analysis of the evolution of reproductive isolation in a cryptic species complex." - COMPETE and by national funds through FCT - PTDC/BIA-BEC/102179/2008; PI: Catarina Pinho.

Resumo

Avanços recentes no campo da filogenética permitem uma inferência abrangente dos processos evolutivos a partir de dados multilocus. Inferir relações evolutivas entre espécies, especialmente quando estas divergiram recentemente, ou muito rapidamente, oferece desafios significativos. Este é o caso das lagartixas do género Podarcis da Península Ibérica e do Norte de África, cuja taxonomia permaneceu controversa mesmo apesar dos vários estudos baseados em DNA mitocondrial (mtDNA), aloenzimas e morfologia.

Neste trabalho foram usados 30 genes nucleares para 170 indivíduos representativos de todos morfotipos e linhagens de DNA mitocondrial conhecidas, a fim de avaliar os níveis de polimorfismo genético e inferir a árvore de espécies das lagartixas do género Podarcis da Península Ibérica e do Norte de África. Para isso, foi utilizada uma variedade de métodos para estimar a árvore das espécies abrangendo métodos estatísticos de distancia (NJst) e máxima pseudo-verosimilhança (MP-EST), métodos Bayesianos de "supertree" (Guenomu) e probabilísticos Bayesianos (*BEAST). Todos estes têm em conta a persistência de polimorfismo ancestral, processo que se pensa ser a principal causa da incongruência entre árvores de genes e árvores de espécies no caso das Podarcis.

As sequências nucleares apresentam altos níveis de partilha de haplótipos em praticamente todos os genes entre as linhagens de DNA mitocondrial previamente definidas. De acordo com estes genes e uma análise de agrupamento Bayesiana que atende à maximização do equilíbrio de Hardy-Weinberg dentro de cada grupo, as 17 linhagens de DNA foram subdivididas em 24 grupos geneticamente distintos. Além disso, também foi detectado fluxo génico entre grupos através da observação de genótipos miscigenados; alguns destes casos correspondem a exemplos previamente documentados (ex. P. bocagei/P. guadarramae lusitanica), mas outros correspondem à descrição de fluxo génico pela primeira vez (P. vaucheri "Spain"/P. carbonelli; P. vaucheri "Spain"/P. virescens; P. hispanica "Galera"/P. hispanica "Albacete/Murcia" e P. liolepis/P. hispanica sensu stricto/P. atrata). Algumas situações de clara discordância citonuclear foram reveladas. A fim de remover evidências claras de fluxo génico recente, indivíduos com uma proporção de menos de 95% do seu genoma atribuído a um único grupo foram excluídos para a inferência filogenética.

As árvores das espécies obtidas corroboraram algumas das relações inferidas pelo mtDNA, mas também revelaram outras completamente diferentes. Um grupo de formas do Este da Península Ibérica (P. hispanica sensu stricto, P. liolepis e P. atrata) foi consistentemente recuperado pelos diferentes métodos, com altos valores de suporte e com apenas pequenas variações nas relações entre linhagens. O clado composto por formas de P. hispanica do Norte

de África, proposto pelo mtDNA, é confirmado pelos dados nucleares, enquanto P. vaucheri tende a ser um clado nas estimativas baseadas nos métodos de distância. Um grupo composto por espécies do Centro e Oeste da Península Ibérica foi recuperado em alguns casos, mas nunca com altos valores de suporte. Um grupo do Sudeste da Península Ibérica, que inclui P. hispanica "Galera" e duas linhagens dentro de P. hispanica "Albacete/Murcia" também foi sempre recuperado com elevado suporte. A diferença mais evidente entre as estimativas filogenéticas apresentadas neste trabalho e as que utilizam o mtDNA é uma troca nas espécies irmãs entre P. liolepis, P. hispanica "Galera", P. hispanica "Albacete/Murcia" e P. hispanica sensu stricto. Os dados nucleares colocam P. liolepis como estreitamente relacionada ou irmã de P. hispanica sensu stricto (mais P. atrata) e P. hispanica "Albacete/Murcia" como irmã de P. hispanica "Galera", apesar do facto de os dois pares serem parentes muito distantes relativamente ao mtDNA. Várias hipóteses são apresentadas para explicar esse padrão.

No geral, os resultados do NJst, MP-EST e * BEAST mostram muitas semelhanças entre si. Os dois primeiros métodos são muito rápidos e perfeitamente utilizáveis em escalas de centenas de loci mas o *BEAST é problemático para conjuntos de dados com muitos genes e precisa de muito mais tempo para atingir a convergência. O Guenomu leva apenas algumas horas usando recursos computacionais padrão, mas claramente não teve um bom desempenho no nosso conjunto de dados, fato para o qual ainda não existe uma explicação clara.

As informações obtidas nesta tese fornecem novas evidências sobre a dinâmica de especiação deste grupo e também hipóteses mais fortes sobre as suas relações evolutivas. Este trabalho permitiu também fazer comparações entre o desempenho de diferentes métodos em relação a dados reais, provenientes de cenários complexos de especiação recente com divergência muito baixa, extensa persistência de polimorfismo ancestral e partilha de haplótipos.

Palavras-chave: *Podarcis*, árvores de genes, ILS, fluxo génico, árvore das espécies, filogenia.

Abstract

Recent advances in the field of phylogenetics allow for a comprehensive inference of evolutionary processes from multilocus data. To infer evolutionary relationships between species, especially when they have diverged recently, or very rapidly, offers significant challenges. This is the case of the Iberian and North African *Podarcis* wall lizards, whose taxonomy has remained controversial even despite several studies based on mitochondrial DNA (mtDNA), allozymes and morphology.

We here used 30 nuclear loci for 170 individuals representative of all currently known morphotypes and mtDNA lineages in order to evaluate the levels of genetic polymorphism and to infer the species-tree of the Iberian and North African *Podarci*s wall lizards. For this, we used a variety of species-tree methods comprising summary statistics distance (NJst) and maximum pseudolikelihood (MP-EST) ones, Bayesian supertree methods (Guenomu) and full probabilistic Bayesian co-estimation ones (*BEAST). All these, take into account incomplete lineage sorting, the process that is thought to be the main cause of incongruence between gene-trees and species-trees in the case of *Podarcis*.

The nuclear sequences show high levels of haplotype sharing at practically all loci between the previously defined mtDNA lineages. According to these loci and a Bayesian clustering approach based on maximizing Hardy-Weinberg equilibrium, the 17 mtDNA lineages were subdivided in 24 genetically distinct groups. Moreover, also gene flow was found. In some cases gene flow had already been detected between those forms/species (*P. bocagei*/*P. guadarramae lusitanica)* but in others it was observed for the first time (*P. vaucheri* "Spain"/*P. carbonelli; P. vaucheri* "Spain"/*P. virescens; P. hispanica* "Galera"/*P. hispanica* "Albacete/Murcia" and *P. liolepis*/*P. hispanica* sensu stricto/*P. atrata)*. Some situations of clear cytonuclear discordance were revealed. In order to remove clear evidences of recent gene flow, individuals with a proportion of less of 95% of its genome assigned to a single group were excluded for phylogenetic inference.

The species-trees obtained corroborated some of the relationships as inferred by mtDNA but also revealed some completely different ones. The Eastern Iberian clade of *P. hispanica* sensu stricto, *P. liolepis* and *P. atrata* is always recovered across methods, with high support and with only small variances in the relationships between lineages. Also the North African *P. hispanica* clade proposed by mitochondrial DNA is confirmed by nuclear data, while *P. vaucheri* tends to be a clade by distance-based methods. The Western-central Iberian group sometimes formed a clade but never well supported. Southeastern Iberian clade of *P. hispanica* "Galera" and the two *P. hispanica* "Albacete/Murcia" is also always recovered and very well supported. The most

evident difference between the present and the ones using mtDNA is a swapping in sister taxa relationships among *P. liolepis*, *P. hispanica* "Galera", *P. hispanica* "Albacete/Murcia" and *P. hispanica* sensu stricto. The current nuclear data places *P. liolepis* as closely related or sister of *P. hispanica* sensu stricto (plus *P. atrata*) and *P. hispanica* "Albacete/Murcia" as sister of *P. hispanica* "Galera", despite the fact that the two pairs are very distantly related concerning mtDNA. Several hypotheses are invoked to explain this pattern.

Overall, the results from NJst, MP-EST and *BEAST show many similarities between them. The first two methods are very fast and perfectly usable at scales of hundreds of loci but *BEAST is problematic for datasets with many loci and needs a much larger amount of time to achieve convergence. Guenomu takes only a few hours using standard computational resources, but clearly did not performed well in our dataset - the explanation for which is still unclear.

The information obtained in this thesis provides new evidences about the speciation dynamics of this group and also stronger hypotheses about their evolutionary relationships. It also allowed us comparisons between performances of different methods using real datasets derived from complex scenarios of incipient speciation with very shallow divergence, extensive incomplete lineage sorting and haplotype sharing.

Keywords: *Podarcis*, gene-trees, ILS, gene flow, specie-tree, phylogeny.

Contents

List of Figures

List of Tables

List of Abbreviations

ACM4 - Acetylcholinergic Receptor M4

AICc - Akaike Information Criterion with correction

BCA - Bayesian Concordance Analysis

BI - Bayesian inference

C-mos - Oocyte maturation factor Mos

ESS - Effective Sample Sizes

GDL - Gene Duplication and Loss

GT - Gene-tree

HGT - Horizontal Gene Transfer

ILS - Incomplete Lineage Sorting

IUCN - International Union for Conservation of Nature

MC1R - Melanocortin Receptor 1

MCMC - Markov Chain Monte Carlo

ML - Maximum Likelihood

MtDNA - Mitochondrial DNA

Mya - Million Years Ago

NFYCint16 - Nuclear Transcription Factor Y, intron 16

Nst - number of Substitution rate

PKM2int5- Muscle Pyruvate Kinase 2, intron 5

RAG1 - Recombination Activating Protein 1

RAG2 - Recombination Activating Protein 2

SIC - Simple Indel Coding

SNPs - Single Nucleotide Polymorphisms

ST - Species-tree

β-fibint7 - Beta-fibrinogen, intron 7

Chapter 1

Introduction

Species definition is an important task in biology, but also one that already lead to more than half century of controversy. Determining what is a species has been difficult due the impossibility in observing some processes directly, and there is still no consensus on a single (and unifying) definition of species. The study of species-complexes, groups of closely related incipient species, such as the wall lizards of the genus *Podarcis*, is one of the best windows into the complexity of processes of diversification, and into the difficulty of defining and inferring the history of species differentiation.

1.1 The Iberian and North Africa clade of *Podarcis*

Podarcis (Lacertidae, Squamata) are diurnal lizards with high morphological and ecological similarity between species. They are habitat generalists and use rocks, trunks, vegetation or bare ground for thermoregulation, foraging and shelter. Ecological modeling revealed that temperature is a key factor for some species (Carretero *et al.*, 2006).

These lizards evolved and diversified in the Mediterranean basin and they are widely distributed in Europe and North Africa (Arnold and Ovenden, 2002). Currently, this genus comprises 23 species according to international databases (Uetz and Hosek, 2014). The object of this thesis is a well-defined, monophyletic group, inside this genus: the Iberian and North Africa clade (Harris *et al.*, 2005). Seven species are currently recognized within this group: *P. bocagei, P. carbonelli, P. hispanica, P. vaucheri, P. liolepis, P. guadarramae* and *P. virescens* (Uetz and Hosek, 2014; Geniez et al., 2014).

In terms of conservation, these species are abundant and widely distributed and thus not considered threatened. The exception is *P. carbonelli* that is classified as “endangered” by the International Union for Conservation of Nature **(**IUCN) Red List of Threatened Species, due its fragmented distribution and loss of habitat (Sá-Sousa *et al.*, 2009).

1.1.1 Taxonomy and species delimitation

The taxonomy of *Podarcis* has for long been controversial. Species delimitation has been difficult and successive studies disagreed on distribution and taxonomic classifications. This is particularly true for the Iberian and North African clade, in which the debate continues even after extensive molecular, morphological, ecological and physiological studies, having been performed (reviewed in Carretero, 2008).

After decades of controversy over morphological variation and classification attempts (e.g. Pérez-Mellado and Galindo, 1986; Geniez, 2001; Sá-Sousa, 2001) the first comprehensive studies regarding genetic variation in this group were based on mitochondrial DNA (mtDNA). These surveys evidenced the existence of several groups highly differentiated and with a strong association with geography and morphology (Harris and Sá-Sousa, 2002; Harris *et al.*, 2002; Pinho *et al.*, 2006). Phylogenetic and biogeographic hypotheses were then constructed, which have been constantly changing as new lineages are discovered (Harris and Sá-Sousa, 2002; Harris *et al.*, 2002; Pinho *et al.*, 2006; Kaliontzopoulou *et al.*, 2011).

Based on mtDNA (Kaliontzopoulou *et al.*, 2011), the distribution of these species is as shown in Figure 1.1. *P. bocagei* inhabits the Iberian Northwest and *P. carbonelli* has a fragmented distribution along the Western Iberian coast, with an isolate in Southern Spain. *P. guadarramae* comprises *P. guadarramae lusitanica* from Northwest and central Iberia and *P. guadarramae guadarramae* from central Iberia (former *P. hispanica* type 1A and 1B, respectively; Geniez *et al.*, 2014). *P. hispanica* type 2 (now called *P. virescens;* Geniez *et al.*, 2014) is found in central and South-west Iberia. *P. guadarramae* and *P. virescens* are visibly complementary, with different ecological affinities (altitude and temperature) and the lack of range overlap suggesting mutual exclusion, at least in Portugal (Sá-Sousa *et al.*, 2000). *P. vaucheri* inhabits the South of the Iberian Peninsula and throughout Morocco and Algeria. *P. liolepis* inhabits the Northeast of the Iberian Peninsula and Southern France and includes the form from the Columbretes islands (Spain) which was formerly known as *P. atrata*. Despite the synonymization of these two species (Renoult *et al.,* 2010), for simplicity in this thesis we use the name *P. atrata* to designate individuals collected in the Columbretes. *P. hispanica* is now considered a paraphyletic complex of mtDNA lineages and, according to the terminology used for e.g. in Kaliontzopoulou *et al.*, 2011 basically includes all forms which have not yet been elevated to the species status or assigned to one of the other species. This includes lineages from Morocco, Algeria and Tunisia in North Africa, and also from Southeastern Spain, where three divergent mtDNA lineages can be found (named “Galera”, “Albacete/Murcia” and “sensu stricto” in Kaliontzopoulou *et al.*, 2011).

The distribution of these species is generally parapatric, with some exceptions. The pair which overlaps the most is *P. guadarramae lusitanica* and *P. bocagei* in the Northwestern Iberia (Pérez-Mellado, 1981a), but there are also cases of sympatry involving *P. carbonelli* and *P. g. lusitanica* and *P.carbonelli* and *P. g. guadarramae* in the Western-central system (Pérez-Mellado, 1981b) *P. virescens* and *P. carbonelli* in the central zone of Portugal (Sá-Sousa, 2001) and *P. hispanica* "Galera" and *P. hispanica* sensu stricto lineages (Kaliontzopoulou *et al.*, 2011). In Doñana, in Southern Spain, sympatric populations of *P. vaucheri* and *P. carbonelli* are found (Harris *et al.*, 2002).

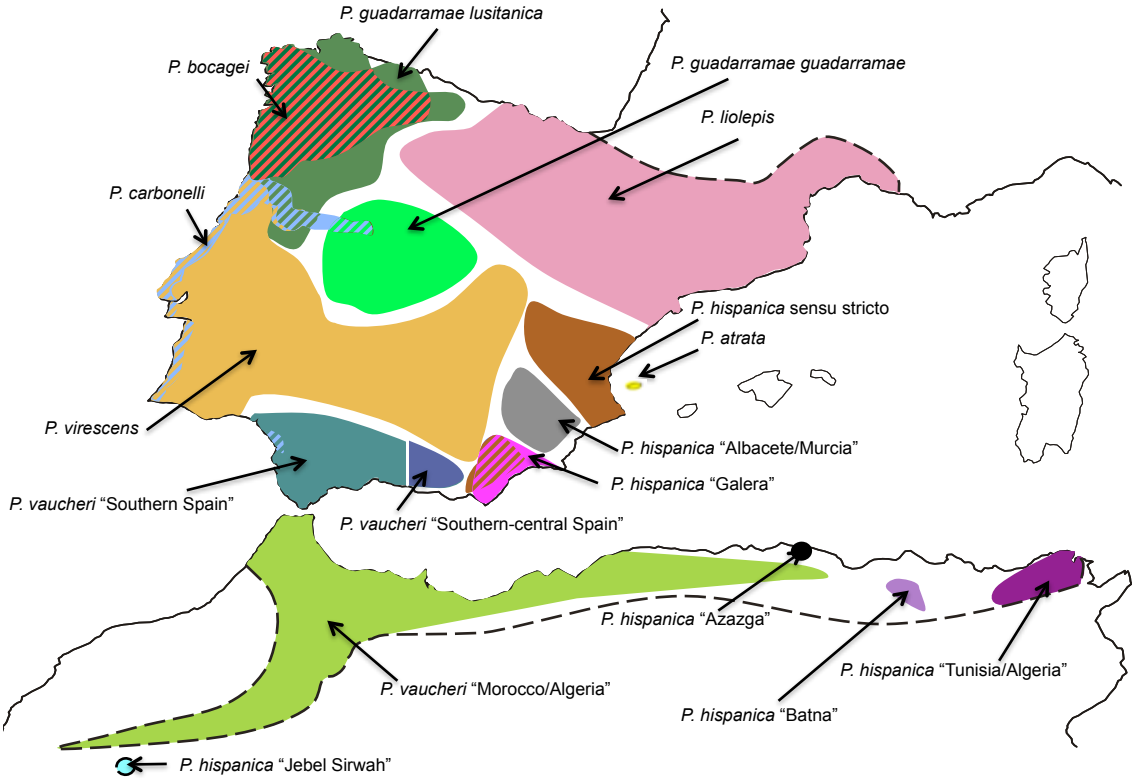


Figure 1.1. Map of the Iberian Peninsula and North African showing the estimated distribution ranges for *Podarcis* mtDNA lineages. Image made according to Pinho *et al.*, 2008 and Kaliontzopoulou *et al.,* 2011.

Nuclear markers, namely protein loci, confirm the evolutionary units inferred based on mtDNA, with some exceptions, giving support to the idea that these entities are really distinct species (Pinho *et al.*, 2007a). Genealogies of two nuclear introns showed high degrees of haplotype sharing, by opposition to the monophyly of mtDNA (Pinho *et al.*, 2008; Renoult *et al.*, 2009; Kaliontzopoulou *et al.*, 2011). This result was interpreted as mostly the outcome of incomplete lineage sorting, coupled with limited gene flow, suggesting recent speciation and incomplete barriers to gene flow. Indeed, variable degrees of introgression have been found, from rare hybridization between e.g. *P. bocagei* and *P. guadarramae lusitanica* (Pinto, 2013) to probably complete nuclear swamping resulting in a striking case of cytonuclear discordance

between forms in Southeast Iberia (Pinho *et al.*, 2007a, 2008; Renoult *et al.*, 2009). A study in an Iberian contact zone (between *P.bocagei* and *P.carbonelli*) indicated that gene flow exists, but the hybridization observed is highly localized and bimodal (Pinho *et al.*, 2009), pointing to the existence of strong reproductive barriers. One of the few comprehensive morphological surveys within the clade (Kaliontzopoulou *et al.*, 2012) showed that there is morphological differentiation consistent with genetic variation, but also that the distinctiveness is subtle and only evident on a pairwise or one-against-all comparative basis.

In short, i) the high differentiation suggested by mtDNA divergence between lineages, ii) the concordant structure inferred by nuclear markers, iii) the morphological differentiation patterns and iv) the existence of barriers to gene flow in contact zones suggest that the forms are most likely distinct species. However, the persistence of high levels of shared polymorphism and the permeability to gene exchange suggest that *Podarcis* species arose quite recently and have not concluded the process of speciation.

Remarkably, although *Podarcis* are nowadays one of the best-studied groups of the Iberian fauna, the evolutionary history of this group remains poorly understood in many aspects. This is partly due to the high rate of cryptic lineage discovery in this system (Kaliontzopoulou *et al.*, 2011), suggesting that there may still be undiscovered "species". Also, recently detected mtDNA lineages such as those of *P. hispanica* from Batna and Azazga, in Algeria (Lima *et al.*, 2009), *P. vaucheri* from South-central Spain or *P. hispanica* from the Albacete/Murcia area (Kaliontzopoulou *et al.*, 2011) or forms with sampling difficulties such as *P. liolepis* from the Columbretes (former *P. atrata*) have not been characterized from a nuclear marker perspective, implying that they may or may not correspond to distinct evolutionary entities. Finally, there is discordant evidence from different studies concerning species distinction and gene flow, particularly in the Southeastern corner of the Iberian Peninsula (Pinho *et al.*, 2007a, 2008; Renoult *et al.*, 2009; Kaliontzopoulou *et al.*, 2011).

1.1.2 Previous phylogenies and biogeographic hypotheses

Previous assessments of the phylogeny of this group were done using allozymes (Pinho *et al.*, 2003, 2007a) and mitochondrial DNA (Harris and Sá-Sousa, 2001; Harris and Sá-Sousa, 2002; Harris *et al.*, 2002; Pinho *et al.*, 2006; Kaliontzopoulou *et al.*, 2011). Allozyme markers, despite their general utility in detecting species boundaries and hybridization in this system, showed a striking lack of resolution in the estimation of relationships between species (Pinho *et al.*, 2007a). The proposed hypotheses recovered most lineages as monophyletic, but failed to suggest any relationship among species/lineages, exhibiting a star-like topology, which could be due to low resolution or be the result of rapid diversification.

On the other hand, analyses using mtDNA recovered, in general, a well supported phylogeny. The most recent study (Kaliontzopoulou *et al.*, 2011) recovered three main clades: one includes all forms from Western and Central Iberia (*P. bocagei, P. guadarramae, P. carbonelli* and *P. virescens*); another comprises the Southeastern Iberian and North African lineages (*P. hispanica* sensu stricto, *P. hispanica* "Albacete/Murcia", *P. vaucheri* with three divergent lineages - one found in North Africa and two others in Southern Spain - and the *P. hispanica* forms from Tunisia, Algeria and Morocco); and a third clade, sister to the other two, that corresponds to *P. liolepis* and *P.hispanica* "Galera", from Southern Spain (Figure 2).

The status of *P. atrata*, endemic to the Columbretes archipelago, is considered doubtful. In a mitochondrial perspective, it is recognized as *P. liolepis* (Harris and Sá-Sousa, 2002).

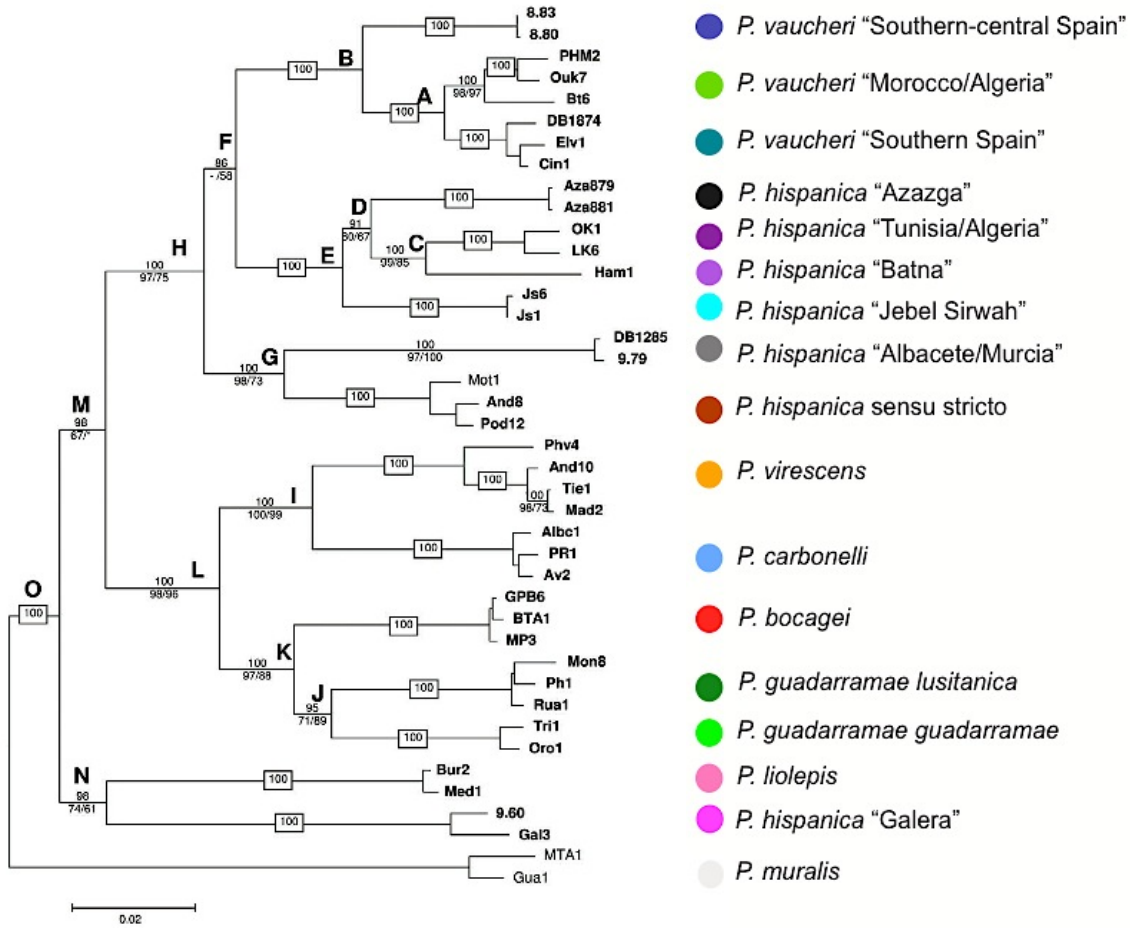


Figure 1.2. Estimate of relationships between Iberian and North African *Podarcis* based on maximum likelihood analyses of mtDNA. Above the node the Bayesian posterior probabilities and below the nodes the maximum likelihood/maximum parsimony bootstraps are given. Image modified from Kaliontzopoulou *et al.*, 2011.

The estimated divergence time between these mtDNA lineages (Kaliontzopoulou *et al.*, 2011) predates the Messinian salinity crisis (~5Mya) and it was proposed that geological events taking place at the end of the Miocene in the area now corresponding to South-eastern Iberia, the rift mountain range in Morocco, and the North of Algeria. This was followed by the abrupt

climatic modifications observed during the Miocene/Pliocene transition influenced the evolution of this group profoundly, since most cladogenic events happened around these periods. In short, it seems that the splitting events leading to major clades occurred in a period that coincides with the opening of the Betic marine corridor and the fragmentation of the Betic region, approximately 8-10 Mya (Kaliontzopoulou *et al.*, 2011). These main splits (nodes M and N in Figure 1.2.) lead to *P. liolepis* and *P. hispanica* "Galera" on one side, and to the Western-central and South-eastern Iberian and North African clades, on the other side. Right after that, the splits between *P. hispanica* sensu stricto and *P. hispanica* "Albacete/Murcia" and also between *P. vaucheri* and the North African *P. hispanica* forms seem to have happened, still considerably before the opening of the Strait of Gibraltar, during a period characterized by land connection and disconnection processes in the areas that are today the Betic and the Rif mountain range. Two later invasions of North Africa from Iberian Peninsula were proposed to explain the phylogeny of the clade of African *P. hispanica* forms and *P. vaucheri*: one by the ancestor of all North African *P. hispanica* forms (6.56 - 7.61 Mya), still predating the opening of the Strait of Gibraltar, and other at 2.3-2.69 Mya, post-dating the Messinian salinity crisis. This latter invasion could additionally have played a role in the fragmentation and isolation within other forms as the Jebel Sirwah and Azazga lineages (Kaliontzopoulou *et al.*, 2011).

Such biogeographical hypotheses, however, have been exclusively based on analyses of mitochondrial DNA variation and there are many reasons why this scenario may not correspond to the real branching pattern (Zhang and Hewitt, 1996, 2003; Ballard and Whitlock, 2004; Galtier *et al.*, 2009). In fact, introgression and mtDNA capture in face of nuclear admixture have been shown to cause cytonuclear discordance in the delimitation of evolutionary units in *Podarcis* (Renoult *et al.*, 2009). Determining the phylogeny of *Podarcis* based on the nuclear genome is thus necessary and crucial to better understand the speciation patterns and biogeography of this complex group.

1.2 Species-tree estimation – State of the art

Reconstructing the evolutionary history of living organisms is one of the ultimate and most challenging aims of Biology. The knowledge of the relationships between species is crucial in evolutionary biology and provides sources of information to help answer important questions in different areas. Phylogenies describe the origins and history of species and thus, are of major importance to understand patterns the formation of the biodiversity that we observe and to predict and manage species´ future. Beyond the obvious importance of phylogenetic reconstructions to the area of systematics, the application of phylogenetic inference overall has

steadily increased and it has become of extreme significance for example 1) to understand speciation processes, providing a trace of the process involved in the origin of the species and an opportunity to reconstructing past speciation events (Barraclough and Vogler, 2000); 2) to define management units (Douglas *et al.*, 2010); 3) in conservation, providing information of how evolutionarily diverse are ecosystems and to access conservation priorities (Douglas *et al.*, 2010). Information on the phylogenetic positions of species can tell us a history about process as extinctions (Purvis *et al.*, 2000), ecosystem functioning (Srivastava *et al.*, 2012) and even ecosystem services (Faith *et al.*, 2010) that can dictate conservation priorities (Willis and Birks, 2006).

The importance of phylogenetic reconstruction to understand the biodiversity that we observe is undeniable and recent advances in the field of phylogenetics and in the collection of molecular data allow for a comprehensive inference of evolutionary processes from multilocus data under increasingly realistic models and assumptions.

1.2.1 Gene-trees/species-tree incongruence

Phylogenetic trees are used to represent organismal evolution since the nineteenth century, dating back to Charles Darwin expressing the concept of the branching divergence of varieties and then species in a process of common descent from ancestors (Darwin, 1859). In the beginning, species were identified and described using morphological characters and these same kinds of traits were used to infer phylogenies (Wiley, 1981). However, morphological characters may suffer from convergent evolution since they often respond directly to selective pressures (Yang and Rannala, 2010). There are also limits on how this information is available for any given taxon and it can be very difficult to apply morphology to the study of some organisms, such as bacteria, and to cryptic species (Whelan, 2011). To solve these difficulties, molecular phylogenies estimated from DNA and protein sequences started being increasingly used.

Until recently, the most common solution was sequencing a gene, generally from mtDNA, or a group of genes, for a collection of species, concatenating them (in the case of several genes), and inferring single or "super" (from concatenation) gene-trees, often equating them with the tree of species. However, it has been noticed for decades that different genes can have different evolutionary histories and may not share the same genealogy as the species (Goodman *et al.*, 1979). This discrepancy between gene-tree (GT) and species-tree (ST) is thus mostly due to well known evolutionary processes such as gene duplication and loss (GDL), horizontal gene transfer (HGT) (or hybridization and introgression), and incomplete lineage sorting (ILS) (Figure 1.3.) (Maddison, 1997; Degnan and Rosenberg, 2009; Kubatko *et al.*, 2011).

Duplication and loss events, wherein a copy of a particular gene from the same genome is inserted or lost, may lead to larges differences in size and phylogenetic distribution of families of homologous genes (Szöllősi and Daubin, 2012). Gene duplication is frequent in plants, fish and insects. Like ILS, this process generates multiple gene lineages coexisting in a species lineage, but unlike ILS, gene duplication does not depend on population sizes or speciation times (Maddison, 1997).

Horizontal gene transfer is the process of transference of genetic material between species and occurs horizontally across a phylogeny. Genes will be more closely related to their ancestors than to those species in which they now reside. HGT is particularly important among prokaryotes and seems to be a crucial mechanism of rapid adaptation (Maddison, 1997). Hybridization with introgression, very common in Eukaryotes (Mallet, 2005), is almost equivalent to HGT with respect to GT discordance and ST inference (difference being in the ploidy and quantity of genes "transmitted" to the receiving species). The impacts of unaccounted gene flow in species-tree inference can be quite severe and can result in biases in species-tree topologies, overestimates of population sizes and underestimates of species divergence times (Leaché *et al.*, 2014).

Finally, incomplete lineage sorting or deep coalescence is the result of retention of polymorphism through speciation events, when recently diverged lineages have not had the time to achieve reciprocal monophyly. It depends on the effective population size and time of divergence and is thus especially problematic for closely related species or species with large population sizes and rapid speciation events, when divergence time was yet insufficient for the fixation of gene lineages by drift (Degnan and Rosenberg, 2009). When ILS is present, distinct gene-trees for different genomic regions are observed and some will differ from the species-tree.

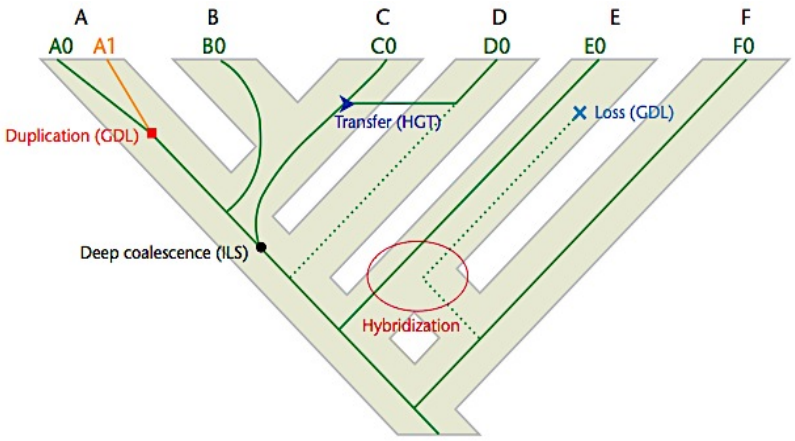


Figure 1.3. Sources of gene-tree/species-tree discordance. Different events are indicated with different colours: incomplete lineage sorting (black), gene duplication (orange) and loss (light blue), horizontal gene transfer (violet) and hybridization (red). Dashed lines represent lost lineages, either by gene loss or replacement by a foreign copy. Image from Mallo *et al.,* 2014.

Due to these phenomena, it is well known that the incorporation of multiple loci in the estimation of evolutionary parameters increases the accuracy and the resolving power of species-tree methods (Edwards *et al.*, 2007; Liu *et al.*, 2009a; Heled and Drummond, 2010). This has lead to a shift in attention away from only mtDNA or microsatellites towards other genomic markers such as anonymous loci, introns or single nucleotide polymorphisms (SNPs) (Brito and Edwards, 2009). The ability of obtaining large amounts of molecular data from multiple gene and individuals per species encouraged the development of new methods to estimate species-trees from estimated gene-trees or sequence data (Whelan, 2011).

1.2.2 Species-trees reconstruction methods

The inference of species-trees from multilocus data requires several computational methodologies adequate to our data and goals. Overall, ST inference methods can be classified into three categories: supermatrix (concatenation), supertree and co-estimation methods.

During many years concatenation was the most used approach to tree estimation and originally was developed from arguments in favor of the "total evidence" (Kluge, 1989). Here, sequences of multiple loci from the same individuals/species are linked to produce a single "supergene", whose phylogeny is equated to the ST (DeGiorgio and Degnan, 2010). The underlying assumption is that the different loci share a single evolutionary history, which is the main "sign" in the data and so, the "supergene" phylogeny is a good proxy of the "species" phylogeny. This approach appears to be accurate only when the ILS signal is low and assumes that gene-trees are congruent with the species-tree, neglecting events such as HGT or gene duplication (Bayzid and Warnow, 2013; Mirarab *et al.*, 2014). Here, it is assumed that all data conform to a single gene-tree, but the reality is most often that each gene has its own history. Further, under certain conditions, the most common gene-tree will be incongruent with the species-tree (Maddison, 1997; Degnan and Rosenberg, 2006).

In contrast, supertree methods try to find the species-tree by combining previously estimated gene phylogenies in a single estimate of the species-tree. Some supertree methods do not model the source of GT discordance, just aiming to infer a ST that minimizes GT discordance. These include 1) methods that try to minimize topological distances (between GT's and the estimated ST), such as the Robinson-Foulds (RF), (Bansal et al., 2010; Chaudhary *et al.*, 2012) and MulRF (Chaudhary *et al.*, 2013) supertrees; 2) consensus methods (reviewed in Degnan *et al.*, 2009) such as the Matrix Representation with Parsimony (MRP) approach (Bininda-Emonds, 2004); and 3) concordance methods that do a Bayesian Concordance Analysis (BCA) like those implemented in BUCKy (Larget *et al.*, 2010) or ASTRAL (Mirarab *et al.*, 2014). BUCKy uses a BCA and given a sample from posterior distribution of gene-trees it outputs a concordance

tree with the clades that have the highest amount of "genomic" support. Other supertree approaches search for the most parsimonious scenario (the one with the smallest reconciliation cost) explicitly taking into account the several possible sources of incongruence (HGT, DL or ILS). This approach is commonly known as Gene-tree Parsimony (GTP) (reviewed in Bansal and Eulenstein 2013), and underlies programs such as iGTP (Chaudhary *et al.*, 2010) and Guenomu (de Oliveira Martins *et al.*, 2014). This latter extends the supertree approach to the Bayesian framework (dealing with GT's and ST's distributions), assuming that all GT's share the same underlying ST and is able to consider reconciliation costs taking into account ILS, GDL and MulRF distances. This method is fast and agnostic regarding the sources of GT's disagreement, taking into account the uncertainty in gene-tree and species-tree estimation. Other supertree methods assume that GT's follow a multispecies coalescent model (Rannala and Yang, 2003), and use this to infer the underlying ST from GT's assuming ILS. These are usually called "summary statistics" methods, and are all very similar (reviewed in Liu *et al.*, 2009b), differing in some implementation details. Distance-based methods such as STAR, STEAC (Liu *et al.*, 2009a; b) iGLASS (Jewett and Rosenberg, 2012) and NJst (Liu and Yu, 2011) use coalescent times as estimates of speciation times. STEM (Kubatko *et al.*, 2009) implements a similar algorithm to GLASS but under a likelihood framework. STELLS (Wu, 2012) uses a maximum likelihood approach and MP-EST (Liu *et al.*, 2010) implements a "pseudo-likelihood" (where the likelihood is not calculated for the full data but some "shortcuts" are used), and use fast heuristic optimization procedures to find the most likely species-tree. Regarding the GT's used, DeGiorgio and Degnan (2014) suggest that estimated coalescent times of GT's seem to be often more recent than the true “speciation” times, and that this underestimation can lead to biases and lack of resolution of the methods based on coalescent times when using ML gene-trees. Using Bayesian gene-trees the problem seems to be less severe. The NJst method is particularly interesting among these because it is capable of working with unrooted gene-trees and missing taxa. MP-EST, NJst and STAR seem to perform similarly (Liu and Yu, 2011) and generally outperform STEAC and GLASS (Liu *et al.*, 2009b). These methods are simple and extremely fast, but require more loci to achieve accurate results than full probabilistic methods. Supertree methodologies seem to be more accurate than concatenation in the presence of high ILS (Mirarab *et al.*, 2014).

A few methods are able to co-estimate gene and species-trees. Bayesian methods as BEST (Liu, 2008) and *BEAST (Heled and Drummond, 2010) rely on the multispecies coalescent model and account for ILS, and PHYLDOG, that implements a maximum likelihood framework (Boussau *et al.*, 2013), accounts for GDL. The computation time of these full probabilistic methods can be a problem and may increase exponentially with the number of sequences; and often they have problems in reaching stationarity. *BEAST can co-estimate the

posterior distribution of gene-trees, the species-tree, and other parameters such as extant and ancestral population sizes and divergence times (whereas BEST does not estimate divergence times neither population sizes). Although computationally intensive and often very hard to achieve convergence, these methods have a better performance and generally are more accurate than the above-mentioned supertree methods (Kubatko *et al.*, 2011; Bayzid and Warnow, 2013).

Simulations indicate that *BEAST produces more accurate trees than BEST, and that BEST and BUCKy produce more accurate trees than STEM (Heled and Drummond, 2010; Leaché and Rannalla, 2010). The accuracy of STEM is only high when gene-trees are known with confidence (or estimated without error), in which case summary statistics methods will also perform well (DeGiorgio and Degnan, 2014). Currently available supertree and co-estimation methods are summarized in Table 1.1.

The methods reviewed above generally take ILS into account. Although all evolutionary processes are important, this seems to be the most preponderant at shallow deep scales (Degnan and Rosenberg, 2009). In addition, numerous studies have attempted to develop methodologies to consider both ILS and hybridization**.** Joly *et al.*, (2009) proposed an approach based on the minimum divergence time between lineage pairs, where a time statistical test will provide an assessment about whether which two (ILS or hybridization) is more likely. With hybridization the divergence may occur after speciation and with ILS the gene divergence time should predate speciation. Other approaches used consist in observing supernetworks (methods that take as input a set of partial trees and produce a set of complete splits) to distinguish between the two processes (Holland *et al.*, 2008) or to use gene-tree topology probabilities under a coalescent model (Huson *et al.*, 2005). The problem is that effective population size and generation time are often difficult to estimate and this can be problematic for these methodologies. In cases where the diversification was quick, the time between divergence and hybridization will be so small that the approaches will have not sufficient power to detect hybridization. On the other hand, the method by Joly *et al.*, (2009) seems to be biased towards detecting very recent hybridization in opposition to historical gene flow. The network approaches on the other side, seem to overestimate hybridization events (Yu *et al.*, 2011b) and return unrealistic estimates (Degnan and Rosenberg, 2006; McBreen and Lockhart, 2006).

In general, despite their limitations, all these methods provide advantages when compared to the most common practices a few years ago, based on a single gene or concatenated gene sequences, because a direct estimation of the species-tree is done. However, choosing the most appropriate specie-tree method for the data at hand is not straightforward due to different data prerequisites, model assumptions, analytical strategies and computational implementations (Mallo *et al.*, 2014).

Most of these methods have been tested on simulated datasets but applications to empirical data are still relatively scarce, particularly for closely related species. In these cases, genetic distances between species can be small (or zero) and completely overlap with intraspecific distances, making it very hard to infer GT's without error, and harder to infer non-flat ST's distributions. Taking into account multiple individuals per "species" thus becomes necessary, but this feature makes GT (and ST) inference to become also exponentially harder. This type of data is incredible challenging and the relative performance of many methods is still not well understood in these cases.

Table 1.1. The most used programs for estimating species-tree and its characteristics. BCA, Bayesian concordance analysis; MSC, Multispecies coalescent model.

	Name	Method	Input	Output
Supertree Methods	**RF**	Distance	Rooted gene-trees	Rooted supertree consistent with maximum number of splits in the input trees, without branch lengths.
	MulRF	Distance	Unrooted gene-trees	Unrooted supertree that minimizes the RF distance to the input multi-labeled trees, without branch lengths.
	MRP	Parsimony	Rooted gene-trees	Unrooted consensus supertree of all multiple trees with the same maximum parsimony score.
	BUCKy	Non-parametric BCA	Unrooted distributions	Unrooted supertree with the clades with the highest amount of genomic support, without branch lengths.
	ASTRAL	Quartet compatibility	Unrooted gene-trees	Unrooted supertree that agrees with the largest number of quartet trees.
	iGTP	Parsimony	Un/rooted gene-trees	Rooted or unrooted supertree without branch lengths or nodal support.
	Guenomu	Bayesian supertree; Parsimony	Unrooted gene-trees	Rooted species-trees with posterior distributions.
	STAR	Distance	Rooted gene-trees	Rooted species-tree without branch lengths and supported by bootstrap analysis.
	STEAC	Distance	Rooted gene-trees	Rooted species-tree without branch lengths and supported by bootstrap analysis.
	iGLASS	Distance	Rooted gene-trees	Rooted species-tree without branch lengths and supported by bootstrap analysis.
	NJst	Distance	Unrooted gene-trees	Unrooted species-tree without branch lengths and supported by bootstrap analysis.
	STEM	Maximum likelihood + Distance	Rooted gene-trees	Rooted species-tree with the highest likelihood and with branch lengths but without posterior probabilities or divergence times.
	MP-EST	Maximum pseudolikelihood	Rooted gene-trees	Rooted species-tree with branch lengths and supported by bootstrap analysis.
	STELLS	Maximum likelihood	Rooted gene-trees	Rooted species-tree with the highest likelihood and branch lengths.
Full Probabilistic Methods	**BEST**	Bayesian analysis; MSC	DNA/protein sequences	Rooted species-tree with branch lengths, posterior probabilities and divergence times.
	***BEAST**	Bayesian analysis; MSC	DNA/protein sequences	Rooted species-tree with branch lengths, posterior probabilities and divergence times.
	PHYLDOG	Likelihood	DNA/protein sequences	Rooted species-tree with branch lengths, posterior probabilities but without divergence times.

1.3 Aims and organization of the thesis

The main goal of this thesis is to explore the utility of several methods of inference of species-trees to infer the phylogeny of the Iberian and North Africa *Podarcis* wall lizards, based on 30 nuclear molecular markers. Several tools were applied to this DNA sequence dataset in order to perform alignment, alignment trimming, allele phasing, inference of gene genealogies and of "population" structure, gene flow estimates and, finally, species-tree inference.

Specifically, it was aimed at:

1) Evaluating the levels of genetic polymorphism for the above-mentioned 30 loci dataset, and its distribution within and among *Podarcis* lineages;

2) Evaluating the evolutionary distinctiveness of mtDNA lineages, particularly of those which had not been studied before, using nuclear DNA markers; and

3) Determining the phylogeny of the Iberian and North Africa *Podarcis* species complex with different methodologies able to take into account disagreement between GT's and the ST due to incomplete lineage sorting.

This thesis is organized in five chapters. The first chapter is a general introduction providing the necessary background information about the biology and controversial taxonomy of the study group and about the inference methodologies currently available to address our questions. In the next three chapters the methods used, the results and the discussion are presented. The 5th chapter corresponds to conclusions and future perspectives, with some further comments about what can be done in the future to address our main questions, what can be improved on current ongoing work, and new questions that have arisen. To finish, the literature references used along this work and Appendix data are presented.

Chapter 2

Material and Methods

2.1 Sample collection, DNA extraction and amplification

The work leading to this report consisted essentially in alignment and statistical analysis of previously produced DNA sequences. Field and laboratory work were therefore not part of this work. DNA extraction, PCR conditions, amplification and sequencing were essentially carried out using the conditions described in Pinho *et al*., (2010) and Pereira *et al.,* (2013).

2.2 Taxon and gene sampling

For this project a set of 30 unlinked nuclear loci was used. This set included loci that were previously sequenced in *Podarcis*, such as β-*fibint7* (Pinho *et al.*, 2008) or the set of markers from Pinho *et al.* 2010; exons widely used in other Squamata, such as *ACM4* (Gamble *et al.*, 2008), *PDC* (Salvi *et al.*, 2013), *C-mos* (Godinho *et al.*, 2005), *RAG1* (Pinho, *unpublished*), *RAG2* (Hoegg *et al.*, 2004) and *MC1R, NFYCint16 e PKM2int5* (Pinho *et al.*, 2010); plus a set of 21 anonymous markers developed for the genus *Podarcis* (Pereira *et al.*, 2013). From a collection of tested markers, these loci were the ones which were successfully amplified and sequenced and for which good quality sequences were obtained.

One hundred and seventy individuals, representative of all currently known species, morphotypes and mtDNA lineages of the Iberian and North African *Podarcis* clade, were Sanger sequenced. This data set included between five and 15 individuals per each of the 16 lineages described in Kaliontzopoulou *et al.*, (2011) plus individuals of *P. atrata.* As outgroups, 20 individuals belonging to five species representative of all other clades of the genus (Salvi, *personnal communication*) were used: *P. muralis* (a species that exists throughout Europe, including the Iberian Peninsula, but that belongs to a different clade than our focal species), *P. sicula*, *P. tiliguerta, P. taurica* and *P. erhardii*.

Sample codes, localities and the mtDNA correspondence for each individual are presented in Table A2 in the Appendix. The geographical origin of each sample is represented in Figure 2.1.

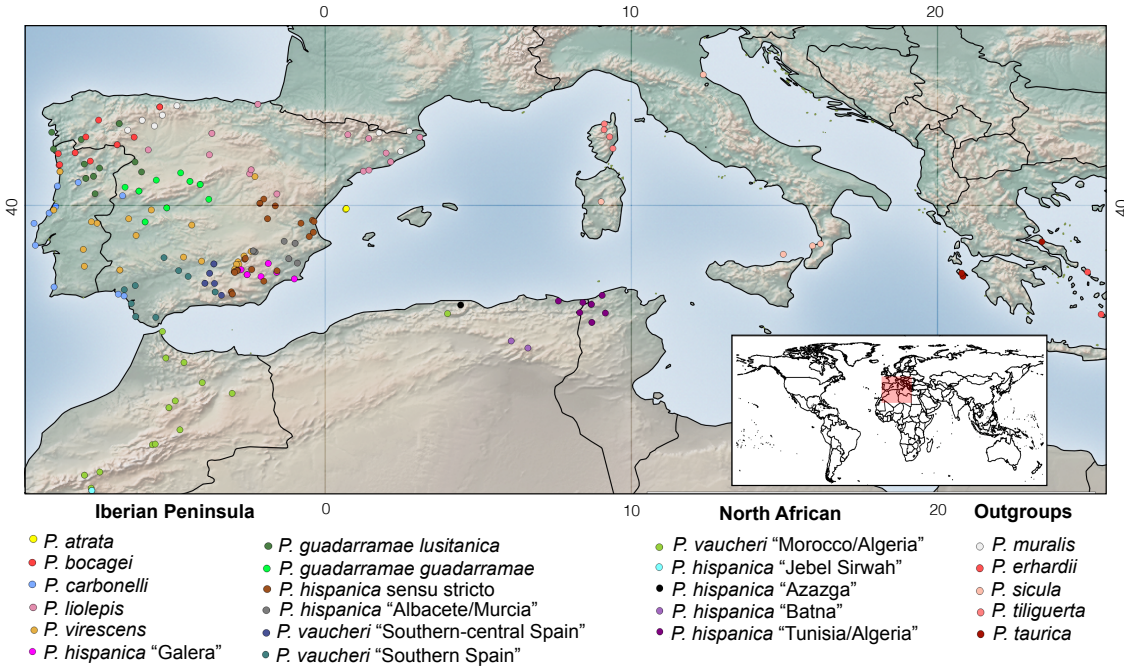


Figure 2.1. Map of the Iberian Peninsula and North African showing the geographical origin of each sample used in this study and its respective mtDNA lineage/species.

2.3 Dataset assembly

Sequences were examined and corrected by eye in Sequencher v.4.1.4. (Gene Codes Corporation). Several sequences were heterozygous for insertion/deletion polymorphisms, and the method outlined by Flot *et al.*, (2006) was used to resolve them. For the majority of the loci, as they contained a considerable amount of indels, alignment was not trivial. We experimented a few alignment algorithms, such as the E/L/G-INS-I implemented on MAFFT v.7.122 (Katoh and Standley, 2013), as well as the automated method implemented in PRANK v.140110, that takes into account the evolutionary distances between sequences while also recognizing insertions and deletions as distinct evolutionary events (Löytynoja and Goldman, 2008). This method seems to outperform the other methods and thus was used for downstream analyses. A few final adjustments, when considered necessary, were made by hand.

Because many of the alignments had large regions with indels, plus some highly variable regions, and this can be problematic for phylogenetic analyses, trimAL v.1.4 (Capella-Gutiérrez

et al., 2009) was used to remove large indels and poorly aligned regions. The heuristic "*automated1"* method was used to automatically decide the best method to trim each specific alignment, between "*gappyout*", "*strict*" and "*strictplus*", depending on the number of sequences, the average identity score among sequences and the average identity score for each most similar pair of sequences. In short, for each column of the alignment a gap-score is calculated and columns are sorted according to this score, producing a plot of gap-score thresholds versus percentage of the alignment below that threshold. The slope of this curve is then used to decide the optimal cut-off point for the *"gappyout"* option. The *"strict"* option removes the columns that would be deleted with *"gappyout"* plus blocks with at least five consecutive columns below a certain similarity cut-off. *"Strictplus"* is similar but selects automatically the block size to be eliminated, defined as 1% of the alignment size between a minimum size of 3 and a maximum size of 12. The main reason to use this approach instead of simply removing all indels was the fact that indels can be useful phylogenetic information (Freudenstein and Mark, 2001; Simmons *et al.*, 2001; Young and John, 2003), and accountable for, especially in distance-related methods.

The Bayesian algorithm implemented in the program PHASE v.2.1.1 (Stephens *et al.*, 2001) was used to recover gametic phases, often assisted by the known haplotype phases determined using the Flot *et al.*, (2006) method. The input files were prepared using DNAsp v.5.0 (Librado and Rozas, 2009) with minor modifications by hand. For each locus, the ingroup was analyzed separately from the outgroup species. Analyses were also performed independently for each of two well sampled outgroups (*P. muralis* and *P. tiliguerta*). Each dataset was analyzed using the general model for recombination rate (-MR) (Li and Stephens, 2003) with 1000 steps for burn-in, one of thinning interval and 1000 main iterations. Each analysis was repeated five times with different random seeds and the consistency of the results was verified. Incongruences between runs were not accepted. The gametic phase of alleles for polymorphic sites was considered for base probabilities ≥ 0.75. This threshold limit was chosen based on the distribution of the probabilities of the inferred alleles for different runs for all loci; was chose a value that allowed us confidence in the results, but minimized the number of positions being discarded. Several in-house perl or python scripts were used to process these data. They were used to produce the file describing the known phases, to summarize the results of different runs and then to replace the ambiguities by the inferred alleles. For *P. erhardii, P. sicula* and *P. taurica*, with fewer sequences, haplotypes were resolved by hand or left unresolved.

BioEdit v.7.2.5 (Hall, 1999) was used to edit alignments when needed. The final number of sequences used for each species/loci is given in Table 2.1.

Table 2.1. Number of sequences for each species and loci. PA, *P. atrata*; PB, *P. bocagei*; PC, *P.carbonelli*, PGL, *P. guadarramae lusitanica*; PGG, *P. guadarramae guadarramae*; PV, *P. virescens*; PHAM, *P. hispanica* "Albacete/Murcia"; PHAZA, *P. hispanica* "Azazga"; PHBAT, *P. hispanica* "Batna", PHGAL, *P. hispanica* "Galera" ; PHJS, *P. hispanica* "Jebel Sirwah", PHSS, *P. hispanica* sensu stricto, PHTA, *P. hispanica* "Tunisia/Algeria", PL, *P. liolepis*, PVMA, *P. vaucheri* "Morocco/Algeria", PVSCS, *P. vaucheri* "Southern-central Spain"; PVSS, *P. vaucheri* "Southern Spain"; PE, *P. erhardii*; PS, *P. sicula*; PTA, *P. taurica*,; and PM, *P. muralis*, PT, *P. tiliguerta*.

Loci	PA	PB	PC	PH1A	PH1B	PH2	PHAM	PHAZA	PHBAT	PHGAL	PHJS	PHSSN	PHSSS	PHTA	PL	PVMA	PVSCS	PVSS	PE	PS	PTA	PM	PT	sum
ACM4	4	2	16	18	18	30	18	4	6	18	10	14	18	20	18	20	16	18	4	10	6	12	8	308
β-fibint7	8	26	22	30	20	40	18	4	6	40	28	22	28	26	34	30	16	24	6	4	4	36	0	472
C-mos	10	20	18	14	16	30	18	4	8	20	16	24	10	20	22	24	20	20	4	4	4	16	10	352
MC1R	8	20	20	18	18	28	18	4	8	18	16	24	10	18	24	20	20	18	4	10	4	16	10	354
NFYCint16	8	18	12	14	18	24	12	4	8	12	16	12	20	20	16	18	10	16	4	4	2	14	10	292
PDC	10	14	20	18	16	28	20	4	6	18	14	24	10	18	22	22	20	20	4	4	4	16	10	342
PKM2int5	10	18	22	18	20	30	20	4	8	20	16	16	18	12	20	22	16	22	4	4	4	16	10	350
RAG1	10	20	18	18	14	30	20	4	6	20	16	22	10	18	22	24	20	20	4	4	4	16	8	348
RAG2	8	20	14	18	14	28	20	4	8	18	16	24	10	20	20	20	20	20	4	4	4	16	4	334
Pod6b	10	14	10	12	14	16	10	4	6	16	14	12	18	18	22	22	6	12	2	2	0	16	0	256
Pod7b	10	18	22	16	16	30	20	4	8	20	14	14	20	20	20	24	2	6	4	4	4	16	8	320
Pod11	10	20	16	18	18	30	20	4	8	18	16	14	20	20	22	18	14	18	4	4	4	14	10	340
Pod12b	10	20	20	18	14	30	20	4	6	18	12	12	16	14	22	22	14	24	0	0	0	0	6	302
Pod13	6	14	14	16	14	26	14	4	8	16	12	16	20	16	22	24	10	22	4	4	4	8	2	296
Pod14	6	20	20	18	14	30	14	4	8	18	12	8	18	16	22	22	16	24	0	0	0	16	0	306
Pod14b	10	20	22	16	20	20	12	4	8	12	16	14	16	18	20	20	10	18	4	4	4	14	10	312
Pod15	6	18	14	18	18	28	20	4	8	18	10	12	20	20	16	26	14	20	4	4	4	16	6	324
Pod15b	10	20	18	18	16	28	20	4	8	18	16	16	20	20	20	20	16	22	4	10	4	16	4	348
Pod16	8	4	16	18	18	28	8	4	8	18	10	16	20	18	22	24	10	14	4	4	0	14	10	296
Pod17	10	20	20	16	10	26	18	4	8	18	16	10	16	16	20	22	14	22	2	4	4	16	8	320
Pod20	10	18	22	18	20	30	20	4	8	20	12	14	20	20	20	22	16	24	4	4	4	16	6	352
Pod21	10	20	20	16	14	24	18	4	8	18	14	12	20	18	20	24	16	22	4	4	4	16	4	330
Pod25	8	8	18	12	16	24	16	4	8	14	6	14	14	12	16	12	10	18	4	4	0	10	6	254
Pod31	10	14	10	14	16	30	18	4	6	18	10	14	20	20	20	22	8	20	4	4	4	2	8	296
Pod33	10	18	16	16	14	26	20	4	8	20	16	14	18	20	16	24	8	24	4	4	0	16	6	322
Pod38	10	16	16	14	14	30	20	4	0	16	0	16	20	0	22	24	14	14	0	0	0	16	0	266
Pod43	8	20	16	14	10	20	16	4	8	14	12	16	18	16	18	18	14	18	4	4	2	12	4	286
Pod55	10	20	20	20	16	30	20	4	8	20	14	14	20	18	22	22	16	22	4	4	0	16	10	350
Pod69	10	20	22	18	20	30	12	4	8	16	14	14	20	18	18	24	16	22	2	2	0	16	8	334
Pod72	6	20	10	18	14	28	16	2	8	20	2	10	20	10	12	22	16	24	4	2	0	2	4	270

2.4 Haplotype networks

Median-joining networks (Bandelt *et al.*, 1999) with MP posterior optimization (Polzin and Daneshmand, 2003) were constructed using Network v.4.612. To build these networks, all columns with indels were removed because each gap-character will count as one different mutation (5th state), thus increasing genetic distance between sequences when long indels occur.

In parallel, we also codified all the indels with the simple indel coding (SIC) method (Simmons and Ochoterena, 2000) implemented in SeqState v.1.4.1 (Müller, 2005). SIC codes indels as separate characters taking into account the start/end of each indel and creates a

character matrix that in added to the alignment. The problem with this approach is that for indels completely comprised within other indels is impossible to determine their state.

2.5 Determining the units of analyses

Although the mtDNA assignment of all samples is known a priori, this shouldn´t be used to delimit units given the possibility of mtDNA introgression among taxa. Consequently, we used STRUCTURE v.2.3.4 (Pritchard *et al.*, 2000) to define units based on the 30 loci multilocus genotype of each individual. For these analyses, we considered each haplotype at each locus to be an independent allele, regardless of the genetic distance between haplotypes. A series of Python scripts were written to convert DNAsp haplotype distribution data files into the STRUCTURE input. *P. muralis*, one of the outgroups, also was included in this analysis since its geographical distribution encompasses the North of the Iberian Peninsula and there are some evidences of gene flow with species from our focal group (namely *P. liolepis*) (Pinho *et al.*, 2008). Individuals with more than 60% of missing data were excluded. STRUCTURE implements a Bayesian model-based clustering algorithm to find clusters of individuals that minimize Hardy-Weinberg and linkage disequilibria without any a priori information about each individual's origin. Given these underlying assumptions, STRUCTURE does not necessarily imply phylogenetic proximity among individuals recovered as belonging to the same cluster. Thus, including the most likely number of groups (K) as estimated by STRUCTURE, from the full data, as units of analyses, might introduce biases in estimated phylogenetic relationships. To minimize these biases, we opted by considering the highest possible subdivision in our sample supported by the data, whether the chosen groups correspond to species or phylogroups within them (the terms "units", "lineages", and "species" are herein used, sometimes interchangeably, and not necessarily with any taxonomic-ranking implication). In order to do so, we applied an iterative procedure to uncover structure in our data set by performing multiple separate analyses, first on the total data set and subsequently on each cluster recovered. Each analysis consisted of three steps: 1) a STRUCTURE run; 2) the choice of the appropriate number of populations present in each data set (K); 3) removal of admixed individuals, if any. All STRUCTURE runs were performed under an admixture model, for 200000 steps after 20000 steps discarded as burn-in, with 5 replicates for every possible K. K was allowed to vary from 1 to a variable number, depending on the dataset and chosen by inspecting trends in the outputs and in the likelihood values. The choice of the value of K more adequate to describe variation in each data set (X) was based on the run with the highest Ln Pr

(X|K) (following the recommendation by the developers of the program; see section 5 of STRUCTURE manual). In order to remove the influence of admixture in our data set, which would violate the assumptions of most phylogenetic methods, we chose to eliminate all individuals presenting less than 95% assignment to a single cluster. This value was chosen after inspection of the overall assignment proportions for the majority of individuals included in the analysis, typically around 99% (see Results). A threshold of 95% can be seen as perhaps too high; however, we assumed this value as a conservative cut-off in order to guarantee as little influence of gene flow as possible, even if pure individuals were inadvertently left out. This procedure was applied iteratively for each cluster defined until one of the three following possibilities were verified: a) the group contained only one individual; b) K=1 was the best K; c) (for K>1) the most part of the individuals appeared with less than 90% of the genome assigned to any of the groups (implying high gene flow or virtually no differentiation among those groups).

2.6 Conforming data to the phylogenetic methods´ assumptions

Most of the methods used in this thesis make three important assumptions: 1) patterns of allele sharing are a result of ILS and not gene flow; 2) there is no recombination within loci; 3) there is free recombination among loci or knowledge about patterns of linkage among loci.

1) Gene flow

As explained in the previous section above, we removed from the data set all the individuals presenting evidence of admixture between groups. Because this approach may fail to detected historical gene flow, was applied the coalescent model of divergence with gene flow, IMa2 (Hey and Nielsen, 2007) in a similar way as Pinho *et al.*, (2008) but some preliminary results showed gene flow among most of the species and because this was a clear violation of the IM model assumptions (Strasburg and Rieseberg, 2009) we could not have confidence in our results and thus, this strategy was abandoned.

2) Recombination

We used RDP3.44 (Martin *et al.*, 2010) to assess whether sequence data were affected by recombination. For this, the possibility of recombination was investigated using three methods: RDP (Martin *et al.*, 2010), GENECONV (Sawyer, 1989) and MaxChi (Smith and Smith, 1998); we used the option "*automask*" for an optimal recombination detection, setting the cut-off P-

value to 0.001. Following author recommendations (Martin, *personnal communication*), if recombination was not inferred for the three methods simultaneously, we assumed recombination-free alignments.

3) Patterns of linkage among loci

We performed an exact test for genotypic disequilibrium using the program Genepop v.4.1.4 (Rousset, 2008) in order to evaluate whether any pair of loci were in physical linkage, thus sharing a common evolutionary tree. This analysis was performed for the five groups, among those recovered by STRUCTURE (see section 2.5), which had a sample size of at least 10 individuals (*P. bocagei*, *P. carbonelli*, *P. virescens, P. vaucheri* "Spain", *P. vaucheri* "Morocco/Algeria 2"; see Results).

2.7 DNA sequence polymorphism

To describe levels of genetic variation, summary diversity statistics for each gene were calculated for our group of interest (that is, after removing all outgroups). This was performed in DNAsp v.5.10.1 (Librado and Rozas, 2009). As this program does not accept ambiguity codes, we re-coded all unphased positions as Ns with an in-house python script. We calculated the number of haplotypes, haplotype and nucleotide diversity, the number of total mutations, the number of segregating sites and the population mutation parameter θ (Watterson, 1975).

2.8 Gene-trees inference

Some of the species-trees inference methods used in this work take gene-trees or gene-trees distributions as input, thus creating a need for gene-tree estimation prior to analysis. This step was carried out under different methods. The best-fit model of evolution for each gene fragment was estimated with jModelTest2 (Guindon and Gascuel, 2003; Darriba *et al.*, 2012), under the Akaike Information Criterion with correction (AICc).

Gene-trees for each loci were estimated using maximum likelihood (ML) in RAxML v.8 (Stamatakis, 2014) and with MrBayes v.3.2.2 (Ronquist *et al.*, 2012) for Bayesian inference (BI). For RAxML the GTR model was always used (as it is the only implemented), with or without G/I, depending on the highest ranked GTR model on the jModeltest2 weight table. The

ML tree was chosen from 50 heuristic replicates and nodal support was assessed through 5000 bootstrap replicates. For some genes, there were some "units"/"species" without any representative (Table 3.3 in Results section). As one of the downstream ST methods does not allow missing data, we ran these adding “dummy” sequences (where all sequence are coded by N´s) so that all groups were represented at least by one sequence. Also, ML searches were carried out in alignments of all the sequences (herein “full” dataset) as well as in alignments reduced to “haplotypes per unit”, i.e., where each haplotype is only represented once in each unit (herein “haplotypes” dataset). In BI, the best-fit model as obtained with jModeltest2 was implemented. The Markov chain Monte Carlo (MCMC) was run for 10 millions generations, sampling each 1000. Two runs were performed and the convergence and congruence were evaluated using AWTY (Nylander *et al.*, 2008). Posterior probabilities of the same splits for the two runs were plotted to assess the congruence of the different runs, and the cumulative posterior probabilities of the most variable splits of each run were plotted across generations, to evaluate convergence and the appropriate burn-in. In a few cases runs of 50 millions of generations were needed to reach convergence.

2.9 Species-tree inference

Species-tree inference was performed under multiple methods that take ILS into account: NJst (Liu and Yu, 2011), MP-EST (Liu *et al.*, 2010), Guenomu (de Oliveira Martins *et al.*, 2014) and *BEAST (Heled and Drummond, 2010).

For NJst (Liu and Yu, 2011) we used the unrooted ML trees and both the “haplotypes” and the “full” ML datasets. One thousand bootstrap replicates were calculated under three different ways: 1) by sampling from the 30 ML trees, with replacement (herein called “locus bootstrap”; 2) by sampling one tree per locus from the ML bootstrap replicates (herein called “sequence bootstrap” and; 3) using a 2-stage bootstrapping technique (Seo, 2008) that combines the two previous, i.e., sampling 30 trees from the 30 ML bootstrap distributions, with replacement (called “multilocus bootstrap”). Extended consensus-trees of the above bootstraps were also calculated.

For MP-EST (Liu *et al.*, 2010) the “haplotypes” ML trees were used. MP-EST requires rooted trees, and further requires that all gene-trees be rooted with tips belonging to the same “species”. As all our “outgroups” had some proportion of “missing data”(i.e., for which there was no sequence for some loci), we rooted all gene-trees with a representative of one of the species with less “missing data”, *P. erhardii*. As such, three of the loci (for which there was no

representative of *P. erhardii*) had to be excluded (*Pod38*, *Pod14*, and *Pod12b*). One hundred independent replicates were run and the one with the highest "pseudolikelihood" score chosen as the MP-EST tree. A consensus of the 100 replicates was used to represent "branch-support".

Guenomu (de Oliveira Martins *et al.*, 2014) takes as input a bayesian distribution of gene-trees or set of bipartitions. Thus we used the "*.trprobs*" file obtained from each locus majority-rule consensus from MrBayes, which contains all the trees that were found during the MCMC search sorted by their posterior probability (which Guenomu takes into account). Guenomu does not take into account branch lengths. Therefore, an extra run of Guenomu was also performed, where a simulator of tree inference uncertainty (usually higher in shorter branches), "*addTreeNoise*" (available in the main program directory) was used. In short, for each tree, branch lengths are used as proxies of estimation error and rearrangements of the topology are performed in the shorter branches, according to a previously defined fraction of intended "*wrong trees*" (it was used 0.7) and an error probability per branch (scaled by the shortest branch; we used 2). For each input tree we created 50 output trees. Then the final outputs of this step were reduced to 9000 trees per gene (for computational reasons only) by re-sampling the trees ordered by their individual probability. The logics behind this step is that, if there is a lot of variation in tree topologies, some due to inference error, and that inference error is related to branch lengths, these rearrangements will increase the "noise", i.e., add variation in "wrong" branches, while the ones more accurately inferred will increase in proportion. This should translate into an "increased" signal for this supermatrix method. Guenomu was then run for 1000 burn-in iterations followed by 100000 generations for the posterior sampling, saving 1000 samples of trees. An extended consensus was built using BayesTrees v.1.3 (Meade, 2011).

Finally, we used *BEAST (Heled and Drummond, 2010), the multispecies coalescent approach implemented in BEAST 2 (Bouckaert *et al.*, 2014). As this method is extremely computationally intensive, we used the "haplotypes" alignments. For all loci, each substitution model was chosen based on jModeltest results (using the closest model available) as well as a strict molecular clock model for all genes. Substitution rate of the "fastest" gene (the one with highest θ_w inferred by DNAsp) was fixed to 1 and the remaining rates were co-estimated. Several runs were made, inclusively down sampling the number of loci (see Results), using several priors for tree-prior (Yule, birth-death, coalescent) as well as for population sizes (linear with constant root and constant). Runs up to one thousand millions of generations (or longer) were set up, sampling at appropriate intervals for a total of 10.000 posterior trees. Log files were inspected using Tracer v1.4 (Rambaut and Drummond, 2007) to check for appropriate effective sample sizes (ESS) and traces of parameters, as well as burn-in. TreeAnnotator was used to obtain the final consensus trees (maximum clade credibility tree) and FigTree v.1.4.2

(Rambaut, 2014) was used to visualize and edit these trees. DensiTree v2.1.11 (Bouckaert, 2010) was used for visualizing the full posterior distribution of trees.

Chapter 3

Results

3.1 Nuclear DNA genealogies

A few examples of the haplotype networks inferred for the 30 nuclear loci are shown in Figure 3.1 and the remaining networks for all loci can be found in the Appendix. These networks were built after removing all columns with indels (see Methods).

The most evident result from the nuclear DNA genealogies is the complete lack of monophyly and the high amount of shared alleles between the distinct mtDNA lineages.

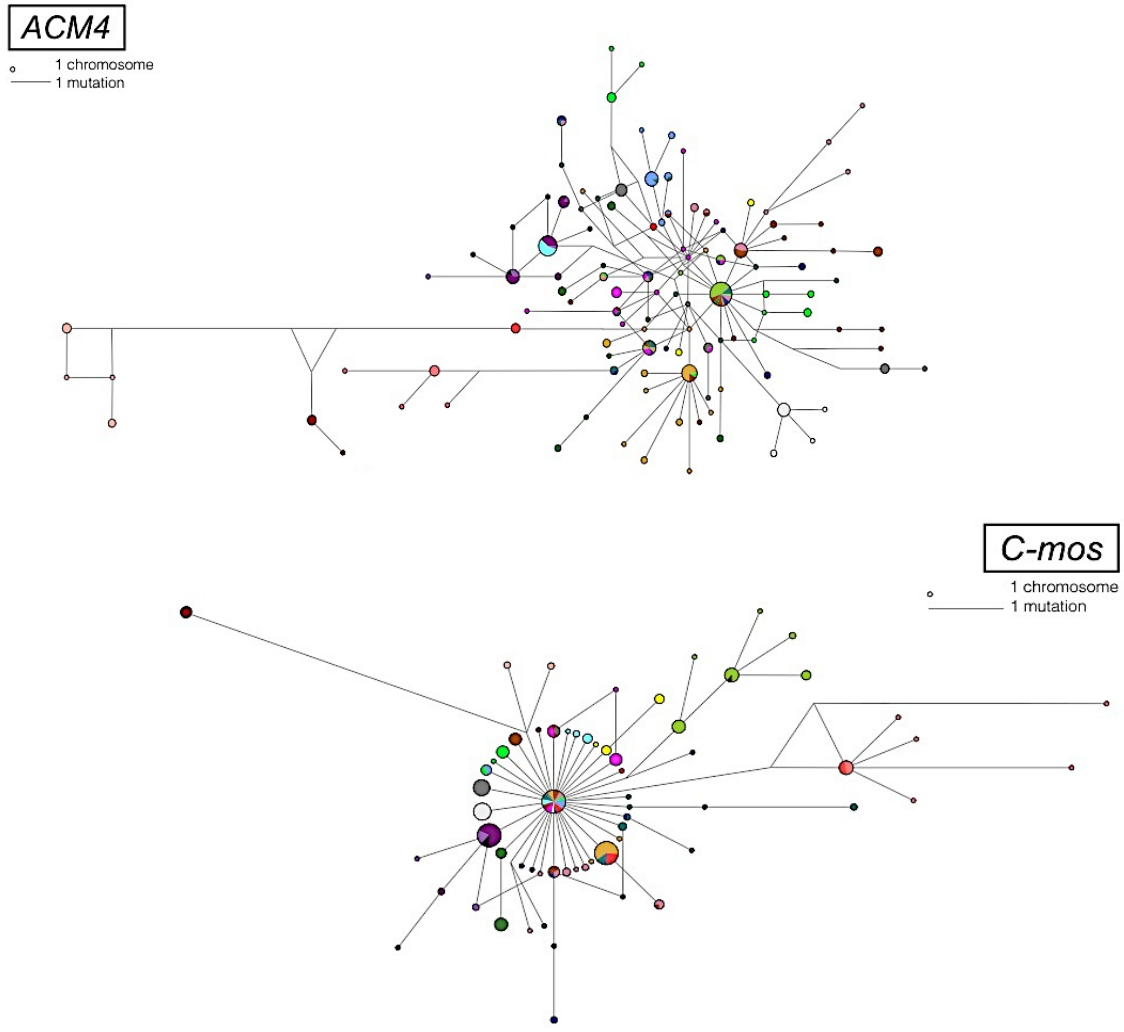


Figure 3.1. Haplotype networks for four nuclear loci analysed in this study. Circle area corresponds to haplotype frequency, and distinct colours to different mtDNA lineages/species (17) and are the same used in Figure 2.1 Branch lengths are proportional to the distance between haplotypes.

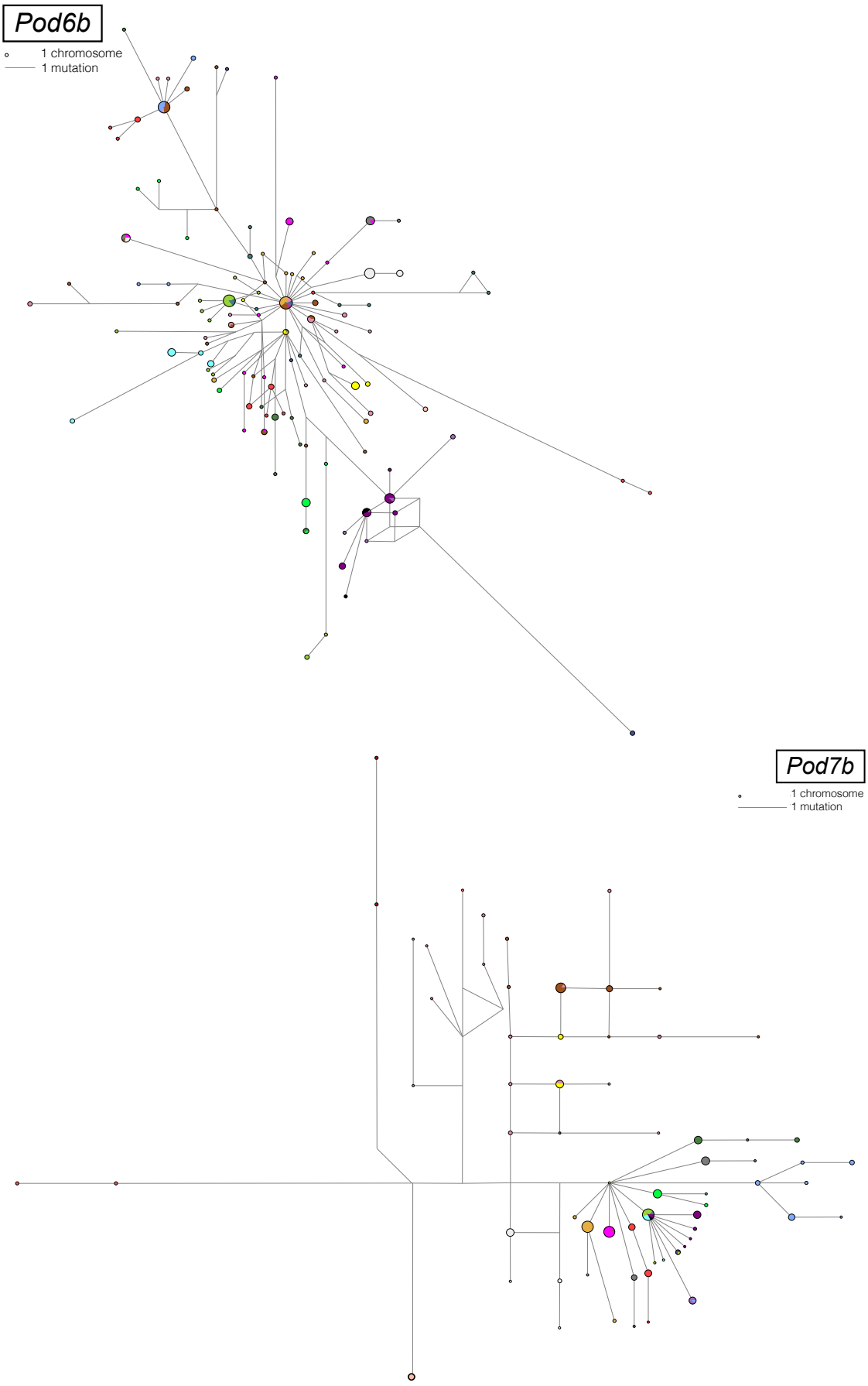


Figure 3.1. (Continuation)

3.2 Individual multilocus genotype analyses

The results from STRUCTURE show that the 17 mtDNA lineages can be organized in 24 genetically distinct "groups" and that most individuals are recovered as "pure" (with more than 95% of its "genome" assigned to a single group). Clusters obtained from successive STRUCTURE runs are shown in Figure 3.1.

The first run, including all individuals, recovered six clusters (K=6): one with *P. atrata, P. hispanica* sensu stricto and *P. liolepis;* another with *P. bocagei, P. guadarramae lusitanica* and *P. guadarramae guadarramae*; another with *P. carbonelli*, *P. virescens,* and all lineages of *P. vaucheri*; a fourth cluster included *P. hispanica* "Albacete/Murcia" and *P. hispanica* "Galera"; another group comprised all North African *P. hispanica* lineages; and lastly, *P. muralis*. In the next steps all these groups were subdivided resulting in clusters roughly comprising each species/mtDNA lineage. Some of these were further subdivided in the following runs, resulting in clusters that probably represent intraspecific groups.

The clusters that were further subdivided corresponded to *P. atrata, P. muralis, P. liolepis*, *P. hispanica* sensu stricto*, P. vaucheri* "Morocco/Algeria" and *P. hispanica* "Albacete/Murcia". In some of these cases, the subdivision probably has some geographic component (although sample size is obviously small to perform any kind of inference): for example, the five individuals of *P. atrata* were subdivided in two different clusters clearly corresponding with the two different groups of islands of the Columbretes archipelago.

Despite resulting in a higher subdivision of the dataset than that considered using mtDNA, in general there is a good agreement between nuclear and mtDNA partitions. The only major exceptions to a clear cytonuclear correspondence were the groups "*P. liolepis* 1", which included individuals carrying both the typical *P. liolepis* and the *P. hispanica* sensu stricto mtDNA lineages, and "*P. vaucheri* Spain", which included the two Iberian *P. vaucheri* lineages (which had not been recovered as sister taxa based on the mtDNA). There was also one individual carrying the *P. hispanica* sensu stricto mtDNA lineage, which was assigned to "*P. hispanica* Galera".

Twenty-two individuals in total exhibited possibly admixed genotypes and were thus excluded from further analyses. These cases of uncertainty/admixture are highlighted in Table 3.1 and Figure 3.2. Remarkably, there are more admixed individuals between *P. liolepis* and *P. hispanica* sensu stricto (three of them including also *P. atrata*) than between remaining groups. Moreover, we detected an individual reflecting admixture between *P. vaucheri* "Spain" and *P. carbonelli*, a pair between which hybridization had never been detected in previous works.

Table 3.1. Number of admixed individuals and designation of the clusters to which they were assigned. PA, *P. atrata*; PB, *P. bocagei*; PC, *P.carbonelli*; PGG, *P. guadarramae guadarramae;* PGL, P. *guadarramae lusitanica;* PHAM, *P. hispanica* "Albacete/Murcia"*;* PHAZA, *P. hispanica* "Azazga"; PHBAT, *P. hispanica* "Batna", PHGAL, *P. hispanica* "Galera; PHJS, *P. hispanica* "Jebel Sirwah", PHSS, *P. hispanica* sensu stricto, PHTA, *P. hispanica* "Tunisia/Algeria", PL, *P. liolepis,* PVMA, *P. vaucheri* "Morocco/Algeria*",* PVSCS*, P. vaucheri "*Southern-central Spain"; PVSS, *P. vaucheri "*Southern Spain"*;* PV*, P. virescens;* and PM, *P. muralis.*

Sample code	**mtDNA lineage**	**Structure**	**Region, Country**
9.32	PHSS	0,009 (PA); 0,314 (PHSS); 0,677 (PL)	Valencia, Spain
B2	PHSS	0,052 (PA); 0,172 (PHSS); 0,777 (PL)	Valencia, Spain
Val1	PHSS	0,169 (PA); 0,309 (PHSS); 0,523 (PL)	Valencia, Spain
1.15	PL	0,181 (PA); 0,151 (PHSS); 0,668 (PL)	Aragón, Spain
Cel1	PGL	0,062 (PB); 0,936 (PGL); 0,001 (PGG)	Ourense, Spain
CR1	PV	0,003 (PVMA); 0,002 (PC); 0,939 (PV); 0,055 (PVSCS/PVSS)	Huelva, Spain
MR16	PV	0,002 (PVMA); 0,004 (PC); 0,173 (PV); 0,821 (PVSCS/PVSS)	Leiria, Portugal
8.26	PVSS	0,006 (PVMA); 0,908 (PC); 0,002 (PV); 0,084 (PVSCS/PVSS)	Huelva, Spain
9.73	PHGAL	0,515 (PHGAL); 0,485 (PHAM)	Murcia, Spain
9.67	PHGAL	0,585 (PHGAL); 0,415 (PHAM)	Murcia, Spain
DB4281	PM	0,917 (PM1); 0,083 (PM2)	León, Spain
DB4294	PM	0,936 (PM1); 0,064 (PM2)	León, Spain
ALB11=9.79	PHAM	0,222 (PHAM1); 0,778 (PHAM2)	Albacete, Spain
ALB2=9.77	PHAM	0,912 (PHAM1); 0,088 (PHAM2)	Albacete, Spain
DB1878	PHAM	0,834 (PHAM1); 0,166 (PHAM2)	Albacete, Spain
7 300	PVMA	0,057 (PVMA1); 0,944 (PVMA2)	Khenifra, Morocco
DB1449	PVMA	0,920 (PVMA1); 0,080 (PVMA2)	Ceuta, Spain
DB8560	PL	0,013 (PL1); 0,811 (PL2); 0,176 (PL3)	Soria, Spain
DB1853	PHSS	0,066 (PHSS1); 0,934 (PHSS2)	Jaén, Spain
DB1748	PHSS	0,940 (PHSS1); 0,060 (PHSS2)	Jaén, Spain
Val1	PHSS	0,905 (PA/PL/PHSS); 0,037 (PHAM/PHGAL); 0,049 (PC/PV/PVMA/PVSCS/PVSS); 0,002 (PHAZA/PHBAT/PHJS/PHTA); 0,003 (PB/PGL/PGG); 0,001 (PM)	Valencia, Spain
Alb8	PC	0,003 (PA/PL/PHSS); 0,003 (PHAM/PHGAL); 0,927 (PC/PV/PVMA/PVSCS/PVSS); 0,060 (PHAZA/PHBAT/PHJS/PHTA); 0,003 (PB/PGL/PGG); 0,002 (PM)	Salamanca, Spain

The number of sequences of each species and locus kept for further analyses after STRUCTURE is shown in Table 3.2.

Table 3.2. Number of sequences for each loci and unit defined with STRUCTURE. PA, *P. atrata*; PB, *P. bocagei*; PC, *P.carbonelli*; PGG, *P. guadarramae guadarramae;* PGL, P. *guadarramae lusitanica;* PHAM, *P. hispanica* "Albacete/Murcia"*;* PHAZA, *P. hispanica* "Azazga"; PHBAT, *P. hispanica* "Batna", PHGAL, *P. hispanica* "Galera; PHJS, *P. hispanica* "Jebel Sirwah", PHSS, P. *hispanica* sensu stricto, PHTA, *P. hispanica* "Tunisia/Algeria", PL, *P. liolepis,* PVMA, *P. vaucheri* "Morocco/Algeria*",* PVSCS*, P. vaucheri "*Southern-central Spain"; PVSS, *P. vaucheri "*Southern Spain"*;* PV*, P. virescens;* PM, *P. muralis;* PE, *P. erhardii;* PS, *P. sicula;* and PTA, *P. taurica.*

Loci	PA1	PA2	PB	PC	PH1A	PH1B	PH2	PHAM1	PHAM2	PHAZA	PHBAT	PHGAL	PHJS	PHSS1	PHSS2	PHTA	PL1	PL2	PL3	PVMA1	PVMA2	PVS	PE	PS	PTA	PM1	PM2	PT	Sum
ACM4	2	2	2	14	16	18	28	8	4	4	6	18	10	4	10	20	12	8	2	2	14	32	4	10	6	8	2	8	274
β-fibint7	4	4	18	14	10	10	24	10	2	4	6	16	14	4	12	16	12	8	2	2	14	38	2	4	4	8	2	0	264
C-mos	6	4	20	14	12	16	28	8	4	4	8	20	16	4	12	20	14	8	2	2	18	38	4	4	4	8	2	10	310
MC1R	6	2	20	18	16	18	28	8	4	4	8	18	16	4	12	18	14	8	2	2	14	36	4	10	4	8	2	10	314
NFYCint16	6	2	18	12	10	18	22	6	2	4	8	12	16	4	12	20	10	6	0	2	14	24	4	4	2	8	0	10	256
PDC	6	4	14	16	16	16	26	10	4	4	6	18	14	4	12	18	14	8	2	2	18	38	4	4	4	8	2	10	302
PKM2int5	6	4	18	18	16	18	28	10	4	4	8	18	16	4	10	12	14	6	2	2	16	36	4	4	4	8	2	10	302
RAG1	6	4	20	16	16	14	28	10	4	4	6	18	16	4	12	18	14	8	2	2	18	38	4	4	4	8	2	8	308
RAG2	6	2	20	12	16	14	28	10	4	4	8	18	16	4	12	20	14	6	2	2	14	38	4	4	4	8	2	4	296
Pod6b	6	4	14	8	12	12	14	6	2	4	6	14	14	4	12	18	12	8	2	2	18	16	2	2	0	8	2	0	222
Pod7b	6	4	18	20	14	16	26	10	4	4	8	18	14	4	12	20	12	8	2	2	18	8	4	4	4	8	2	8	278
Pod11	6	4	20	16	16	16	26	10	4	4	8	16	16	4	12	20	14	8	2	2	16	34	4	4	4	8	2	10	306
Pod12b	6	4	20	20	16	14	26	10	4	4	6	18	12	4	10	14	12	8	2	2	16	36	0	0	0	0	0	6	270
Pod13	2	4	14	12	14	14	22	6	4	4	8	16	12	4	12	16	12	8	2	2	18	30	4	4	4	4	0	2	254
Pod14	4	2	20	18	16	12	26	8	2	4	8	16	12	4	10	16	14	8	2	2	18	38	0	0	0	8	2	0	270
Pod14b	6	4	20	20	14	18	18	8	2	4	8	12	16	4	10	18	12	8	2	2	16	26	4	4	4	6	2	10	278
Pod15	6	0	18	12	16	16	26	10	4	4	8	18	10	4	12	20	10	6	2	2	18	32	4	4	4	8	2	6	282
Pod15b	6	4	20	18	16	16	24	10	4	4	8	16	16	4	12	20	14	6	2	2	14	38	4	10	4	8	2	4	306
Pod16	4	4	4	14	16	18	24	4	0	4	8	16	10	4	12	18	14	8	2	2	18	24	4	4	0	8	2	10	256
Pod17	6	4	20	18	16	10	22	8	4	4	8	16	16	2	10	16	12	8	2	2	16	34	2	4	4	8	2	8	282
Pod20	6	4	18	20	16	18	26	10	4	4	8	18	12	4	12	20	14	6	2	2	18	38	4	4	4	8	2	6	308
Pod21	6	4	20	20	14	14	22	10	4	4	8	16	14	2	12	18	14	6	2	2	18	36	4	4	4	8	2	4	292
Pod25	4	4	8	16	12	14	22	10	2	4	8	10	6	2	12	12	10	6	2	0	8	28	4	4	0	4	2	6	220
Pod31	6	4	14	8	12	14	26	8	4	4	6	16	10	4	12	20	12	8	2	2	18	26	4	4	4	0	0	8	256
Pod33	6	4	18	16	14	14	22	10	4	4	8	18	16	4	10	20	14	6	2	2	18	28	4	4	0	8	2	6	282
Pod38	6	4	16	14	12	12	26	10	4	4	0	16	0	4	12	0	14	8	2	2	18	28	0	0	0	6	2	0	220
Pod43	4	4	20	16	12	10	18	4	4	4	8	14	12	4	12	16	10	8	2	2	14	32	4	4	2	6	2	4	252
Pod55	6	4	18	22	16	14	28	12	4	4	8	18	14	6	14	18	14	8	2	2	18	36	4	4	0	10	2	10	316
Pod69	6	4	18	22	16	18	26	8	2	4	8	14	14	4	12	18	14	4	2	2	18	36	2	2	0	8	2	8	292
Pod72	2	4	20	6	16	14	26	10	2	2	8	18	2	2	12	10	8	2	2	2	18	38	4	2	0	0	2	4	236

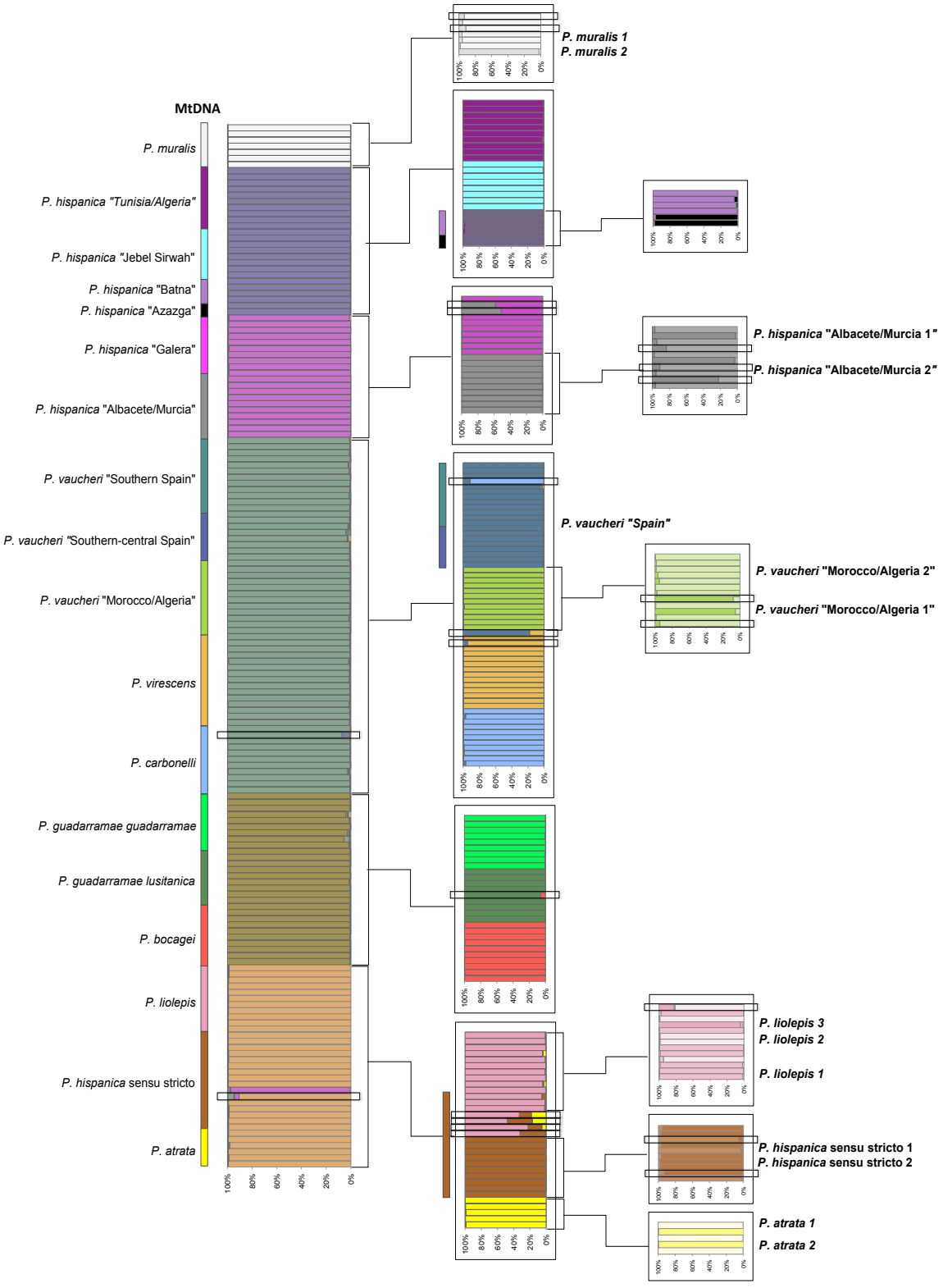


Figure 3.2. Estimated probability of ancestry of the Iberian and North African *Podarcis* species complex and of the outgroup *P.muralis,* as calculated with STRUCTURE. Each horizontal bar represents one individual and is divided into K segments shown in different colours, with sizes proportional to the portion of the genome of each individual inferred to have originated from each of the K inferred clusters. Transparent boxes highlight admixed individuals, who were excluded from further analyses. The species assignment/mtDNA lineages of the individuals in question are shown in the vertical bars on the left, with the same colours used in Figure 2.1 and 3.1. The new units resulting from subdivisions by STRUCTURE are shown on right in bold.

3.3 Nuclear loci variability

Seven out of the 30 newly sequenced loci had large indel regions, sometimes amounting to over 400bp. Initial alignment length varied between 320bp and 1392bp, and after trimming, final alignments vary between 306bp and 676bp.

Summary diversity statistics for the 30 loci and considering all individuals of the Iberian and North African *Podarcis* group are given in Table 3.3.

Table 3.3. Summary statistics and neutrally test for the 30 loci analysed in this study. Nseqs, number of sequences; NSites, total sequence lengths – () excluding sites with gaps / missing data; h, number of haplotypes; Hd, haplotype diversity; Eta, total number of mutations; S, number of segregating sites, π, nucleotide diversity, θw, Theta-Waterson.

Loci	Nseqs	NSites	Polymorphism						Indels
			h	Hd	Eta	S	π	θw	
ACM4	232	432 (385)	38	0,782	40	37	0,004670	0,01596	0
β-fibint7	244	519 (407)	102	0,981	110	101	0,011793	0,04087	11
C-mos	274	550 (537)	44	0,738	44	44	0,002672	0,01324	0
MC1R	272	694 (648)	45	0,795	38	38	0,002784	0,00949	0
NFYCint16	228	646 (405)	72	0,941	82	78	0,010893	0,03208	17
PDC	268	349 (327)	36	0,779	26	26	0,004037	0,01289	0
PKM2int5	270	459 (289)	42	0,807	41	36	0,004983	0,02018	9
RAG1	278	454 (431)	27	0,571	27	24	0,001690	0,00898	0
RAG2	268	676 (615)	48	0,896	51	51	0,003132	0,01345	0
Pod6b	208	489 (339)	53	0,845	81	73	0,007764	0,03642	11
Pod7b	248	385 (343)	39	0,931	43	41	0,011255	0,01963	4
Pod11	274	435 (264)	53	0,792	45	41	0,008175	0,02510	8
Pod12b	264	410 (241)	93	0,97	101	87	0,018868	0,05869	7
Pod13	236	360 (204)	44	0,644	45	40	0,006056	0,03247	2
Pod14	260	530 (402)	65	0,931	72	67	0,009209	0,02716	8
Pod14b	248	400 (242)	67	0,907	74	65	0,008356	0,04411	25
Pod15	254	420 (282)	48	0,895	47	43	0,006518	0,02495	3
Pod15b	274	504 (357)	78	0,891	71	66	0,006441	0,02987	15
Pod16	228	306 (192)	81	0,969	64	61	0,026945	0,05291	8
Pod17	254	336 (250)	51	0,681	55	50	0,004832	0,03272	3
Pod20	280	396 (212)	31	0,601	33	30	0,005253	0,02279	2
Pod21	266	316 (150)	55	0,873	54	44	0,013613	0,04763	7
Pod25	200	276 (241)	30	0,798	30	28	0,007636	0,01978	3
Pod31	236	507 (380)	89	0,968	78	75	0,010491	0,03268	8
Pod33	258	339 (158)	38	0,832	35	33	0,013199	0,03408	10
Pod38	212	461 (326)	17	0,566	21	20	0,004096	0,01034	0
Pod43	232	353 (320)	28	0,507	29	29	0,001952	0,01505	0
Pod55	286	421 (377)	30	0,493	33	32	0,001951	0,01362	0
Pod69	270	388 (294)	32	0,587	31	31	0,003009	0,01708	2
Pod72	224	462 (328)	67	0,936	63	59	0,015985	0,03005	7

Nucleotide diversity (π) ranged between 0,00169 (*RAG1*) and 0,018868 (*Pod12b*) while haplotype diversity (Hd) varied between 0,571 (*RAG1*) and 0,981 (*β-fibint7*). These levels of polymorphism are similar to the ones previously observed in the two nuclear introns ($\pi_{\beta\text{-}fibint7}$*=0.01269;* $\pi_{6\text{-}Pgdint7}$=0.01721) and, as expected, considerably lower than that detected in mtDNA (Pinho *et al.*, 2008). The population mutation rate (θ_w) ranged between 0,00898 (*RAG1*) and 0,05869 (*Pod12b*).

3.4 Gene-trees

Models of sequence evolution selected using the AICc in jModeltest2 for each locus are given in Table 3.4, as well as the respective models as implemented in *BEAST.

All ML and BI gene-trees were very shallow, and usually poorly resolved, mostly comprised of polytomies. Often, the ingroup was not even recovered as monophyletic.

3.5 Species-tree Inference

Species-trees inferred by the supertree summary statistics method NJst are shown in Figure 3.3 a) and b). We here present the topology of the "best" NJ species-tree (built using the ML trees for all loci) from the "haplotypes" dataset and the bootstrap NJst consensus, with multilocus bootstraps calculated following the 2-stage bootstrap procedure from Seo (2008). Remaining tree topologies (based on the dataset including all sequences ("full") and bootstrap NJst consensus) can be found on Appendix. Well-supported groups are coincident in all cases.

Tree topologies as estimated with NJst reveal overall concordance in many aspects. The ingroup (the Iberian and North African *Podarcis* excluding *P. muralis*) is in all cases monophyletic with great support, as well as *P. muralis*. A few groups are in all cases well supported: 1) the eastern Iberian group of *P. atrata*, *P. liolepis* and *P. hispanica* sensu stricto; 2) a clade with the three clusters of *P. vaucheri*; 3) all *P. hispanica* forms from North Africa; 4) the southeastern Iberian *P. hispanica* "Albacete/Murcia" and *P. hispanica* "Galera" and 5) *P. bocagei* plus *P. guadarramae lusitanica*. The sister-relationship between *P. vaucheri* and the African forms of *P. hispanica* was recovered by the consensus of bootstrap NJst estimates of both datasets ("full" and "haplotype"), and also by the "ML" NJst estimate of the "full" dataset, but not highly supported. In all cases except in the bootstrap consensus estimate using the "full"

dataset, the western Iberian species *P. virescens, P. carbonelli, P. guadarramae* and *P. bocagei* were also recovered as a clade, although with low support.

Table 3.4. Models of sequence evolution obtained with jModeltest2 and the models used in *BEAST for each gene. Nst, number of substitution rate categories; AICc, Akaike Information Criterion with correction.

Loci	jModeltest2				*BEAST
	All individuals		Haplotypes per species		
	Model_AICc	Nst	Model_AICc	Nst	Model_AICc
ACM4	K80+I+G	2	K80+I	2	HKY+I
β-fibint7	JC	1	K80+G	2	HKY+G
C-mos	JC	1	K80+G	2	HKY+G
MC1R	JC	1	HKI+I+G	2	HKY+I+G
NFYCint16	K80+G	2	HKI+G	2	HKY+G
PDC	TrNef+I+G	6	K80+I	2	HKY+I
PKM2int5	K80+G	2	K80+G	2	HKY+G
RAG1	K80+I	2	K80+I	2	HKY+I
RAG2	K80	2	K80+I+G	2	HKY+I+G
Pod6b	JC	1	K80+G	2	HKY+G
Pod7b	K80	2	K80	2	HKY
Pod11	TPM3+I+G	6	JC	1	JC69
Pod12b	K80+G	2	K80+G	2	HKY+G
Pod13	K80+I+G	2	K80+G	2	HKY+G
Pod14	JC	1	K80+G	2	HKY+G
Pod14b	TPM2uf+G	6	JC	1	JC69
Pod15	K80+G	2	K80+G	2	HKY+G
Pod15b	TPM3uf+I+G	6	K80+G	2	HKY+G
Pod16	TIM1+I+G	6	K80+G	2	HKY+G
Pod17	HKY+G	2	HKI	2	HKY
Pod20	HKY+I+G	2	JC	1	JC69
Pod21	TrNef+G	6	JC	1	JC69
Pod25	TrNef+G	6	TIM1ef+G	6	GTR+G
Pod31	JC	1	TrNef+G	6	GTR+G
Pod33	HKY+I+G	2	JC	1	JC69
Pod38	JC	1	K80+G	2	HKY+G
Pod43	K80	2	K80	2	HKY
Pod55	K80+G	2	K80	2	HKY
Pod69	K80	2	K80	2	HKY
Pod72	JC	1	K80+I+G	2	HKY+I+G

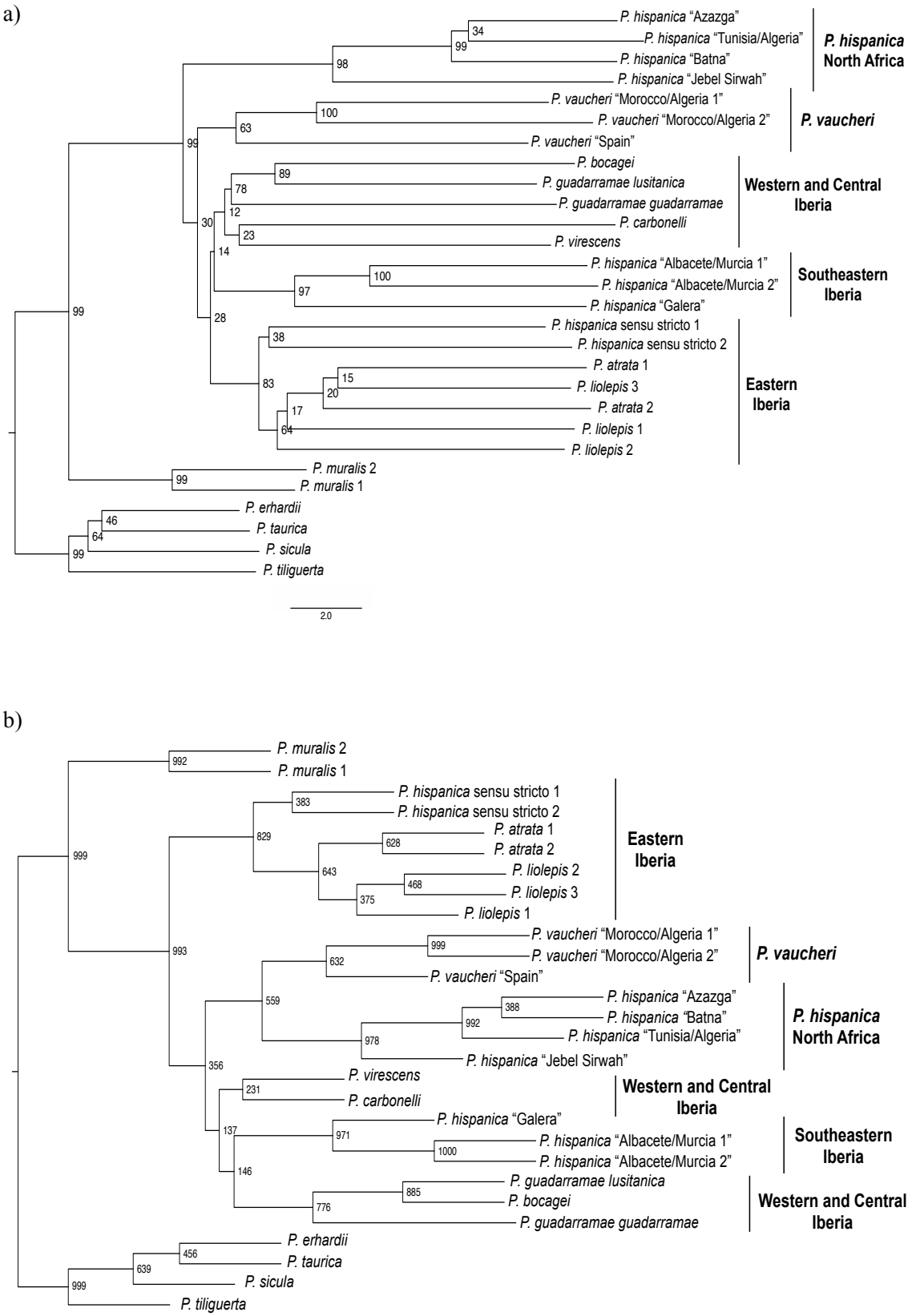


Figure 3.3. Phylogeny of Iberian and North African *Podarcis* species as estimated with the NJst method using each gene ML tree topology from the "haplotypes" dataset; a) "best" NJ species-tree; b) NJst consensus (branch lengths are transformed and do not reflect any amount of evolution). Trees were inferred unrooted and rooted for visualization in the branch leading to *P. erhardii*, *P. taurica*, *P. sicula* and *P. tiliguerta*. The numbers on the nodes indicate multilocus bootstrap support values for branches calculated following Seo, 2008.

Species-trees inferred by MP-EST are shown in Figure 3.4.

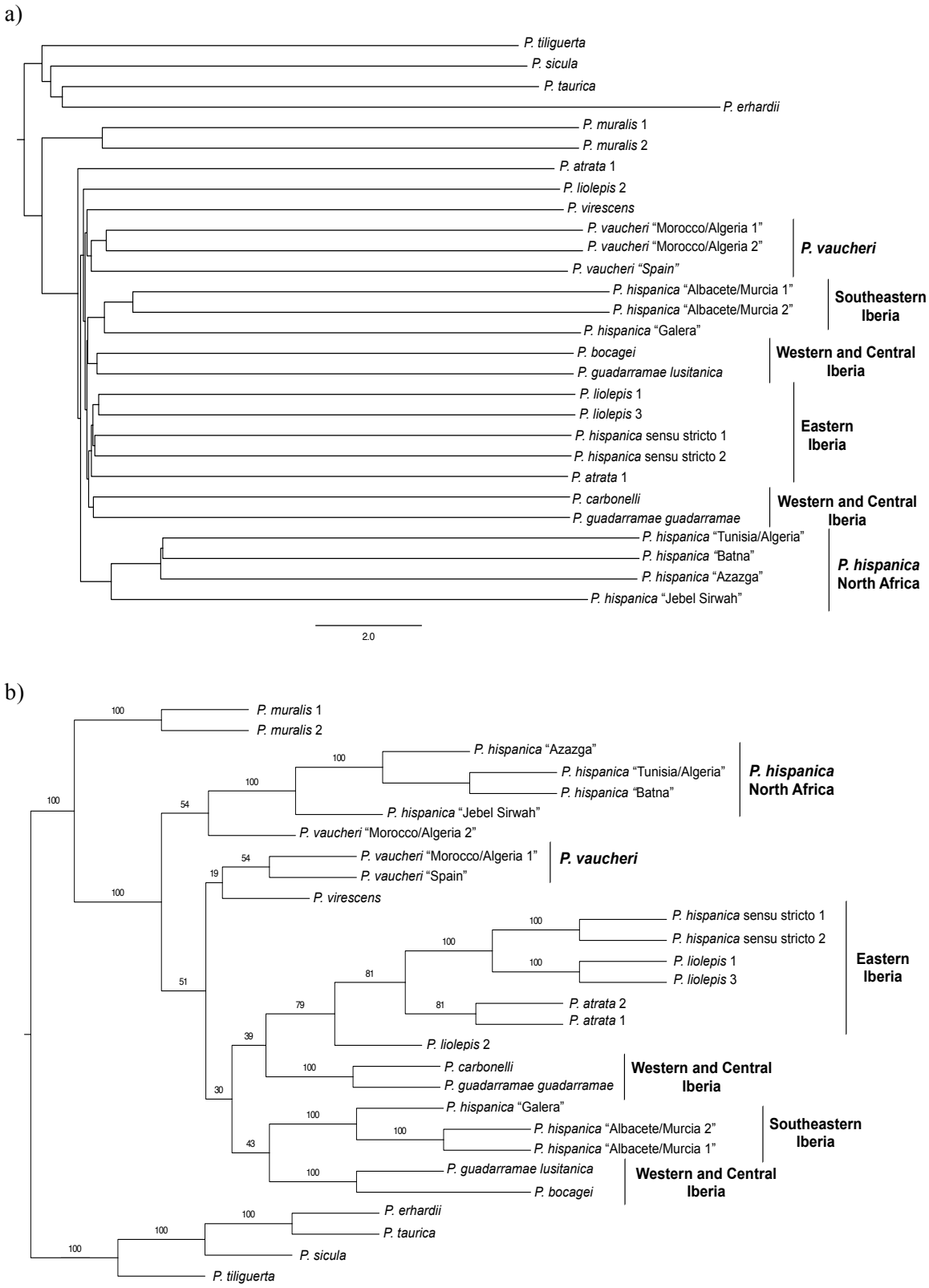


Figure 3.4. Phylogeny of Iberian and North African *Podarcis* species as estimated using the maximum pseudolikelihood coalescent method, MP-EST, using the 27 loci ML tree: a) maximum pseudo-likelihood estimate of 100 independent searches (branches are in coalescent units, $T = 2\tau/\theta$); b) extended consensus of the 100 searches (numbers on the nodes indicate the number of times a given bipartition was found in the 100 searches; branch lengths are transformed and do not reflect any amount of evolution).

The MP-EST tree topology recovered largely coincides with previous estimates from NJst. Focusing on the consensus-tree, previously described groups 1, 3, 4 and 5 are again recovered with maximum "support". The main differences are that *P. vaucheri* from Morocco (2) is not recovered as in the same clade as its remaining conspecifics, but instead groups with the African *P. hispanica* forms (in the consensus tree, with low support), and also that here a clade with *P. carboneli* and *P. guadarramae guadarramae* is recovered in all tree searches. This specific bipartition was never found in any of the NJst tree searches.

The species-tree estimate (extended consensus) obtained with the Bayesian gene-trees distributions and Guenomu is presented in Figure 3.5. In this case, no clade is supported. The strict consensus tree (not shown) is a polytomy between all units, showing that in the 1000 species-trees used to construct the consensus no common clade exists in the big majority of the gene-trees.

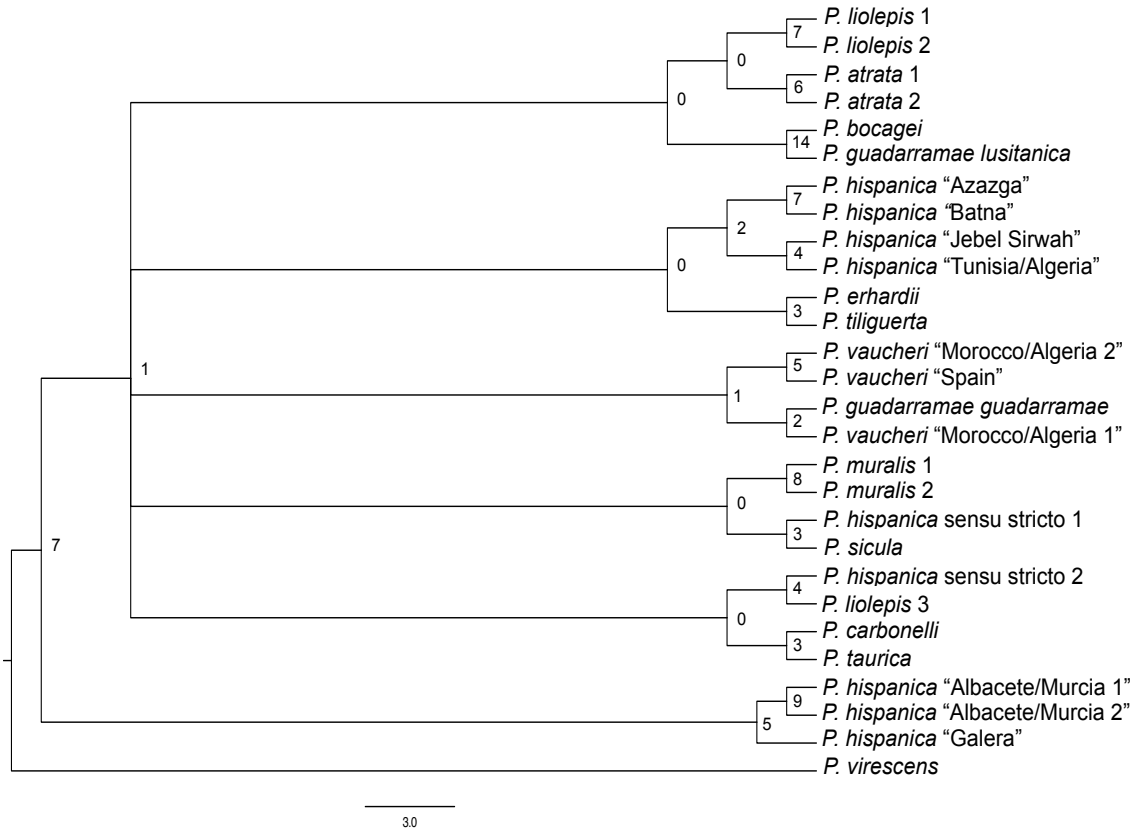


Figure 3.5. Extended consensus of the posterior distribution of species-trees obtained with Guenomu. Values indicate the posterior probabilities.

Some results from *BEAST are shown in Figures 3.6 and 3.7. The distributions of tree topologies based only on species-trees consensus (consensus per topology within run) can be found on Appendix. Convergence revealed to be extremely hard to achieve for many of the parameters, especially for gene-tree likelihoods and gene-tree heights, and, despite the very long lengths of the Markov chains already sampled, no single run can be yet considered to present satisfactory effective sample size (ESS) values, stationarity or convergence.

Overall, across runs, sequence model parameters such as base frequencies, substitution rates and heterogeneity parameters (G and I), as well as clock rates, do apparently achieve convergence and high ESS values, while gene-tree likelihoods, gene-tree heights, as well as overall likelihood and posterior almost never do. Likewise, convergence at models for tree priors and population sizes was never achieved despite the large number of (large) runs performed.

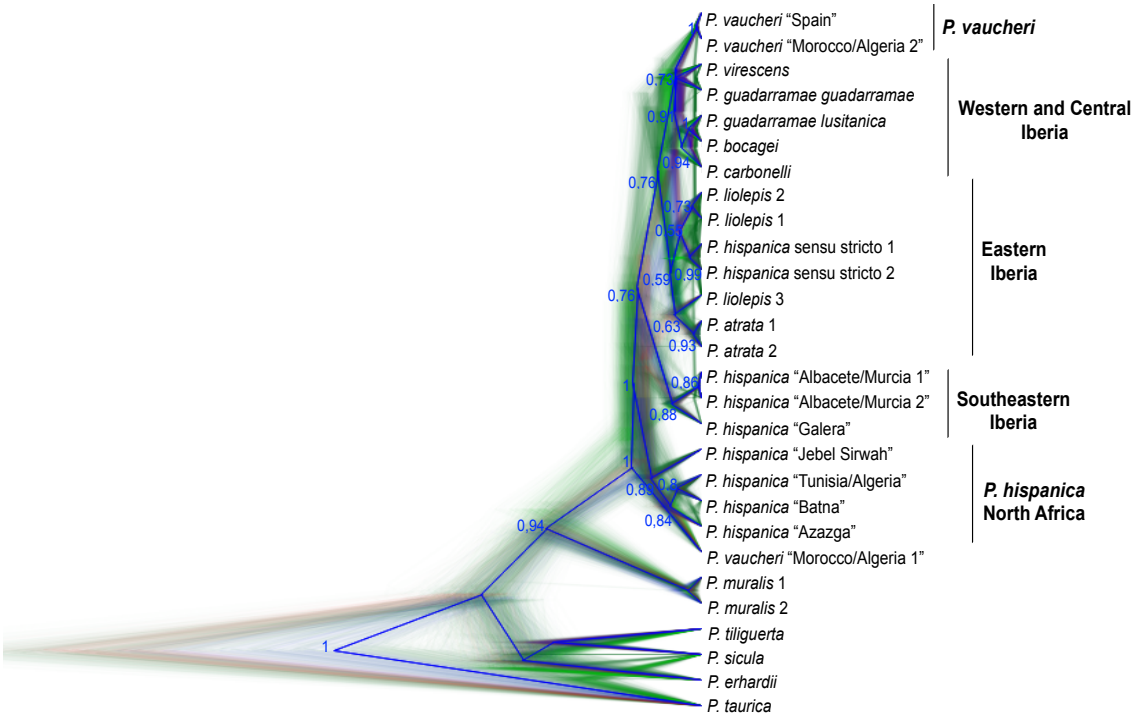


Figure 3.6. Posterior density of species-trees (cloudogram) from *BEAST analyses for the Iberian and North African *Podarcis* species for all loci, for a chain length/tree prior of 341M/Yule. Each thin line corresponds to a sampled tree, so darker areas correspond to higher density of trees in agreement. Blue sets of trees represent those with the same topology as the most popular tree, the next most popular set appearing in red, and the third most popular green. Remaining trees are all dark green. Uncertainty in node heights is shown by smears around the mean node height. Maximum clade credibility tree and posterior probabilities of support above 50 are show in blue.

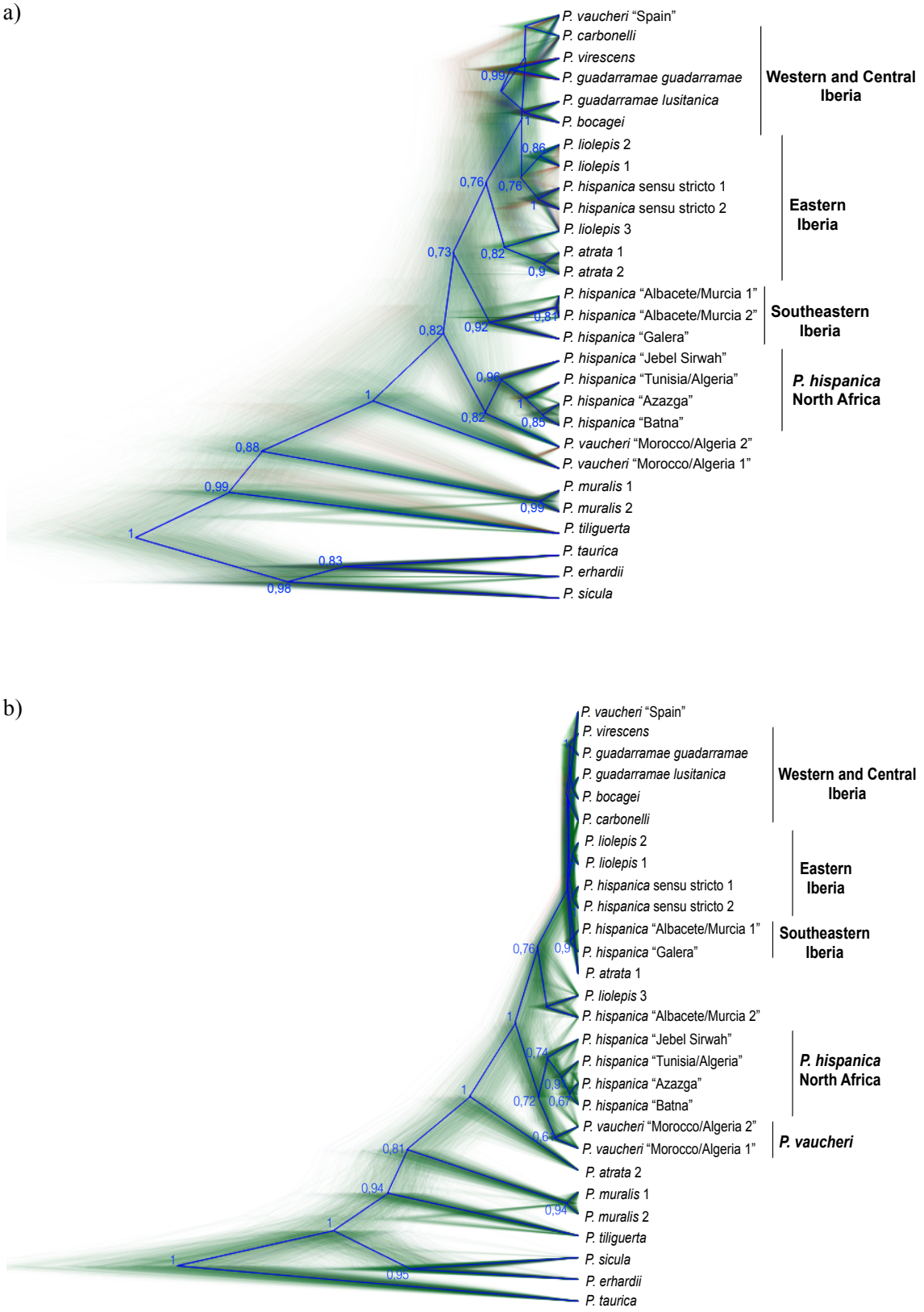


Figure 3.7. Posterior density of species-trees (cloudogram) from *BEAST analyses for the Iberian and North African *Podarcis* species for 21 loci, for a chain length/tree prior of a) 605M/Yule; b) 168M/coalescent. Each thin line corresponds to a sampled tree, so darker areas correspond to higher density of trees in agreement. Blue sets of trees represent those with the same topology as the most popular tree, the next most popular set appearing in red, and the third most popular green. Remaining trees are all dark green. Uncertainty in node heights is shown by smears around the mean node height. Maximum clade credibility tree and posterior probabilities of support above 50 are show in blue.

Overall, runs with the 21 loci (after removing the 9 loci for which ESS for tree likelihoods did not improve in the course of the runs) subset behave much better, with many of the parameters seeming to reach stationarity (605 million generations), including some tree likelihoods and tree heights. Yet, many tree likelihoods still had very low ESS values indicating non-stationarity of the Markov chain.

Runs using a coalescent tree prior on the species-tree also seem to behave better (higher ESS values for tree priors, tree likelihoods, thee heights and posterior) than runs using Yule or birth death priors, although current chain length is still low to access stationarity and stability of these values (168 millions generations).

Yet, some overall common aspects and trends can be observed so far: 1) the tree is generally rooted with *P. taurica*, *P. erhardii* and *P. sicula* (or these plus *P. tiliguerta*); 2) *P. muralis* is sister to but well differentiated from the Iberian and North African species, which define a well supported clade; 3) the forms of *P. hispanica* from North Africa always define a clade, sometimes also including two or one of the North African *P. vaucheri* lineages. *P. vaucheri* "Morocco/Algeria 1" alternates between being placed basal to the whole "ingroup" and related to the North African *P. hispanica* forms (sometimes with *P. vaucheri* "Morocco/Algeria 2"). On the other side, *P. vaucheri* "Spain" is often placed within the "Western and Central Iberian" group; 4) *P. hispanica* "Albacete/Murcia 1" and *P. hispanica* "Galera" always define a clade (*P. hispanica* "Albacete/Murcia 2" tends to a group with this clade except in the tree of 168M of 21loci, in which this group is placed as sister to *P. liolepis* 3); 5) *P. liolepis* 1 and 2 from Northeast Spain, and *P. hispanica* sensu stricto from Southeast of Spain tend to define a clade; as well as 6) *P. liolepis* 3 and *P. atrata* 1 and 2 (except in the tree of 168M of 21loci, where *P. atrata* 2 is placed as sister to the whole "ingroup"). In the run with all loci, this Eastern Iberian group (5+6) is actually recovered with considerable support and 7) *P. bocagei* and *P. guadarramae lusitanica* are always recovered as sister taxa and also closely related to *P. carbonelli, P. guadarramae guadarramae* and *P. virescens.* These last species seem to be closely related but relationships at the base of the group, as well as basal relationships between all clades remain largely variable.

Sampling from the prior only (without data), with both datasets, results in different values and distributions, showing that the data, and not the prior, are responsible for the posterior distributions (Appendix).

Chapter 4

Discussion

This thesis focused on the study of the evolutionary history of *Podarcis* species from the Iberian Peninsula and North African based entirely on nuclear DNA markers. This assessment had been previously impossible due to i) the lack of suitable markers; ii) the unavailability of species-tree determination methods thoroughly accounted for causes of gene-tree/species-tree incongruence, mainly ILS. This was a critical step in our understanding of the patterns of speciation in this group of organisms.

Phylogenies inferred for this group had been so far based on allozymes, suffering from some lack of resolution in the estimate of relationships, or mitochondrial DNA, which clearly had phylogenetic signal but portrayed a mere single locus perspective. In fact, mtDNA has been in general the major source used to describe genetic variability. The easiness of amplification, even for non-model organisms, its high mutation rate, its prompt acquirement of reciprocal monophyly due to its lower effective population size, and the absence of recombination make it an attractive marker to use. Nevertheless, mtDNA is also largely dependent on stochasticity and selection (Hudson and Turelli, 2003) and, with a few exceptions, exclusively of maternal transmission, making the use of the single mtDNA locus to infer evolutionary relationships often questionable and with many possible drawbacks. By contrast, nuclear genes reveal the history of males and females, plus recombine, so that genes on the same chromosome may have different evolutionary histories. They are thus an unlimited source of information and make possible to address different evolutionary questions. In this work our goal was to use data from multiple unlinked sequence nuclear loci to estimate phylogenetic relationships among Iberian and North African *Podarcis* wall lizards, while evaluating the ability of current methods to deal with real and challenging datasets.

4.1 Gene genealogies: extensive allele sharing among species and mtDNA lineages

The results show that *Podarcis* species share a great deal of nuclear genetic variation by opposition to was observed with mtDNA in previous studies (Harris et al., 2002; Pinho et al., 2006; Kaliontzpoulou et al., 2011). Despite the fact that some lineages/groups were previously defined based on mtDNA reciprocal monophyly (e.g. *P. guadarramae guadarramae* vs. *P. guadarramae lusitanica;* the various unnamed forms of *P. hispanica*), others are for long recognized as different species, not based only on mtDNA exclusiveness (e.g. *Podarcis bocagei*, *P. carbonelli*), and even these exhibit, for most loci, high degrees of haplotype sharing. The same high degrees of haplotype sharing had already been detected by Pinho *et al.*, (2008) at two nuclear introns, in a pattern very similar to the one now detected at 29 additional markers. A notable exception to this pattern is the *Pod7b* marker that shows almost monophyletic clades for several mtDNA lineages, with some degree of haplotype sharing between individuals carrying different (but closely related) mitochondrial lineages. Furthermore, *P. muralis*, that was used as outgroup and represents a relatively distinct lineage according to previous phylogenies and to the species-trees inferred here, for some loci is not monophyletic, sharing alleles with lineages from the Iberian Peninsula and North Africa.

Such remarkable discrepancy between mtDNA and nuclear data could suggest that mtDNA lineages do not correspond to true evolutionary entities and their differentiation could be related to effects such as selection acting only in the mitochondrial genome. However, this does not seem to be the case due the fact that morphological analyses (Sá-Sousa *et al.*, 2002; Kaliontzopoulou *et al.*, 2012) and other nuclear markers (allozymes) clearly distinguish these lineages, or at least the majority of them (Pinho *et al.,* 2007). In fact, even if each single genealogy does not show any obvious differentiation among evolutionary groups, the data herein presented allow a clear distinction of each species/lineage (see Results) when multilocus genotypes are considered. The phylogenetic relationships between these groups are not yet clear for all cases, probably because both fast diversification with high degree of incomplete lineage sorting still observed, and some potential gene flow between lineages. Disentangling between these two sources of allele sharing is difficult and, of course, both can be present simultaneously.

4.2 Iberian and North African *Podarcis* lineages as defined by nuclear markers

The results from STRUCTURE showed that with the 30 nuclear loci, we could distinguish all the groups that had already been recognized before with mtDNA (and other nuclear) markers plus a few more, which resulted from further subdivisions of *P. atrata, P. muralis, P. liolepis*, *P. hispanica* sensu stricto*, P. vaucheri* "Morocco/Algeria" and *P. hispanica* "Albacete/Murcia". Some of these cases are concordant with geography; a case in point is that of *P. atrata,* that was subdivided in two clusters corresponding to two groups of islands within the Columbretes, but others may also be true (the groups within *P. liolepis* or *P. vaucheri* from Morocco and Algeria, for example). Obviously, such results would need confirmation using larger sample sizes. Previous groups not yet evaluated with nuclear data (*P. hispanica* "Albacete/Murcia"; *P. hispanica* "Azazga"; *P. hispanica* "Batna") can also be distinguished using multilocus genotypes. One exception in the overall concordance between nuclear and mitochondrial DNA was the lack of clear boundaries between *P. liolepis* and *P. hispanica* sensu stricto, given the inclusion of some individuals of the latter mtDNA lineage in the former nuclear cluster and the existence of several individuals admixed between these clusters and also with affinities to *P. atrata,* in a pattern suggestive of intraspecific variability. Another discordant feature is the inferred lack of differentiation between *P. vaucheri* "Southern Spain" and *P. vaucheri* "Southern-central Spain" mtDNA lineages.

Again, we note that we refer to these groups interchangeably as "groups" or "units" or "lineages". For species-tree methods, these are the designated "species" (which do not need to consist of "species" in the taxonomic sense, but may be any kind of recognizable population structure, which many of the groups here are). Whether they should be recognized as distinct species (in the taxonomic sense) or not is not possible to ascertain without more information, and was not the goal of this thesis. Yet, current new results from STRUCTURE coupled with inferred phylogenetic relationships can be used to make some taxonomic considerations (described in section 4.6 below).

4.3 Incomplete lineage sorting versus gene flow between Iberian and North African *Podarcis*

Incomplete lineage sorting arises from the retention of ancestral polymorphism through speciation events. One of the aspects the data now collected further highlights is that *Podarcis* species have diverged very rapidly, with lineages not having time to achieve reciprocal monophyly at most loci. Incomplete lineage sorting seems thus to be the most widely accepted hypothesis for the abundant haplotype sharing, as also inferred from previous work (Pinho *et al.*, 2008).

Nevertheless, gene flow was also inferred. Results from STRUCTURE obtained in this thesis showed cases of gene flow among forms/species, some of which had never been reported before (as well as gene flow between probably intraspecific forms). Twenty-two individuals were identified as having a possibly admixed genotype, and a few more were assigned to a different group than suggested by its mitotype Some of these cases were already expected because they involve species that are found in sympatry: for example, gene flow was detected between *P. bocagei* and *P. guadarramae lusitanica*, in similarity to various previous reports (Pinho *et al.*, 2007a, 2008; Pinto, 2013). There was also one individual carrying the *P. hispanica* sensu stricto mtDNA lineage that was assigned to *P. hispanica* "Galera", a situation that had also been reported before (Pinho *et al.*, 2008; Renoult *et al.*, 2009). Other cases were totally new, such as the finding of an individual admixed between *P. vaucheri* "Spain" and *P. carbonelli* in the sympatric zone of Doñana. Individuals presenting signs of admixture between *P. vaucheri* "Spain" and *P. virescens* were also found, although far from the putative contact zone between these species. Individuals showing signs of admixture were also detected between the geographically close *P. hispanica* "Galera" and *P. hispanica* "Albacete/Murcia". Probably one of the most striking results was the high number of admixed individuals between *P. liolepis, P. hispanica* sensu stricto and *P. atrata*. These individuals are hard to explain taking into account hybridization, not only because of its triple nature but also because *P. atrata* is an island form. Other cases of gene flow were identified between clusters found within mtDNA lineages. All these cases most likely reflect the overall lack of differentiation between these clusters and their probably intraspecific nature. Additionally, there is one case where it was observed one individual of *P. carbonelli* mtDNA lineage with signs of admixture with the North African *P. hispanica* group. This result can only be an artifact because is impossible that gene flow occurs between these forms.

As the species-tree inference methods here used all assume absence of gene flow between "species" and because the approach used of removing admixed individuals as detected by

STRUCTURE may fail to detect (older) historical gene flow, we further applied the coalescent model of divergence with gene flow, IMa2 (Hey and Nielsen, 2007), to these groups of *Podarcis* in a pairwise manner, in a similar way as Pinho *et al.* (2008). Surprisingly, gene flow was now detected among most of the species. Yet, these results have to be considered with caution as our inference scheme, performing pairwise tests between all units (even those which are clearly not sister taxa), is a clear violation of the IM model assumptions (Strasburg and Rieseberg, 2009). In many of the pairwise comparisons, the possibility of unaccounted gene flow from external taxa may strongly limit the validity of our inferences. Moreover, a thorough evaluation of levels of historical gene flow between Iberian and North African *Podarcis* was outside the scope of this thesis. Thus, we provisionally decided not to take these results into account until we can more accurately test for the presence of gene flow without the possibility of too many false positive estimates. Nevertheless, we cannot disregard the possibility of unaccounted gene flow in "species-tree" inference, which effect can be both false sister relationships, and the inference of divergence times biased towards more recent times (Leaché *et al.*, 2014).

4.4 Phylogenetic relationships of *Podarcis* "species"

This work is the first attempt so far at inferring Iberian and North African *Podarcis* phylogeny from a multilocus nuclear DNA sequence data. For this, a number of different approaches were used, representative of the main categories of these methods; a distance method (NJst), a maximum "pseudolikelihood" method (MP-EST), and a probabilistic method (Guenomu) (super-tree methods), and finally a fully probabilistic bayesian co-estimation method (*BEAST).

Starting by the "outlier", the topology inferred by Guenomu is the less resolved one and does not seem to be biologically realistic in many aspects. In fact, the strict consensus of the recovered species-trees distribution was a full polytomy. Despite the fact that some of the inferred groups and/or sister relationships (extended consensus) do make sense and agree with other methods (eg. *P. liolepis* and *P. atrata*; *P. bocagei* and *P. guadarramae lusitanica*; the African *P. hispanica* forms, *P. hispanica* "Albacete-Murcia" and *P. hispanica* "Galera"; *P. muralis*), no relationship is actually supported, showing that there is almost no common clade in the 1000 species-trees used to calculate the consensus. Further, and contrary to all other methods, the ingroup is not recovered as monophyletic. Possible reasons for the clear lower performance of this method are explored below (section 4.7)

Regarding inferences from other methods, overall, the results from NJst, MP-EST, and *BEAST show many similarities between them and also with the mtDNA phylogeny (Kaliontzopoulou *et al.*, 2011), as well as some particular differences.

For a start, all estimates agree in a highly supported "ingroup", to the exclusion of *P. muralis*, whose two "groups" also define a monophyletic clade. Then, basal relationships within the ingroup are mostly unresolved, and thus different (and never supported) across methods, but some groups of closely related "species" were commonly recovered, with high support, and are described below. One group that is always recovered with high support by NJst and by MP-EST and which *BEAST runs tend to recover also, is the Eastern Iberian clade of *P. hispanica* sensu stricto, *P. liolepis* and *P. atrata*. Within this group, the two lineages of *P. hispanica* sensu stricto and of *P. atrata* are, respectively, recovered with support as monophyletic, but remaining relationships are recovered differently and unsupported across methods. The interesting exception is MP-EST (consensus), which recovers very well supported relationships within this group, with a non-monophyletic *P. liolepis*, and P. *liolepis* 2 lineage as basal to the whole group. Nevertheless, and especially given that this MP-EST consensus does not represent real "bootstrap" values, this results must be taken with caution and remain as an hypothesis to be further tested.

The remaining species from the Iberian Peninsula (Western-central Iberian group) were sometimes recovered as a clade (although never well supported) – NJst – or at least part of its "species"- MP-EST, *BEAST – generally with unsupported basal relationships. Again, within the group some sister-relationships do are well supported across methods, as the sister relationship between *P. bocagei* and *P. guadarramae lusitanica*. With MP-EST, *P. carbonelli* is also recovered with high support as sister taxa of *P. guadarramae guadarramae*, something which is not recovered with any other method.

The other highly consistent result across all methods was the inference of the clade containing all *P. hispanica* mtDNA lineages from North Africa. In the relationships within this group there were some differences across methods, yet, the lineage from Jebel Sirwah was always recovered as the most basal within this clade.

The most evident difference between the present estimates of relationships and the ones using mtDNA is a swapping in sister taxa relationships among *P.liolepis, P.hispanica* "Galera", *P.hispanica* "Albacete/Murcia" and *P.hispanica* sensu stricto. Indeed, current nuclear data places *P. liolepis* as closely related (perhaps conspecific or sister) to *P. hispanica* sensu stricto and *P. hispanica* "Albacete/Murcia" as sister to *P. hispanica* "Galera". However, mtDNA recovers as sister taxa the pairs *P. hispanica* sensu stricto/*P. hispanica* "Albacete/Murcia" and *P.liolepis/P. hispanica* "Galera", the two pairs very distantly related.

It is not the first time that this close relationship between *P.liolepis* and *P.hispanica* sensu stricto is suggested. It was also inferred with allozymes (Pinho *et al.*, 2007a), low divergence time estimated from nuclear introns (Pinho *et al.*, 2008) and from the lack of morphological and genetic differentiation across the mtDNA lineage's contact zone (Renoult *et al.*, 2009). Using allozymes, these two mtDNA lineages could not be well distinguished, in similarity to the present analyses using STRUCTURE, which despite revealing some differentiation recovers fuzzy boundaries between these taxa, with prevalent cytonuclear discordance and admixture, It thus seems possible that they are in fact conspecifics. A few hypotheses can explain these discordances: 1) it could imply ancestral (or not so) gene flow between divergent taxa with a probable dilution of the nuclear genome but not the mitochondrial one (with the mtDNA genealogy representing the "original" relationships between taxa) or 2) the capture of a foreign mtDNA lineage, corresponding to a now extinct nuclear "unit" (without any corresponding obvious signature in the nuclear genome). This last scenario was also one of the hypotheses proposed by Renoult *et al.*, (2009), which suggests an ancient mitochondrial introgression originating from an evolutionary unit presently absent from the study area.

Remarkably, the close relationship inferred between *P. hispanica "*Albacete/Murcia" and *P. hispanica "*Galera" implies that a similar phenomenon of an ancient mtDNA capture happened in parallel in this species pair. Another possibility is, then, that it was the same mtDNA lineage, the ancestral to one of the pairs (e.g. *P. hispanica* sensu stricto and *P. hispanica* Albacete/Murcia) that was captured by both species of the other pair (e.g. by *P. liolepis* and *P. hispanica "*Galera"), and that it has diverged in the two species (generating the two divergent mtDNA lineages) since then. If this was the case, it is difficult to explain such double introgression without invoking an adaptive nature.

Curiously, the idea above can be seen as the recycling of one proposal by Renoult *et al.*, (2009), which our data actually dismiss. Indeed, these authors also suggested a double introgression, but much more recent, in both cases involving the "modern" *P. hispanica* sensu stricto: one into *P. liolepis* (as our data also may suggest) and another into *P. hispanica* "Galera", which our data do not support. Although we do find instances of cytonuclear discordance involving the *P. hispanica* sensu stricto mtDNA lineage and *P. hispanica* "Galera", the discordance is sporadic, not general, as the two forms can be clearly distinguished even when occupying nearby localities.

Another difference between nDNA and mtDNA estimates of relationships, but this time involving relationships within a clade and not between major clades, is that *P. guadarramae* (previously *P. hispanica* 1A and 1B) is probably paraphyletic since *P. bocagei* appears always highly supported as the sister taxa of *P. g. lusitanica*.

Inconsistencies between our nuclear trees were also observed. *Podarcis vaucheri* "Spain" (that correspond to mtDNA lineages *P. vaucheri* "Southern-central Spain" and *P. vaucheri* "Southern Spain") is recovered by NJst (both consensus plus "full" dataset) as closely related to *P. vaucheri* "Morocco/Algeria" 1 and 2 and sister of the all other lineages from North Africa (although not strongly supported) but in the *BEAST results this taxa appears most often as closely related to species from Western and central Iberia. With MP-EST (consensus), one of the African *P. vaucheri* lineages is inferred as closely related to the North African *P. hispanica* forms, while the other two lineages forms a separate group (without a supported relationship to any other group). At *BEAST runs, some of the topologies recovered related *P. vaucheri* (African) lineages with the North African forms of *P. hispanica*, but also *P. vaucheri* from Spain was often related to the West-central Iberian clade. Overall, this seems to support a relationship between *P. vaucheri* and other North African forms, with the placement of *P. vaucheri* from Spain close to the Western Iberian group possibly caused by some degree of unaccounted-for gene flow (possibly with *P. carbonelli* or *P. virescens*).

As a last mention to aspects of the presented phylogenies, at some *BEAST runs *P. vaucheri* "Morocco/Algeria 1" and *P. atrata* 2 were sometimes clearly distant from all the others species, as also *P. liolepis 3* tended to be inferred as related to quite different taxa across runs. These "species" include only one (*P. vaucheri* "Morocco/Algeria 1" and *P. liolepis* 3) or two individuals (*P. atrata* 2) and we hypothesize that is the cause of their behavior and of their position being more difficult to estimate with this co-estimation method. Due to the fact that none of the *BEAST runs here presented achieved satisfactory converge, we do not have much confidence in these topologies overall, although we believe the results support some of the inferences made from other methods, especially regarding consistently recovered relationships. Moreover, sampling from the prior only yielded very different results from those obtained with data, indicating also that there is some information in the trees so far been obtained.

4.5 Biogeographic implications

Another interesting aspect of the evolutionary history of the Iberian and North African *Podarcis* is the biogeography of the group around the Strait of Gibraltar. For some species, the opening of the Strait was the probable cause for separation between Iberian and North African species (Maia-Carvalho *et al.*, 2014), but in the case of *Podarcis* the successive mtDNA phylogenies suggest otherwise. Indeed, the Strait has not worked as a complete barrier to migration, and the distribution of genetic variation requires two independent events, either two

transmarine colonizations or one vicariant event followed by the crossing of the Strait (Harris *et al.*, 2002; Pinho *et al.*, 2006; Kaliontzopoulou *et al.*, 2011). Either we accept or not that *P. vaucheri* and the North African *P. hispanica* forms are related, the results are concordant with the two-event scenario suggested by the mtDNA, since there is one Iberian form (*P. vaucheri* "Spain") grouped within a North African clade. This observation thus implies, again, at least one transmarine colonization. However, the directionality of the colonization cannot be inferred from the present estimates of relationships.

4.6 Taxonomic implications

A taxonomic reevaluation of the clade is clearly beyond the scope of this thesis. However, it is possible to make some reflections on this subject.

In general our results support the distinctiveness of currently accepted species. Particularly, *P. bocagei, P. carbonelli* and *P. virescens* are clearly diagnosable genetically and morphologically. *P. vaucheri* is also clearly distinct from other species, although it is difficult to evaluate the taxonomic status of forms inhabiting South Iberia and North Africa.

P. guadarramae, on the other hand, is recovered as paraphyletic, since *P. guadarramae lusitanica* (former *P. hispanica* type 1A) is placed as sister to *P. bocagei* in our analyses. If true, this suggests that the taxonomy of these forms, which was recently redefined (Geniez *et al.*, 2014), will require a new reevaluation. However, the sister taxa relationship between *P. bocagei* and *P. guadarramae lusitanica* could eventually be biased by high levels of gene flow between these sympatric species.

With respect to *P. liolepis*, the genetic proximity and in some cases paraphyly with respect to *P. atrata* may indicate that these two forms are closely related or even conspecific, as suggested previously by Harris and Sá-Sousa, (2002) and Renoult *et al.*, (2010). However, the distinctiveness of *P. atrata* (and of different island groups within it) raises the possibility that it may well deserve a different taxonomic status. *P. liolepis* seems to be also closely related to populations exhibiting the *P. hispanica* sensu stricto mtDNA lineage. This is true for populations from the Northern area of the distribution of this lineage, in the provinces of Cuenca and Valencia, as suggested by Renoult *et al.*, (2009) but also for the South of its distribution, in the provinces of Jaén and Granada. Given the low level of differentiation and the prevalence of admixed individuals between clusters ascribed to *P liolepis*, *P. atrata* and *P. hispanica* sensu stricto, it is probably more conservative to assume that these clusters all correspond to phylogroups within *P. liolepis*. Further studies of the contact zones between

clusters will be highly valuable in order to understand whether there are reproductive barriers among them.

The recovered phylogenetic relationships also highlight that the forms that still remain under the designation of *P. hispanica* are unrelated. It includes at least two groups, one from Southern Iberian Peninsula and another from North Africa. Within each group, it is difficult to assess whether or not the taxa deserve the species status. On one hand, all the mitochondrial DNA lineages correspond to genetically diagnosable groups; on the other hand, both the pair "Albacete/Murcia" and "Galera" and the set "Tunisia and Algeria", "Jbel Siroua", "Batna" and "Azazga" are obviously closely related. It is this difficult to evaluate, with the data in hands, their taxonomic status. Again this would require an extensive study of contact zones, which might be possible in the first case given their contiguous distribution but likely impossible in the second, given the fragmented distribution patterns of these lineages.

4.7 Comparison between methods: factors affecting species-tree inference

The choice of the method can often have a large impact in the analyses, as different methods have different accuracies in different kinds of datasets. It is thus of most importance to be aware of the expected error rates and specific biases of the methods we use in the specific characteristics of the datasets we have on hands.

NJst, MP-EST and Guenomu are simple and extremely fast methods compared with full probabilistic approaches such as *BEAST and seem to provide good approximations for large amounts of data, often outperforming the other methods of the same class. Constructing a species phylogeny with NJst and MP-EST methods take only a few minutes (given, of course, gene-trees are already obtained from a separate method at the users choice). Yet, from a practical point of view, both often require the manipulation of large amounts of data, and sometimes the performance of not-so-trivial operations such as rooting a big amount of gene-trees, often not with the same taxa (MP-EST), therefore requiring some scripting abilities, and thus, the time-investment. Yet, they are very fast, and perfectly usable at scales of hundreds of loci; contrary to *BEAST, for which datasets of few tens of loci and "species" already reveal problematic.

Guenomu takes slightly more time (especially if we account for the Bayesian gene-tree search), but yet runs in a few hours, which is remarkable compared with *BEAST that may need months to achieve convergence.

The authors of these methods have tested them extensively on simulated data sets, which typically include few species and few individuals per species (especially true in the case of NJst and MP-EST; (Liu *et al.*, 2010; Liu and Yu, 2011). Further research with multilocus data has been done to compare these methods in the presence of various levels of incomplete lineage sorting, showing some problems that could influence their accuracy (Yang and Warnow, 2011; Bayzid and Warnow, 2013; Mirarab *et al.*, 2014). These studies concluded that in general the results can be consistently improved as the number of genes increases, but also that the methods are highly sensitive to violations of their assumptions caused for example by error in gene-tree inference and introgressive hybridization (Leaché *et al.*, 2014). Yet, simulations also show that for trees involving very short branch lengths (or at least a proportion of branches which is very short) and/ or for larger population sizes (which increase the average proportion of gene-tree/species-tree discordance), it may not be possible to infer the "true" species-tree (Leaché and Rannala, 2011; Mirarab *et al.*, 2014).

In the case of *Podarcis* we are dealing with closely related "species", where genetic distances are very small and often zero. This is likely a reason for high levels of gene-tree error in our estimates and consequently high species-tree estimation error. This was perhaps one of the reasons of Guenomu to have failed in our dataset, for example, given that this program performs better at simulations than other supertree methods, including STAR (Liu *et al.*, 2009b), which is almost identical to NJst. Additional reasons for the lower performance of Guenomu may involve the lack of convergence of Bayesian gene-tree distributions. Although runs were quite long and convergence was carefully examined, decision about runs convergence and stabilization is often not straightforward, Finally, it may also be that the method itself (built for datasets involving both duplications and losses and ILS) does not perform well when only ILS is present in the dataset - being a "reconciliation" method (in the sense that it tries to optimize some cost function), dup-losses are almost always very informative. Yet, some of the simulations performed with the method (de Oliveira Martins *et al.*, 2014), and across which this method still outperformed others, involved very low rates of dup-loss and very high rates of ILS. It is certainly interesting to further explore the performance of this method in other cases of ILS only and to investigate the reasons for its apparent very low performance in this dataset.

*BEAST is expected to have better performance than supertree methods and also seems to produce more accurate trees than the other methods of the same class, like BUCKY and STEM (Kubatko *et al.*, 2009). In our case, although *BEAST seems to be walking towards a similar solution to that obtained using NJst or MP-EST, and many common aspects can be found, it is clearly much slower, and so far, after multiple runs of several weeks, most parameters still did not stabilize or converge. This was expected since this approach is computationally intensive and the size of the dataset seems to be a crucial factor for full probabilistic methods. Dealing

with this problem and improving the scalability of *BEAST is the object of current research and some strategies are being developed (Zimmermann *et al.*, 2014), which may be definitely worth to try.

As this thesis clearly illustrates, the applicability of these methods in cases of very shallow divergence with extensive lineage sorting and haplotype sharing can be very challenging. Because these methods are relatively new and hence poorly tested on empirical data, it remains to be seen whether the difficulties inherent to the analysis of this dataset are shared by other case studies, suggesting that is difficult to apply these methods on real data, under complex scenarios like those represented here.

Chapter 5

Conclusions and Future Perspectives

Iberian and North African *Podarcis* wall lizards are an example of a complex group that clearly illustrates how fast diversification coupled with gene flow can cause complex patterns that make the evolutionary history extremely difficult to infer.

The nuclear data here analysed increased our knowledge about the history of this group and allowed us to obtain another perspective into the levels of genetic polymorphism and distinctiveness of evolutionary units, as well as into their phylogenetic relationships. Inferences based on mtDNA or morphological characters were in some cases corroborated and in others rejected. In summary, our results suggest that:

1) Evolutionary lineages still share an important proportion of alleles and have not had time to achieve reciprocal monophyly at the majority of the genome.

2) Despite this lack of single-locus differentiation lineages can be easily diagnosed using a multilocus framework.

3) Gene flow among lineages is prevalent, particularly among sympatric forms (*P. bocagei* and *P. guadarramae lusitanica*, *P. carbonelli* and *P. vaucheri*, *P. hispanica "*Galera" and populations carrying the *P. hispanica* sensu stricto mitotype) but also between parapatric forms such as *P. hispanica* "Galera"/*P. hispanica* "Albacete/Murcia", or *P. vaucheri* and *P. virescens*. A high number of admixed individuals were also observed for *P. liolepis/P. hispanica* sensu stricto/*P. atrata,* which may indicate conspecificity.

4) A well-supported Eastern Iberian clade, composed by *P. hispanica* sensu stricto, *P. liolepis* and *P. atrata*, is typically recovered by the different methods, in similarity to a Southeastern Iberia clade placing *P. hispanica* "Galera" and the phylogroups within *P. hispanica* "Albacete/Murcia" as sister taxa. This is a clear difference compared with mtDNA that recovers as sister taxa the pairs *P. hispanica* sensu stricto/*P. hispanica* "Albacete/Murcia" and *P.liolepis/P. hispanica* "Galera", suggesting repeated, perhaps adaptive, phenomena of mtDNA capture and/or extensive nuclear swamping.

5) A Western-central Iberian clade composed by *P. bocagei, P. carbonelli, P. virescens, P. g. guadarramae* and *P. g. lusitanica*, was recovered by most methods, in similarity to mtDNA estimates, but with low support. Relationships within this clade are probably different than those suggested by mtDNA.

6) The various forms of *P. hispanica* from North Africa form a clade, with some differences in the relationships between forms; the lineage from Jebel Sirwah is consistently recovered as sister to the remaining forms.

7) The biogeography of *Podarcis* across the Strait of Gibraltar is complex, as suggested by mtDNA variation, and requires an explanation involving either two colonizations or one episode of vicariance and a transmarine colonization.

8) *P. bocagei, P. carbonelli, P. virescens* and *P. vaucheri* are indeed distinct and well-defined species. However, *P. guadarramae* is recovered as paraphyletic, since *P. guadarramae lusitanica* is sister of *P. bocagei* suggesting that a taxonomic reevaluation concerning this species is probably needed. *P. liolepis*, *P. atrata* and *P. hispanica* sensu stricto could correspond to phylogroups within the same species, although more studies are needed to understand the level of differentiation between these forms.

This work also highlighted the advantage of applying different methods and comparing inferences across methods as measures of reliability, particularly in cases where the real history is not known. Our results suggest that:

1) Full probabilistic methods like *BEAST may fail to achieve convergence when data sets are complex, even after run times in the order of a few months under nearly optimal conditions.

2) Faster approaches like NJst and MP-EST are promising in the sense that, when compared with the preliminary results of *BEAST, show several common points which are very likely to represent true inferences. Guenomu, on the other side, has indeed a low performance in our dataset; the reasons behind this failure are probably worth investigating.

Relying on the results above, new questions have arisen and further research along several lines could be done:

1) It is important to further investigate on the putative evidences of gene flow here found, particularly between the forms where it had never been detected before. As such, and given the application of IMa revealed so time-consuming (and always possibly affected by the violation of its assumptions), alternative ways of testing gene flow have to be explored, either applying it to multiple "species" using subsets of the current dataset, either using simulation tools and hypotheses testing (Heled *et al.*, 2013; Hoban, 2014). Clearly, gene flow is important in this system, and only methods that are capable of co-estimating "population structure" and migration can ever fully capture the complexity of the data.

2) A question we deliberately did not pursue in this thesis was "species delimitation". Across all our analyses, "species" do not necessarily refer to "species" but to any other kind of recognizable population structure, often interchangeably. It would be most interesting in the future, to attempt to use genetic data for species delimitation. Most current methods for species-delimitation are still very hard (if not impossible) to apply to a huge number of putative species, as they basically work through testing alternative hypothesis of numbers of "species", which can be really hard to define as its number becomes elevated (Carstens *et al.*, 2013). Maybe now, that some clades seem to be possible to be defined with confidence; this methodologies can be applied in subsets of the whole group with success. Also, new and more scalable methods are currently being developed, that may be applicable to larger datasets, in general, and require less a-priori hypothesis statements (see Carstens *et al.*, 2013).

3) Particularly interesting from the point of view raised above are methods that require no *a-priori* species assignment, and that can co-estimate "populations" (Choi and Hey, 2011) or "species" (O' Meara, 2010) assignments and trees, or that can deal with a high number of putative hypothesis in a rapid manner (see www.brianomeara.info/phrapl). All these are very promising avenues of research to better understand *Podarcis* evolutionary history.

4) Simulations using some of the methods mentioned above, as well as Approximatte Bayesian Computation (ABC) techniques (Csilléry *et al.*, 2010) will be probably worth to explore to try to distinguish between the alternative hypotheses that can explain the *P. liolepis*/*P. hispanica* Galera – *P. hispanica* sensu stricto/*P. hispanica* "Albacete/Murcia" cytonuclear discordance.

5) Also in terms of phylogenetic inference, a very promising approach is the use of biallelic markers as single nucleotide polymorphisms (SNPs), for example from RAD-sequencing or other genotyping-by-sequencing variants (Davey *et al.*, 2011) and programs such as SNAPP (Bryant *et al.*, 2012). This way, a considerable larger amount of information can potentially be treated in acceptable time, something that will be potentially be very useful for difficult internal nodes.

6) Finally, further studies of the contact zones between clusters will be important to analyze reproductive barriers, evaluate taxonomic status, and fully understand the history of the speciation of this group.

References

Arnold EN, Ovenden D (2002). Reptiles and Amphibians of Britain and Europe. Hong Kong HarperCollins Publ.

Ballard JWO, Whitlock MC (2004). The incomplete natural history of mitochondria. *Mol Ecol* **13**: 729–744.

Bandelt HJ, Forster P, Röhl A (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**: 37–48.

Barraclough TG, Vogler AP (2000). Detecting the Geographical Pattern of Speciation from Species-Level Phylogenies. *Am Nat* **155**: 419–434.

Bayzid S, Warnow T (2013). Naive Binning Improves Phylogenomic Analyses. *Bioinformatics* **9**: 1–8.

Bininda-Emonds ORP (2004). The evolution of supertrees. *TRENDS Ecol Evol* **19**: 315–322.

Bouckaert R (2010). DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* **26**: 1372–1373.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, *et al.* (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLOS Comput Biol* **10**: e1003537.

Boussau B, Szollosi GJ, Duret L, Gouy M, Tannier E, Daubin V (2013). Genome-scale coestimation of species and gene trees. *Genome Res* **23**: 323–330.

Brito PH, Edwards S V (2009). Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica* **135**: 439–55.

Bryant D, Bouckaert R, Felsenstein J, Rosenberg N a, RoyChoudhury A (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol Biol Evol* **29**: 1917–32.

Capella-gutiérrez S, Silla-martínez JM, Gabaldón T (2009). TrimAl : a tool for automated alignment trimming in large-scale phylogenetic analyses. **25**: 1972–1973.

Carretero MA (2008). An integrated Assessment of a group with complex systematics: the Iberomaghrebian lizard genus *Podarcis* (Squamata, Lacertidae). *Integr Zool* **3**: 247–66.

Carretero MA, Marcos E, Prado P (2006). Intraspecific variation of preferred temperatures in the NE form of *Podarcis hispanica. Mainland and Insular Lacertid Lizards: a Mediterranean Perspective. Firenze University Press, Florence*: 55–64.

Carstens BC, Pelletier T a, Reid NM, Satler JD (2013). How to fail at species delimitation. *Mol Ecol* **22**: 4369–83.

Chaudhary R, Bansal MS, Wehe A, Fernández-Baca D, Eulenstein O (2010). iGTP: a software package for large-scale gene tree parsimony analysis. *BMC Bioinformatics* **11**: 574.

Chaudhary R, Burleigh JG, Fernández-Baca D (2012). Fast local search for unrooted Robinson-Foulds supertrees. *Comput Biol Bioinformatics, IEEE/ACM Trans* **9**: 1004–1013.

Chaudhary R, Burleigh JG, Fernández-baca D (2013). Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance. *Algorithms Mol Biol* **8**: 28.

Choi SC, Hey J (2011). Joint inference of population assignment and demographic history. *Genetics* **189**: 561–77.

Csilléry K, Blum MGB, Gaggiotti OE, François O (2010). Approximate Bayesian Computation (ABC) in practice. *Trends Ecol Evol* **25**: 410–8.

Darriba D, Taboada G, Doallo R, Posada D (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* **9**: 696–704.

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

Davey JW, Hohenlohe P a, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* **12**: 499–510.

Degiorgio M, Degnan JH (2010). Fast and consistent estimation of species trees using supermatrix rooted triples. *Mol Biol Evol* **27**: 552–569.

DeGiorgio M, Degnan JH (2014). Robustness to Divergence Time Underestimation When Inferring Species Trees from Estimated Gene Trees. *Syst Biol* **63**: 66–82.

Degnan JH, DeGiorgio M, Bryant D, Rosenberg N a (2009). Properties of consensus methods for inferring species trees from gene trees. *Syst Biol* **58**: 35–54.

Degnan JH, Rosenberg N a (2006). Discordance of species trees with their most likely gene trees. *PLoS Genet* **2**: e68.

Degnan JH, Rosenberg NA (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *TRENDS Ecol Evol Evol* **24**: 332–340.

Douglas ME, Douglas MR, Schuett GW, Beck DD, Sullivan BK (2010). Conservation phylogenetics of helodermatid lizards using multiple molecular markers and a supertree approach. *Mol Phylogenet Evol* **55**: 153–67.

Edwards S V, Liu L, Pearl DK (2007). High-resolution species trees without concatenation. *Proc Natl Acad Sci USA* **104**:5936-5941.

Faith DP, Magallón S, Hendry AP, Conti E, Yahara T, Donoghue MJ (2010). Evosystem services: an evolutionary perspective on the links between biodiversity and human well-being. *Sci Direct* **2**: 66–74.

Flot J-F, Tillier A, Samadi S, Tillier S (2006). Phase determination from direct sequencing of length-variable DNA regions. *Mol Ecol Notes* **6**: 627–630.

Freudenstein J V, Mark WC (2001). Analysis of mitochondrial nad1 bc intron sequences in Orchidaceae: utility and coding of length-change characters. *Syst Bot* **26**: 643–657.

Galtier N, Nabholz B, Glémin S, Hurst GDD (2009). Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Mol Ecol* **18**: 4541–50.

Gamble T, Bauer AM, Greenbaum E, Jackman TR (2008). Evidence for Gondwanan vicariance in an ancient clade of gecko lizards. *J Biogeogr* **35**: 88–104.

Gene Codes Corporation MU Sequencher® version 4.1.4 sequence analysis software.

Geniez F (2001). Variation géographique des lézards du genre *Podarcis* (Reptilia, Sauria, Lacertidae) dans la péninsule Ibérique, l'Afrique du Nord et le sud de la France. *Mémoire présenté pour l'obtention du diplôme l'École Prat des Hautes Etudes, Montpellier*.

Geniez P, Sá-Sousa P, Guillaume C, Cluchier A, Crochet P (2014). Systematics of the *Podarcis hispanicus* complex (Sauria, Lacertidae) III: valid nomina of the western and central Iberian forms. **3794**: 1–51.

Godinho R, Crespo EG, Ferrand N, Harris DJ (2005). Phylogeny and evolution of the green lizards, *Lacerta* spp.(Squamata: Lacertidae) based on mitochondrial and nuclear DNA sequences. *Amphibia-Reptilia* **26**: 271–285.

Goodman MJ, Czelusniak GW, Moore AE, Romero-Herrera, Matsuda G (1979). Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst Biol* **28**: 132–163.

Guigó R, Uchnik I, Smith TF (1996). Reconstruction of Ancient Molecular Phylogeny. *Mol Phylogenet Evol* **6**: 189–213.

Guindon S, Gascuel O (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst Biol* **52**: 696–704.

Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp* **41**: 95–98.

Harris DJ, Carranza S, Arnold EN, Pinho C, Ferrand N (2002). Complex biogeographical distribution of genetic variation within *Podarcis* wall lizards across the Strait of Gibraltar. *J Biogeogr* **29**: 1257–1262.

Harris DJ, Pinho C, Carretero MA, Corti C, Böhme W (2005). Determination of genetic diversity within the insular lizard *Podarcis tiliguerta* using mtDNA sequence data , with a reassessment of the phylogeny of *Podarcis*. *Amphibia-Reptilia* **26**: 401–407.

Harris DJ, Sá-Sousa P (2002). Molecular phylogenetics of Iberian wall lizards (*Podarcis*): is *Podarcis hispanica* a species complex? *Mol Phylogenet Evol* **23**: 75–81.

Harris D, Sá-Sousa P (2001). Species distinction and relationships of the Western Iberian *Podarcis* lizards (Reptilia, Lacertidae) based on morphology and mitochondrial DNA sequences. *Herpetol J* **11**: 129–136.

Heled J, Bryant D, Drummond AJ (2013). Simulating gene trees under the multispecies coalescent and time-dependent migration. *BMC Evol Biol* **13**: 44.

Heled J, Drummond AJ (2010). Bayesian inference of species trees from multilocus data. *Mol Biol Evol* **27**: 570–580.

Hey J, Nielsen R (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci USA* **104**: 2785–90.

Hoban S (2014). An overview of the utility of population simulation software in molecular ecology. *Mol Ecol* **23**: 2383–401.

Hoegg S, Vences M, Brinkmann H, Meyer A (2004). Phylogeny and comparative substitution rates of frogs inferred from sequences of three nuclear genes. *Mol Biol Evol* **21**: 1188–2000.

Holland BR, Benthin S, Lockhart PJ, Moulton V, Huber KT (2008). Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evol Biol* **8**: 202.

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

Huson DH, Klöpper T, Lockhart PJ, Steel MA (2005). Reconstruction of Reticulate Networks from Gene Trees. *Research in Comput Mol Biol. Springer Berlin Heidelb*: 233–249.

Jewett EM, Rosenberg NA (2012). iGLASS: an improvement to the GLASS method for estimating species trees from gene trees. *J Comput Biol* **19**: 293–315.

Joly S, Mclenachan PA, Lockhart PJ (2009). A Statistical Approach for Distinguishing Hybridization and Incomplete Lineage Sorting. *Am Nat* **174**: E54–E70.

Kaliontzopoulou A, Carretero MA, Llorente GA (2012). Morphology of the *Podarcis* wall lizards (Squamata: Lacertidae) from the Iberian Peninsula and North Africa: patterns of variation in a putative cryptic species complex. *Zool J Linn Soc* **164**: 173–193.

Kaliontzopoulou A, Pinho C, Harris DJ, Carretero MA (2011). When cryptic diversity blurs the picture: a cautionary tale from Iberian and North African *Podarcis* wall lizards. *Biol J Linn Soc* **103**: 779–800.

Katoh K, Standley DM (2013). MAFFT: Multiple Sequence Alignment Software Version 7 : Improvements in Performance and Usability Article Fast Track. *Mol Biol Evol* **30**: 772–780.

Kluge AG (1989). A Concern for Evidence and a Phylogenetic Hypothesis of Relationships Among Epicrates (Boidae, serpentes). *Syst Zool* **38**: 7–25.

Kubatko LS, Carstens BC, Knowles LL (2009). STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* **25**: 971–973.

Kubatko LS, Gibbs HL, Bloomquist EW (2011). Inferring species-level phylogenies and taxonomic distinctiveness using multilocus data in sistrurus rattlesnakes. *Syst Biol* **60**: 393–409.

Larget BR, Kotha SK, Dewey CN, Ané C (2010). BUCKy: Gene tree / species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* **26**: 2910–2911.

Leaché AD, Harris RB, Rannala B, Yang Z (2014). The influence of gene flow on species tree estimation: a simulation study. *Syst Biol* **63**: 17–30.

Leaché AD, Rannala B (2011). The accuracy of species tree estimation under simulation : A comparison of methods. *Syst Biol*: syq073.

Li N, Stephens M (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genet Soc Am* **165**: 2213–2233.

Librado P, Rozas J (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**: 1451–1452.

Lima A, Pinho C, Larbes S, Carretero MA, Brito JC, Harris DJ (2009). Relationships of *Podarcis* wall lizards from Algeria based on mtDNA data. *Amphibia-Reptilia* **30**: 483–492.

Liu L (2008). BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* **24**: 2542–2543.

Liu L, Yu L (2011). Estimating species trees from unrooted gene trees. *Syst Biol* **60**: 661–667.

Liu L, Yu L, Edwards S V (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol Biol* **10**: 302.

Liu L, Yu L, Kubatko L, Pearl DK, Edwards S V (2009). Coalescent methods for estimating phylogenetic trees. *Mol Phylogenet Evol* **53**: 320–8.

Liu L, Yu L, Pearl DK, Edwards S V (2009). Estimating species phylogenies using coalescence times among sequences. *Syst Biol* **58**: 468–77.

Löytynoja A, Goldman N (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**: 1632–5.

Maddison WP (1997). Gene trees in species trees. *Syst Biol* **46**: 523–536.

Maia-Carvalho B, Gonçalves H, Ferrand N, Martínez-Solano I (2014). Multilocus assessment of phylogenetic relationships in *Alytes* (Anura, Alytidae). *Mol Phylogenet Evol* **79**: 270–8.

Mallet J (2005). Hybridization as an invasion of the genome. *Trends Ecol Evol* **20**: 229–37.

Mallo D, de Oliveira Martins L, Posada D (2014). Estimation of Species Trees. *eLS*

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P (2010). RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**: 2462–3.

McBreen K, Lockhart PJ (2006). Reconstructing reticulate evolutionary histories of plants. *Trends Plant Sci* **11**: 398–404.

Meade A (2011). BayesTrees v.1.3. *Sch Biol Sci Univ Reading, Reading, United Kingdom*.

Mirarab S, Bayzid MS, Warnow T (2014). Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Syst Biol* **0**: 1–15.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**: i541–i548.

Müller K (2005). SeqState-primer design and sequence statistics for phylogenetic DNA data sets. *Appl Bioinformatics* **4**: 65–69.

Nylander JAA, Wilgenbusch JC, Warren DL, Swofford DL (2008). AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics* **24**: 581–3.

O'Meara BC (2010). New heuristic methods for joint species delimitation and species tree inference. *Syst Biol* **59**: 59–73.

de Oliveira Martins L, Mallo D, Posada D (2014). A Bayesian supertree model for genome-wide species tree reconstruction. *Syst Biol*: syu082.

Pereira C, Couto A, Luís C, Costa D, Mourão S, Pinho C (2013). Twenty-one new sequence markers for population genetics, species delimitation and phylogenetics in wall lizards (*Podarcis* spp.). *BMC Res Notes* **6**: 299.

Pérez-Mellado V (1981). La lagartija de Bocage, *Podarcis bocagei* (Seoane, 1884): Primeros datos sobre su distribución, colorido y ecología. *Amphibia–Reptilia* **3**: 253–268.

Pérez-Mellado V (1981). Nuevos datos sobre la sistemática y distribución de *Podarcis bocagei* (Seoane, 1884) (Sauria, Lacertidae) en la Península Iberica. *Amphibia-Reptilia* **265**: 259–265.

Pérez-Mellado V, Galindo MP (1986). Sistemática de *Podarcis* (Sauria, Lacertidae) ibéricas y norteafricanas mediante técnicas multidimensionales. In: *Série Manuales Universitários, (Spain: Univ. Salamanca)*.

Pinho C, Ferrand N, Harris DJ (2006). Reexamination of the Iberian and North African *Podarcis* (Squamata: Lacertidae) phylogeny based on increased mitochondrial DNA sequencing. *Mol Phylogenet Evol* **38**: 266–73.

Pinho C, Harris DJ, Ferrand N (2003). Genetic polymorphism of 11 allozyme loci in populations of wall lizards (*Podarcis* sp.) from the Iberian Peninsula and North Africa. *Biochem Genet* **41**: 343–59.

Pinho C, Harris DJ, Ferrand N (2007a). Comparing patterns of nuclear and mitochondrial divergence in a cryptic species complex : the case of Iberian and North African wall lizards (*Podarcis*, Lacertidae). *Biol J Linn Soc* **91**: 121–133.

Pinho C, Harris D, Ferrand N (2007b). Contrasting patterns of population subdivision and historical demography in three western Mediterranean lizard species inferred from mitochondrial DNA variation. *Mol Ecol* **16**: 1991–1205.

Pinho C, Harris DJ, Ferrand N (2008). Non-equilibrium estimates of gene flow inferred from nuclear genealogies suggest that Iberian and North African wall lizards (*Podarcis* spp.) are an assemblage of incipient species. *BMC Evol Biol* **8**: 63.

Pinho C, Kaliontzopoulou A, Carretero MA, Harris DJ, Ferrand N (2009). Genetic admixture between the Iberian endemic lizards *Podarcis bocagei* and *Podarcis carbonelli*: evidence for limited natural hybridization and a bimodal hybrid zone. *J Zool Syst Evol Res* **47**: 368–377.

Pinho C, Rocha S, Carvalho BM, Lopes S, Mourão S, Vallinoto M, *et al.* (2010). New primers for the amplification and sequencing of nuclear loci in a taxonomically wide set of reptiles and amphibians. *Conserv Genet Resour* **2**: 181–185.

Pinto C (2013). Hybridization between *Podarcis hispanica* type 1A and *Podarcis bocagei* in areas of syntopy. University of Porto.

Polzin T, Daneshmand SV (2003). On Steiner trees and minimum spanning trees in hypergraphs. *Oper Res Lett* **31**: 12–20.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Purvis A, Gittleman JL, Cowlishaw G, Mace GM (2000). Predicting extinction risk in declining species. *Proc Biol Sci* **267**: 1947–52.

Rambaut A (2014). Figtree version 1.4.2. Available from http://beast.bio.ed.ac.uk/figtree.

Rambaut A, Drummond A (2007). Tracer v1.4. Available from http://beast.bio.ed.ac.uk/tracer.

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

Renoult JP, Geniez P, Bacquet P, Guillaume CP, Crochet PA (2010). Systematics of the *Podarcis hispanicus*-complex (Sauria, Lacertidae) II: the valid name of the north-eastern Spanish form. Zootaxa **2500**: 58-68.

Renoult JP, Geniez P, Bacquet P, Benoit L, Crochet PA (2009). Morphology and nuclear markers reveal extensive mitochondrial introgressions in the Iberian Wall Lizard species complex. *Mol Ecol* **18**: 4298–315.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, *et al.* (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* **61**: 539–42.

Rousset F (2008). genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol Ecol Resour* **8**: 103–106.

Sá-Sousa P (2001). Comparative chorology between *Podarcis bocagei* and *P. carbonellae* (Sauria: Lacertidae) in Portugal. *Rev Española Herpetol* **15**: 85–97.

Sá-Sousa P, Pérez-Mellado V, Martínez-Solano I (2009). *Podarcis carbonelli.* The IUCN Red List of Threatened Species. *Version 20142 <www.iucnredlist.org> Downloaded 24 Sept 2014*.

Sá-Sousa P, Vicente L, Crespo E (2002). Morphological variability of *Podarcis hispanica* (Sauria: Lacertidae) in Portugal. *Amphibia-Reptilia* **23**: 55–69.

Sá-Sousa P, Almeida AP, Rosa H, Vicente L, Crespo EG (2000). Genetic and morphological relationships of the Berlenga wall lizard (*Podarcis bocagei berlengensis:* Lacertidae). *J Zool Syst Evol Res* **38**: 95–102.

Salvi D, Harris DJ, Kaliontzopoulou A, Carretero M a, Pinho C (2013). Persistence across Pleistocene ice ages in Mediterranean and extra-Mediterranean refugia: phylogeographic insights from the common wall lizard. *BMC Evol Biol* **13**: 147.

Sawyer S (1989). Statistical tests for detecting gene conversion. *Mol Biol* **6**: 526–538.

Seo T-K (2008). Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Mol Biol Evol* **25**: 960–71.

Simmons MP, Ochoterena H (2000). Gaps as characters in sequence-based phylogenetic analyses. *Syst Biol* **49**: 369–81.

Simmons MP, Ochoterena H, Carr TG (2001). Incorporation, relative homoplasy, and effect of gap characters in sequence-based phylogenetic analyses. *Syst Biol*: 454–462.

Smith JM, Smith NH (1998). Detecting recombination from gene trees. *Mol Biol Evol* **15**: 590–599.

Srivastava DS, Cadotte MW, MacDonald a AM, Marushia RG, Mirotchnick N (2012). Phylogenetic diversity and the functioning of ecosystems. *Ecol Lett* **15**: 637–48.

Stamatakis A (2014). RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Stephens M, Smith NJ, Donnelly P (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**: 978–89.

Strasburg J, Rieseberg L (2009). How robust are "Isolation with Migration" analyses to violations of the IM model? A simulation study. *Mol Biol Evol* **27**: 297–310.

Szöllősi GJ, Daubin V (2012). Modeling gene family evolution and reconciling phylogenetic discord. *Evol Genomics Humana Press*: 29–51.

Uetz PP, Hosek J (2014). The Reptile Database. <*www.reptile-database.org*>.

Watterson G (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**: 276.

Whelan NV (2011). Species tree inference in the age of genomics. *Trends Evol Biol* **3**: e5.

Wiley EO (1981). Phylogenetics: the theory and practice of phylogenetic systematics. *John Wiley Sons*.

Willis KJ, Birks HJB (2006). What is natural ? The need for a long-term perspective in biodiversity conservation. *Science* **314**: 1261–1265.

Wu Y (2012). Coalescent-based species tree from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution* **66**: 763–775.

Yang Z, Rannala B (2010). Bayesian species delimitation using multilocus sequence data. *Proc Natl Acad Sci USA* **107**: 9264–9.

Yang J, Warnow T (2011). Fast and accurate methods for phylogenomic analyses. *BMC Bioinformatics* **12**: S4.

Young ND, John H (2003). GapCoder automates the use of indel characters in phylogenetic analysis. *BMC Bioinformatics* **6**: 1–6.

Yu Y, Than C, Degnan JH, Nakhleh L (2011). Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Syst Biol* **60**: 138–49.

Zhang D-X, Hewitt GM (1996). Nuclear integrations: challenges for mitochondrial DNA markers. *Tree* **11**: 247–251.

Zhang D-X, Hewitt GM (2003). Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Mol Ecol* **12**: 563–584.

Zimmermann T, Mirarab S, Warnow T (2014). BBCA: Improving the scalability of *BEAST using random binning. *BMC Genomics* **15**: S11.

Appendix

Table A1. List of samples used in this study (mtDNA lineage, sample code and the GenBank accession numbers) is accessible in https://www.dropbox.com/s/wbijjldyarud69j/Appendix_GenBank.pdf?dl=0

Table A2. Samples of *Podarcis* used in this study. mtDNA lineage, sample code, locality information, including province/district and country and latitude/longitude.

mtDNA lineage	Sample Code	Locality	Province/District	Country	Latitude	Longitude
P. atrata	4H	Columbrete Grande	Castellón	Spain	39.898760	0.684864
	10	Columbrete Grande	Castellón	Spain	39.898760	0.684864
	F8	Foradada, Columbretes	Castellón	Spain	39.875071	0.670547
	L3	Lobo, Columbretes	Castellón	Spain	39.875041	0.671748
	M13	Mancolibre, Columbretes	Castellón	Spain	39.895511	0.690119
P. bocagei	3.120	Madalena	Porto	Portugal	41.103983	-8.661383
	3 166	Montesinho	Bragança	Portugal	41.979267	-6.795317
	Sar1	Sarria	Lugo	Spain	42.783333	-7.400000
	3.221	Subportela	Viana do Castelo	Portugal	41.687433	-8.718117
	3.123	Vila Pouca de Aguiar	Vila Real	Portugal	41.445833	-7.672183
	3.56	Gião	Porto	Portugal	41.312950	-8.691633
	3.341	Gerês	Braga	Portugal	41.718333	-8.166667
	Tab1	Taboadela	Ourense	Spain	42.233333	-7.816667
	3 281	Tanes	Astúrias	Spain	43.211167	-5.402533
	DB4292	Torneros de la Valdería	León	Spain	42.225422	-6.239401
	Mad11	Madalena	Porto	Portugal	41.103983	-8.661383
	M6	Montesinho	Bragança	Portugal	41.979267	-6.795317
	bta5	Tanes	Astúrias	Spain	43.211167	-5.402533
	V12	Vairão	Porto	Portugal	41.330383	-8.672400
P. carbonelli	4.97	Cabo Raso	Lisboa	Portugal	38.700000	-9.466667
	4 159	El Acebuche	Huelva	Spain	37.047740	-6.565696
	MC16	Monte Clérigo	Faro	Portugal	37.339880	-8.853810
	SPM9	S. Pedro de Moel	Leiria	Portugal	39.750000	-9.016667
	N292	S.Pedro do Sul	Viseu	Portugal	40.750000	-8.066667
	VR16	Villasrubias	Salamanca	Spain	40.316667	-6.616667
	Av8	Aveiro	Aveiro	Portugal	40.631750	-8.746350
	4.43	Berlengas	Peniche	Portugal	39.415211	-9.508529
	ATL6	Carriço	Leiria	Portugal	39.966667	-8.800000
	Esm18	Esmoriz	Aveiro	Portugal	40.61666	-8.750000
	Alb8	La Alberca	Salamanca	Spain	40.466667	-6.083333
	Albc7	La Alberca	Salamanca	Spain	40.466667	-6.083333
	MC3	Monte Clérigo	Faro	Portugal	37.339880	-8.853810
	PR3	Playa del Rompeculos	Huelva	Spain	37.106177	-6.756996
	SPM2	S. Pedro de Moel	Leiria	Portugal	39.750000	-9.016667

P. erhardii	DB3820	Menites	Andros	Greece	37.826416	24.900352
	DB3819	Thirasia islet	Santorini	Greece	36.447830	25.344021
P. g. lusitanica	5 143	Alvão	Vila Real	Portugal	41.350000	-7.866667
	5 259	Ledesma	Salamanca	Spain	41.091750	-5.997900
	Anc5	Los Ancares	León	Spain	42.669633	-6.726967
	5.23	Moledo	Viana do Castelo	Portugal	41.838567	-8.874067
	5 121	Vale de Rossim, Serra da Estrela	Guarda	Portugal	40.383333	-7.516667
	5 150	Vila de Rua	Viseu	Portugal	40.950000	-7.566667
	Cel1	Celanova	Ourense	Spain	42.150000	-7.966667
	5 180	Gerês	Braga	Portugal	41.718333	-8.166667
	Pen21	Pendilhe	Viseu	Portugal	40.883333	-7.816667
	5 225	Tudera	Zamora	Spain	41.416890	-6.210430
	PG2	Atlantic Islands	ONS	Spain	42.386001	-8.930933
	Anc2	Los Ancares	León	Spain	42.669633	-6.726967
	Mon1	Montesinho	Bragança	Portugal	41.979267	-6.795317
	Mon2	Montesinho	Bragança	Portugal	41.979267	-6.795317
	Mon8	Montesinho	Bragança	Portugal	41.979267	-6.795317
	Pen2	Pendilhe	Viseu	Portugal	40.883333	-7.816667
	Pen8	Pendilhe	Viseu	Portugal	40.883333	-7.816667
	FT12	Tua	Bragança	Portugal	41.218333	-7.368333
	Rua1	Vila de Rua	Viseu	Portugal	40.950000	-7.566667
P. g. guadarramae	5 206	Alba de Tormes	Salamanca	Spain	40.825590	-5.515460
	5 194	Ciudad Rodrigo	Salamanca	Spain	40.592950	-6.536333
	GuaI1	Guadarrama	Madrid	Spain	40.683333	-4.083333
	Oro1	Oropesa	Toledo	Spain	39.919900	-5.174650
	TC1	Torrejón de la Calzada	Madrid	Spain	40.200000	-3.800000
	6 291	Trujillo	Cáceres	Spain	39.460667	-5.881500
	Are1	Arévalo	Ávila	Spain	41.062071	-4.720288
	HLA2	La Alberca	Salamanca	Spain	40.466667	-6.083333
	HLA5	La Alberca	Salamanca	Spain	40.466667	-6.083333
	Vil6	Villacastín	Segóvia	Spain	40.783333	-4.416667
	HAlb1	La alberca	Salamanca	Spain	40.466667	-6.083333
	Trj1	Trujillo	Cáceres	Spain	39.460667	-5.881500
	Vil3	Villacastin	Segóvia	Spain	40.783333	-4.416667
	Vil8	Villacastin	Segóvia	Spain	40.783333	-4.416667
P. hispanica **"Albacete/Murcia"**	ALB1=9.76	Cañada de Provencio	Albacete	Spain	38.518033	-2.353150
	DB1841	El Pardal	Albacete	Spain	38.485309	-2.287438
	DB1878	Montealegre del Castillo	Albacete	Spain	38.830755	-1.339061

	ALB11=9.79	Sierra de la Oliva	Albacete	Spain	38.764883	-0.982583
	9.83	Sierra de la Oliva	Albacete	Spain	38.764883	-0.982583
	DB1286	Sierra de la Pila	Murcia	Spain	38.264363	-1.189820
	ALB2=9.77	Cañada de Provencio	Albacete	Spain	38.518033	-2.353150
	DB3861	Sierra de Callosa del Segura	Alicante	Spain	38.120213	-0.906044
	ALB21=9.89	Sierra de la Oliva	Albacete	Spain	38.764883	-0.982583
	DB1285	Sierra de la Pila	Murcia	Spain	38.264363	-1.189820
***P. hispanica* "Azazga"**	881	Azazga	Tizi Ouzou	Algeria	36.753433	4.424833
	879	Azazga	Tizi Ouzou	Algeria	36.753433	4.424833
***P. hispanica* "Batna"**	DjeA31	Djebel Aurés	Batna	Argelia	35.350117	6.621867
	DjeA34	Djebel Aurés	Batna	Argelia	35.350117	6.621867
	Ham1	Hamla	Batna	Algeria	35.579700	6.076250
	Ham2	Hamla	Batna	Algeria	35.579700	6.076250
***P. hispanica* "Galera"**	9.73	Caravaca de la Cruz	Murcia	Spain	38.107450	-1.859050
	9.55	Cartagena	Murcia	Spain	37.603283	-1.007517
	9.19	Galera	Granada	Spain	37.741783	-2.549317
	DB2961	Castril river source	Granada	Spain	37.908395	-2.749222
	9.10	Sierra de Espuña	Murcia	Spain	37.819850	-1.582683
	9.11	Velez Blanco	Almería	Spain	37.681633	-2.096217
	9.67	Caravaca de la Cruz	Murcia	Spain	38.107450	-1.859050
	9.48	Cartagena	Murcia	Spain	37.603283	-1.007517
	9.18	Galera	Granada	Spain	37.741783	-2.549317
	9.7	Sierra de Espuña	Murcia	Spain	37.819850	-1.582683
	Gal2BF+BF8	Galera	Granada	Spain	37.741783	-2.549317
	Gal5x	Galera	Granada	Spain	37.741783	-2.549317
	Gal7x	Galera	Granada	Spain	37.741783	-2.549317
	Gal4	Galera	Granada	Spain	37.741783	-2.549317
	Gal6x	Galera	Granada	Spain	37.741783	-2.549317
	Gal8x	Galera	Granada	Spain	37.741783	-2.549317
	Gal1	Galera	Granada	Spain	37.741783	-2.549317
	Gal1x	Galera	Granada	Spain	37.741783	-2.549317
	Gal3x	Galera	Granada	Spain	37.741783	-2.549317
	BEV7353	Puebla de Don Fadrique	Granada	Spain	37.912800	-2.400100
***P. hispanica* "Jebel Sirwah"**	7.82	Jebel Siroua	Taroudannt	Morocco	30.746966	-7.609283
	7.68	Jebel Siroua	Taroudannt	Morocco	30.746966	-7.609283
	7.78	Jebel Siroua	Taroudannt	Morocco	30.746966	-7.609283
	DB1663	road to Jebel Siroua	Taroudannt	Morocco	30.788062	-7.593582

	DB11031	Tizi-n'-Melloul	Ouarzazate	Morocco	30.808100	-7.583670
	PH70	Jebel Siroua	Taroudannt	Morocco	30.704069	-7.621469
	PH28	Jebel Siroua	Taroudannt	Morocco	30.746966	-7.609283
	PH60	Jebel Siroua	Taroudannt	Morocco	30.746966	-7.609283
	JS1	Jebel Siroua	Taroudannt	Morocco	30.746966	-7.609283
	JS2	Jebel Siroua	Taroudannt	Morocco	30.746966	-7.609283
	JS3	Jebel Siroua	Taroudannt	Morocco	30.746966	-7.609283
	JS6	Jebel Siroua	Taroudannt	Morocco	30.746966	-7.609283
	PH184	Jebel Siroua	Taroudannt	Morocco	30.746966	-7.609283
	PH186	Jebel Siroua	Taroudannt	Morocco	30.746966	-7.609283
	JS8	Jebel Siroua	Taroudannt	Morocco	30.746966	-7.609283
***P. hispanica* sensu stricto**	9.22	Boniche	Cuenca	Spain	39.985383	-1.628717
	Burj692	Burjasot	Valencia	Spain	39.516667	-0.416667
	DB8644	Cuenca	Cuenca	Spain	40.066667	-2.133333
	Mot1	Motilla del Palancar	Cuenca	Spain	39.566667	-1.883333
	9.32	Xátiva	Valencia	Spain	38.985750	-0.520017
	DB8646	Bunyol	Valencia	Spain	39.418258	-0.790769
	G1	Ciudad Encantada	Cuenca	Spain	40.207397	-2.005227
	B2	La Casella, Alzira	Valencia	Spain	39.125110	-0.386231
	B4	La Casella, Alzira	Valencia	Spain	39.125110	-0.386231
	Val1	Valencia	Valencia	Spain	39.475000	-0.364000
	Cue2	Cuenca	Cuenca	Spain	40.066667	-2.133333
	Cue1	Cuenca	Cuenca	Spain	40.066667	-2.133333
	Mot1	Motilla del Palancar	Cuenca	Spain	39.566667	-1.883333
	And8	Puebla de Don Fadrique	Granada	Spain	37.912800	-2.400100
	BEV7337	Sta. Maria de Nieva	Condorcanqui	Spain	37.530000	-2.004000
	DB1879	Barranco de Guadalentín	Jaén	Spain	37.871042	-2.876590
	9.8	Castillo de la Calahora	Granada	Spain	37.183467	-3.065133
	10.46	Cazorla	Jaén	Spain	37.891867	-2.864717
	10.29	Puerto de la Ragua, Sierra Nevada	Granada	Spain	37.112467	-3.033117
	DB1853	Rio Madera	Jaén	Spain	38.245624	-2.628203
	DB1791	500m from los negros camping	Jaén	Spain	38.271465	-2.606100
	DB1899	Barranco de Guadalentín	Jaén	Spain	37.871042	-2.876590
	DB1748	Guadalquivir river source	Jaén	Spain	37.839145	-2.974465
	DB3053	Pico Cabanyas	Jaén	Spain	37.814097	-2.954600
	DB3858	Sierra Espuña-Zona Pozos de Nieve	Murcia	Spain	37.869971	-1.571928
	Pod12	Granada	Granada	Spain	37.176377	-3.592951
	SN10	Puerto de la Ragua, Sierra Nevada	Granada	Spain	37.113505	-3.030699

	SN11	Puerto de la Ragua, Sierra Nevada	Granada	Spain	37.113506	-3.030700
	SN2	Puerto de la Ragua, Sierra Nevada	Granada	Spain	37.113507	-3.030701
***P. hispanica* "Tunisia/Algeria"**	8 536	Ain Draham	Jendouba	Tunisia	36.772975	8.685613
	8 553	Cap Negro	Jendouba	Tunisia	37.068566	9.046450
	Edo33	Edough	Annaba	Algeria	36.883333	7.616667
	8 480	Feidja NP	Jendouba	Tunisia	36.504433	8.313717
	LK5	Le Kef	Kef	Tunisia	36.184680	8.710300
	8 521	Ain Draham	Jendouba	Tunisia	36.772975	8.685613
	8 557	Cap Negro	Jendouba	Tunisia	37.068566	9.046450
	Elk32	El Kala	El Tarf	Algeria	36.834439	8.415283
	DB1625	Jbel Goraa	Teboursouk	Tunisia	36.490671	9.151754
	DB1595	Jbel Goraa	Teboursouk	Tunisia	36.490671	9.151754
	LK6	Le Kef	Kef	Tunisia	36.184680	8.710300
	OK11	Oued Kébir	Skikda	Tunisia	36.777000	8.695930
	OK1	Oued Kébir	Skikda	Tunisia	36.777000	8.695930
	OK8	Oued Kébir	Skikda	Tunisia	36.777000	8.695930
P. liolepis	AP1	Alcolea del Pinar	Guadalajara	Spain	41.033333	-2.466667
	Bur4	Burgos	Burgos	Spain	42.350000	-3.700000
	DB8613	Castrillo de la Vega	Burgos	Spain	41.651982	-3.781494
	Get14	Getaria	Guipúzcoa	Spain	43.300000	-2.200000
	DB1731	Monasterio de Moreruela	Zamora	Spain	41.811860	-5.778050
	DB8605	Sopeira	Huesca	Spain	42.309327	0.740439
	DB1732	Torredembarra	Tarragona	Spain	41.151750	1.431183
	DB1716	Aiguamolls del Emporda	Girona	Spain	42.222000	3.092340
	1.15	Calomarde	Aragón	Spain	40.371917	-1.577733
	DB1762	Les Solans	Barcelona	Spain	41.708984	1.989240
	DB8560	Medinaceli	Soria	Spain	41.162250	-2.417767
	Barc4	Barcelona	Barcelona	Spain	41.422793	2.139054
	Vall1	Vall d´Alynià	Lerida	Spain	42.182800	1.421950
	Barc5	Barcelona	Barcelona	Spain	41.422793	2.139054
	Bur2	Burgos	Burgos	Spain	42.350000	-3.700000
	Get1	Getaria	Guipúzcoa	Spain	43.300000	-2.200000
	Med1	Medinaceli	Soria	Spain	41.162250	-2.417767
	PhT1	Tarragona	Tarragona	Spain	41.116667	1.250000
P. sicula	DB5936	Cervia	Ravenna	Italy	44.263549	12.347682
	DB9103	Pizzo	Vibo Valentia	Italy	38.745858	16.179450
	DB1713	Sardinia	Sardinia	Italy	40.120875	9.012893
	DB9105	Tropea	Vibo Valentia	Italy	38.682415	15,917993

	DB771	Vulcano Island	Vulcano Island	Italy	38.406550	14,962554
P. tiliguerta	PT5	Between Col de Lavezzo and Monetta	Corsica N	France	42.658415	9.117760
	PT9	Between Corte and Aleria, close to Altiani	Corsica E	France	42.236667	9.290586
	PT6	Close to S. Quilico	Corsica NNW	France	42.469307	9.109948
	PT11	Solenzarea	Corsica E	France	41.857072	9.398407
P. taurica	DB3805	Strofilia	Eubeia	Greece	38.815370	23.409438
	DB11623	Zakynthos	Zakynthos	Greece	37.792020	20.789296
	DB11622	Zakynthos island, near lake Keri	Zakynthos	Greece	37.680000	20.820000
P. virescens	6.59	Beja	Beja	Portugal	38.016667	-7.866667
	6.96	Casar de Cáceres	Cáceres	Spain	39.566667	-6.416667
	MR16	Monte Real	Leiria	Portugal	39.850000	-8.866667
	DB1890	Rio Linares- Riba de Saelices	Guadalajara	Spain	40.940402	-2.292685
	DB2866	Sierra de Segura, 5km W of Embalse del Tranco	Jaén	Spain	38.161933	-2.847769
	DB2846	SW of Embalse del Tranco	Jaén	Spain	38.082325	-2.817273
	6.34	Virgen de la Cabeza, Andujar	Jaén	Spain	38.181833	-4.037983
	DB2911	Area Recreativa de Gil Cobo	Jaén	Spain	38.080451	-2.899910
	DB1779	Area recreativa de los estrechos	Toledo	Spain	38.323485	-2.633985
	6 153	Arroyo Brezoso	Castilla la Mancha	Spain	39.360417	-4.358367
	CR1	Castaño de Robledo	Huelva	Spain	37.893667	-6.705933
	DB1736	Cueva del Santillo	Jaén	Spain	37.924807	-2.952072
	6 128	Parque Natural de Cornalvo	Badajoz	Spain	39.019867	-6.176150
	6 311	Riopar el Viejo	Albacete	Spain	38.499867	-2.418967
	VC3	Villanueva de Córdoba	Córdoba	Spain	38.315350	-4.625317
	And10	Benatae	Jaén	Spain	38.350000	-2.650000
	CR1	Castaño de Robledo	Huelva	Spain	37.893667	-6.705933
	CV1	Castelo de Vide	Portalegre	Portugal	39.416667	-7.450000
	EV4	Évora	Évora	Portugal	38.566667	-7.900000
	MadA	Madalena	Porto	Portugal	41.103983	-8.661383
	MadB	Madalena	Porto	Portugal	41.103983	-8.661383
	SM1	S. Mamede	Portalegre	Portugal	39.467967	-7.634767
	And9	Saucedilla	Cáceres	Spain	39.851721	-5.676911
P. vaucheri **"Morocco/Algeria"**	Dju939	Djurjura	Tizi Ouzou	Algeria	36.471050	3.996633
	DB1449	Ceuta	Ceuta	Spain	35.888270	-5.316160
	7 334	Ketama	Al-Hoceima	Morocco	34.878233	-4.610867
	7 300	Midelt	Meknès-Tafilalet	Morocco	32.682367	-4.742650
	7 409	Tislit Lake	Errachidia	Morocco	32.196400	-5.642933
	DB1047	Tizi-n-Tleta	Taroudannt	Morocco	30.780900	-7.643540
	7 385	Debdou	Oujda	Morocco	33.872467	-3.038783

	DB1140	Imouzer-des-Glaoua	Tizi-n-Titchka	Morocco	31.306972	-7.362605
	DB1587	Imouzzer Kandar to Annoceur	Fès-Boulemane	Morocco	33.625820	-4.896278
	DB76	Lac Iseli	Meknès-Tafilalet	Morocco	32.216467	-5.549717
	7 137	Mischliffen	Meknès-Tafilalet	Morocco	33.405433	-5.103317
	7.86	N Oukaimeden	Marrakech	Morocco	31.203550	-7.861720
	7.26	Taza	Taza	Morocco	34.221190	-4.015857
	Bt6	Bab Taza	Taounate	Morocco	35.022583	-5.204483
	Mis3	Mischliffen	Meknès-Tafilalet	Morocco	33.405433	-5.103317
	Ouk7	Oukaimeden	Marrakech	Morocco	31.203550	-7.861717
P. vaucheri **"Southern-central Spain"**	AR5	Alcalá la Real	Jaén	Spain	37.461400	-3.928583
	AR11	Alcalá la Real	Jaén	Spain	37.461400	-3.928583
	J16	Jaen Ciudad	Jaén	Spain	37.786350	-3.775117
	DB1873	Linares	Jaén	Spain	38.093620	-3.635844
	DB1754	Pradollano, Sierra Nevada	Granada	Spain	37.071684	-3.388135
	AR4=8.83	Alcalá la Real	Jaén	Spain	37.461400	-3.928583
	DB1208	Between Benalua de las Villas e Iznalloz	Jaén	Spain	37.456823	-3.614522
	J2	Jaen Ciudad	Jaén	Spain	37.786350	-3.775117
P. vaucheri **"Southern-Spain"**	DB1811	Bonanza	Cádiz	Spain	36.800000	-6.333333
	COR2	Cordoba Ciudad	Córdoba	Spain	37.873750	-4.786500
	DB1874	Linares	Jaén	Spain	38.093620	-3.635844
	8.26	Matalascañas	Huelva	Spain	37.253617	-6.561182
	PP1	Peñarroya/Pueblonuevo	Córdoba	Spain	38.298050	-5.264933
	DB1251	Castro del Rio	Córdoba	Spain	37.683198	-4.474988
	DB1380	Granada	Granada	Spain	37.176377	-3.592951
	8.22	La Barrosa	Cádiz	Spain	36.373300	-6.186950
	DB2863	Linares	Jaén	Spain	38.093620	-3.635844
	DB2869	Linares	Jaén	Spain	38.093620	-3.635844
	DB446	Parque Nacional Los Alcornocales	Cádiz	Spain	36.350000	-5.533333
	DB3871	Sanlucar La Mayor	Sevilla	Spain	37.386906	-6.203470
	LB2	La Barrosa	Cádiz	Spain	36.373300	-6.186950
	LB4	La Barrosa	Cádiz	Spain	36.373300	-6.186950
	LB5	La Barrosa	Cádiz	Spain	36.373300	-6.186950
	LB9	La Barrosa	Cádiz	Spain	36.373300	-6.186950

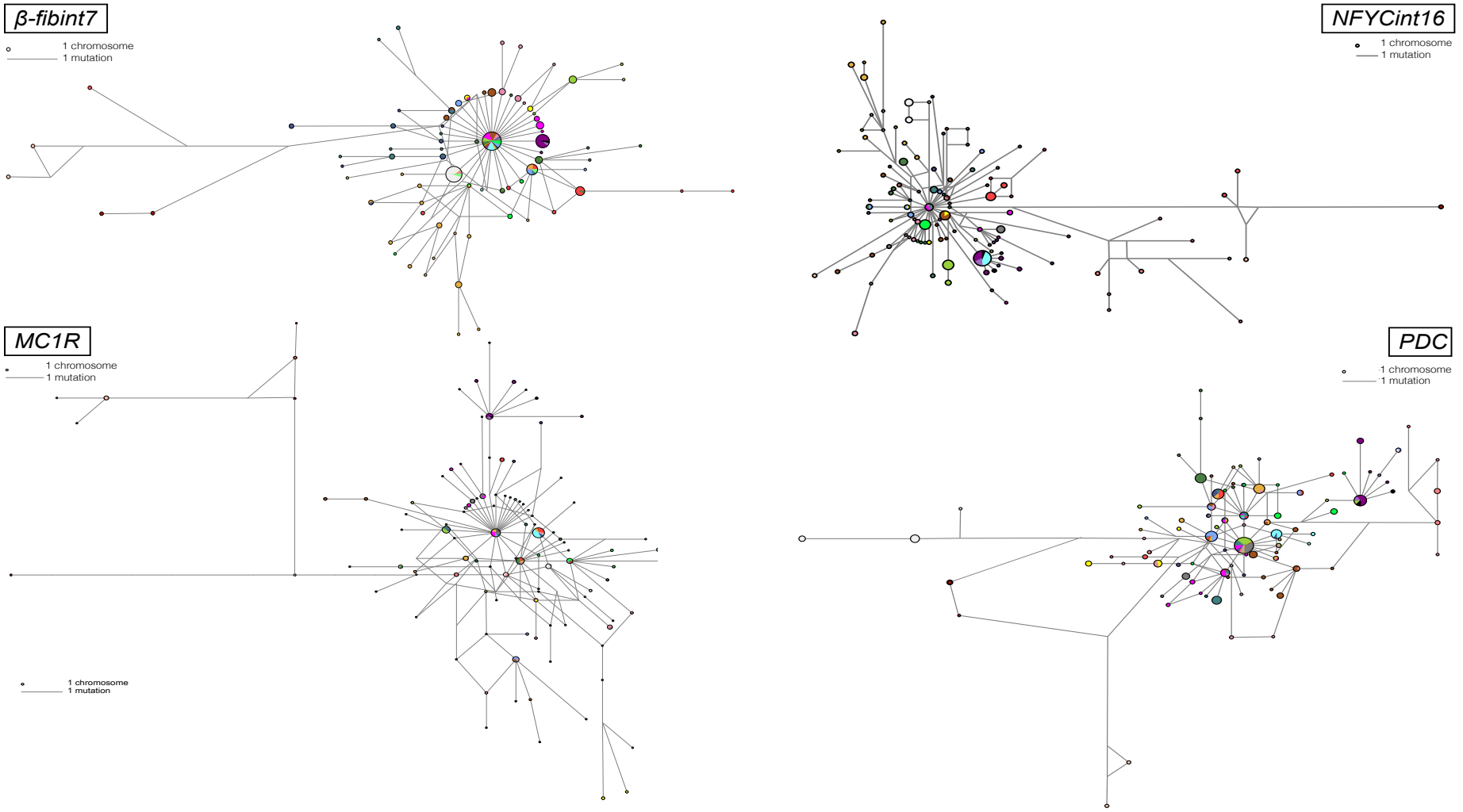


Figure A1. Haplotype networks for twenty-six nuclear loci analysed in this study. Circle area corresponds to haplotype frequency, and distinct colours to different mtDNA lineages/species (17) and are the same used in Figure 2.1 Branch lengths are proportional to the distance between haplotypes.

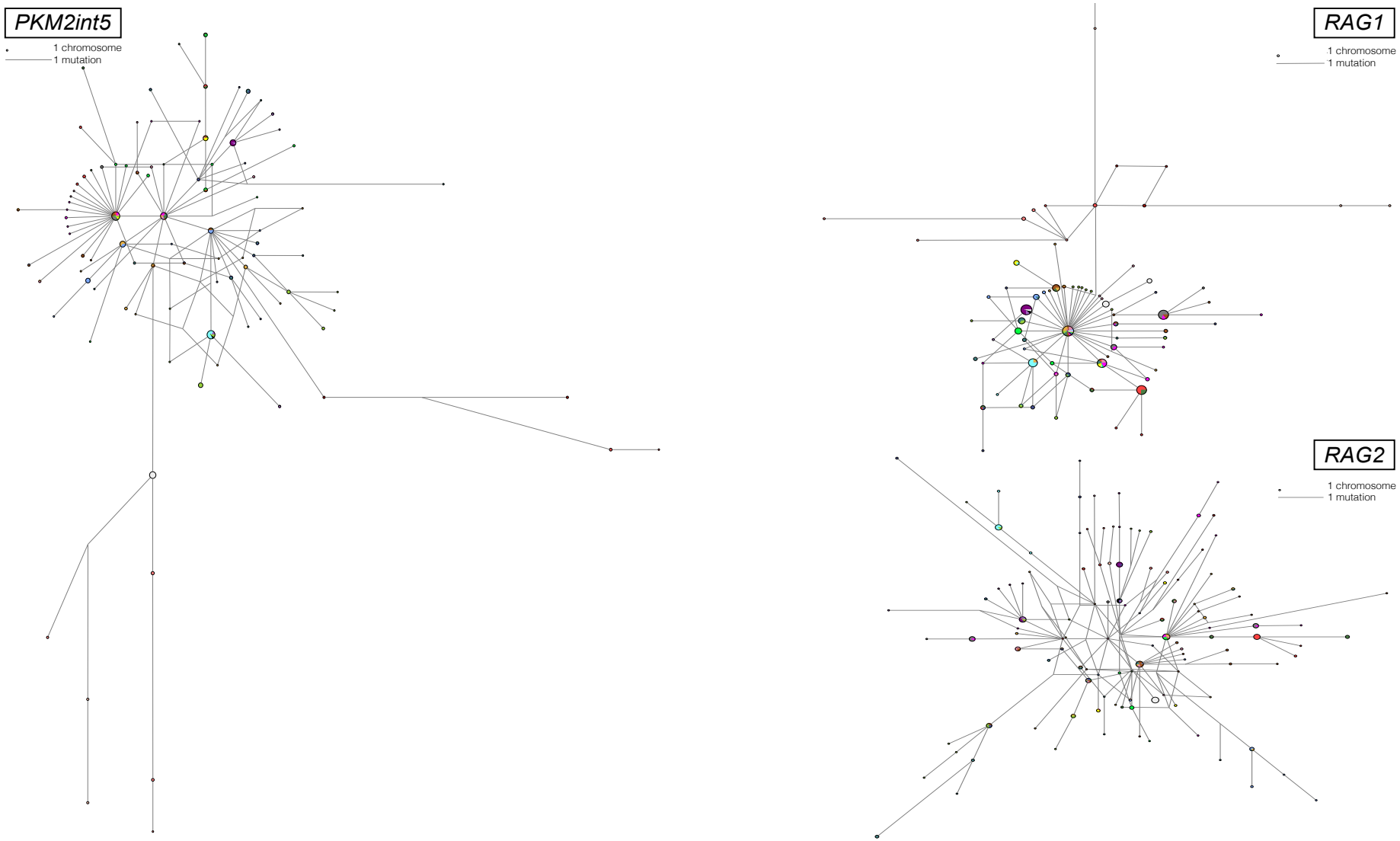


Figure A1. (Continuation)

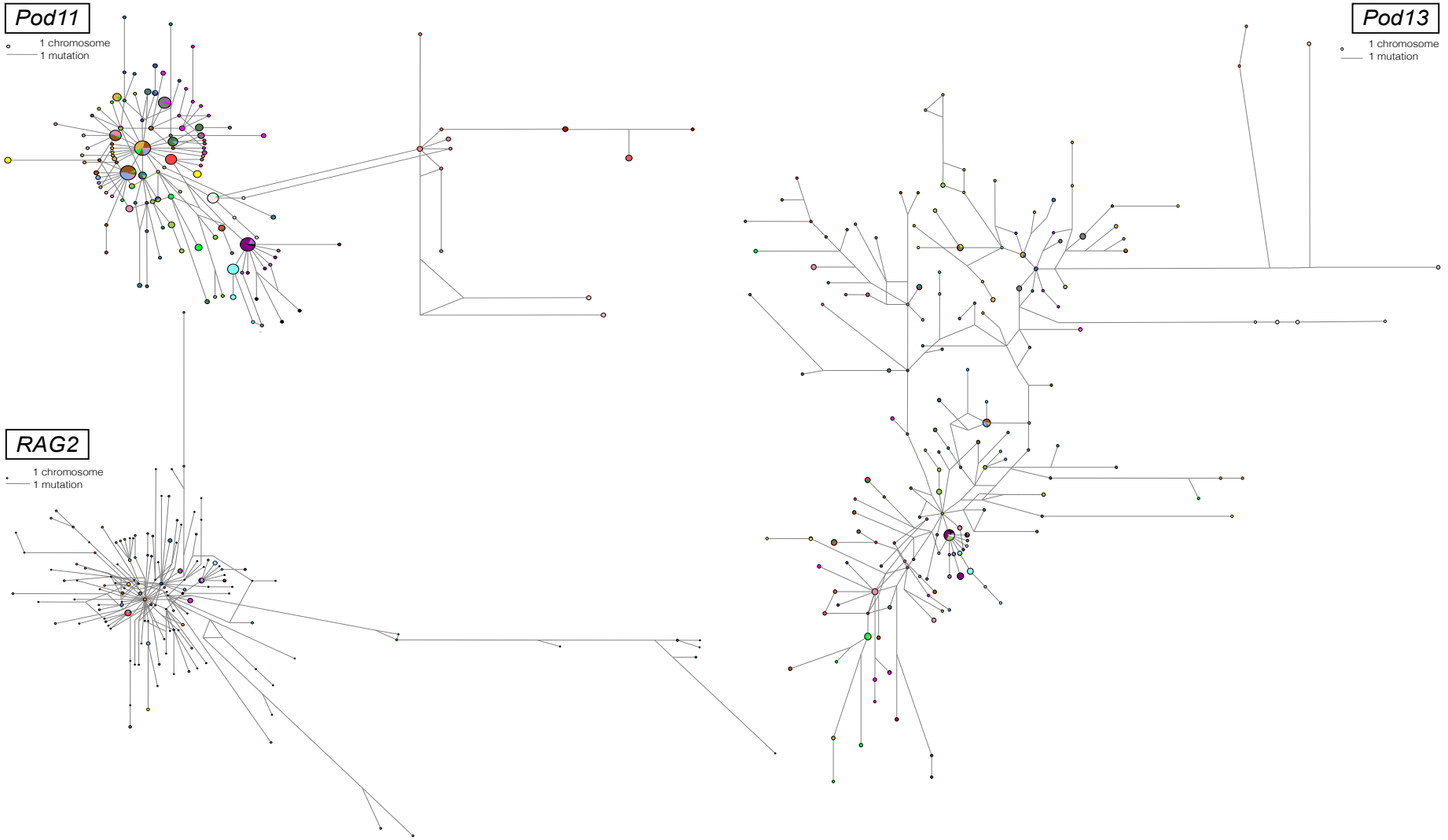
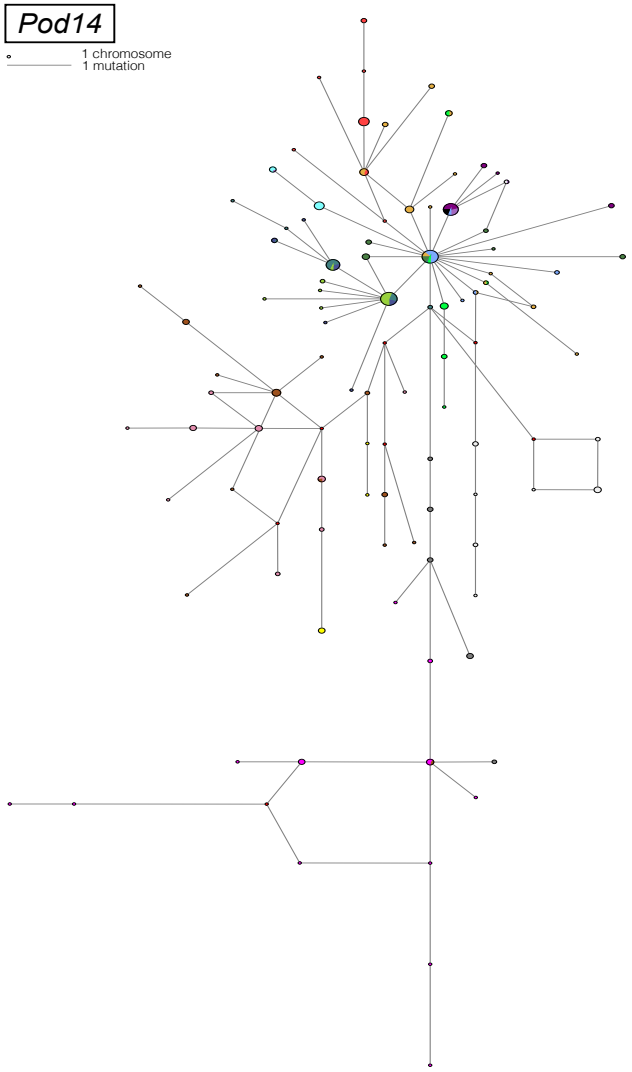


Figure A1. (Continuation)



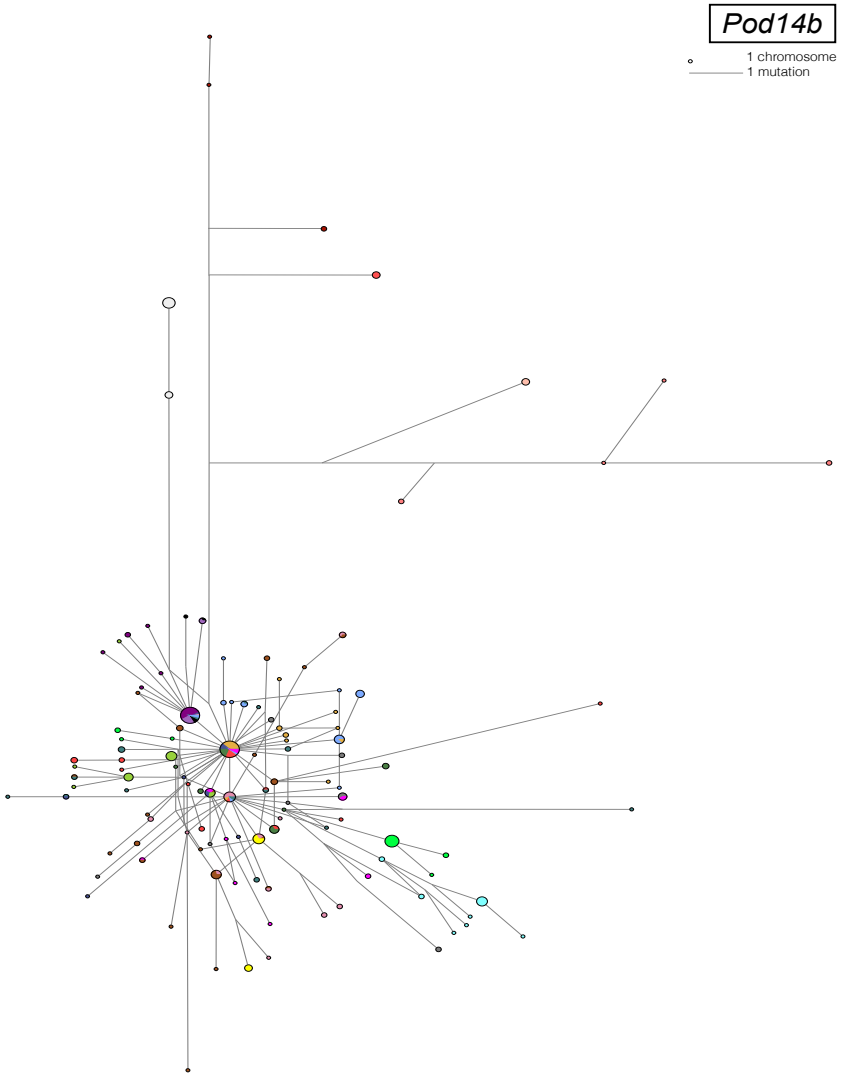


Figure A1. (Continuation)

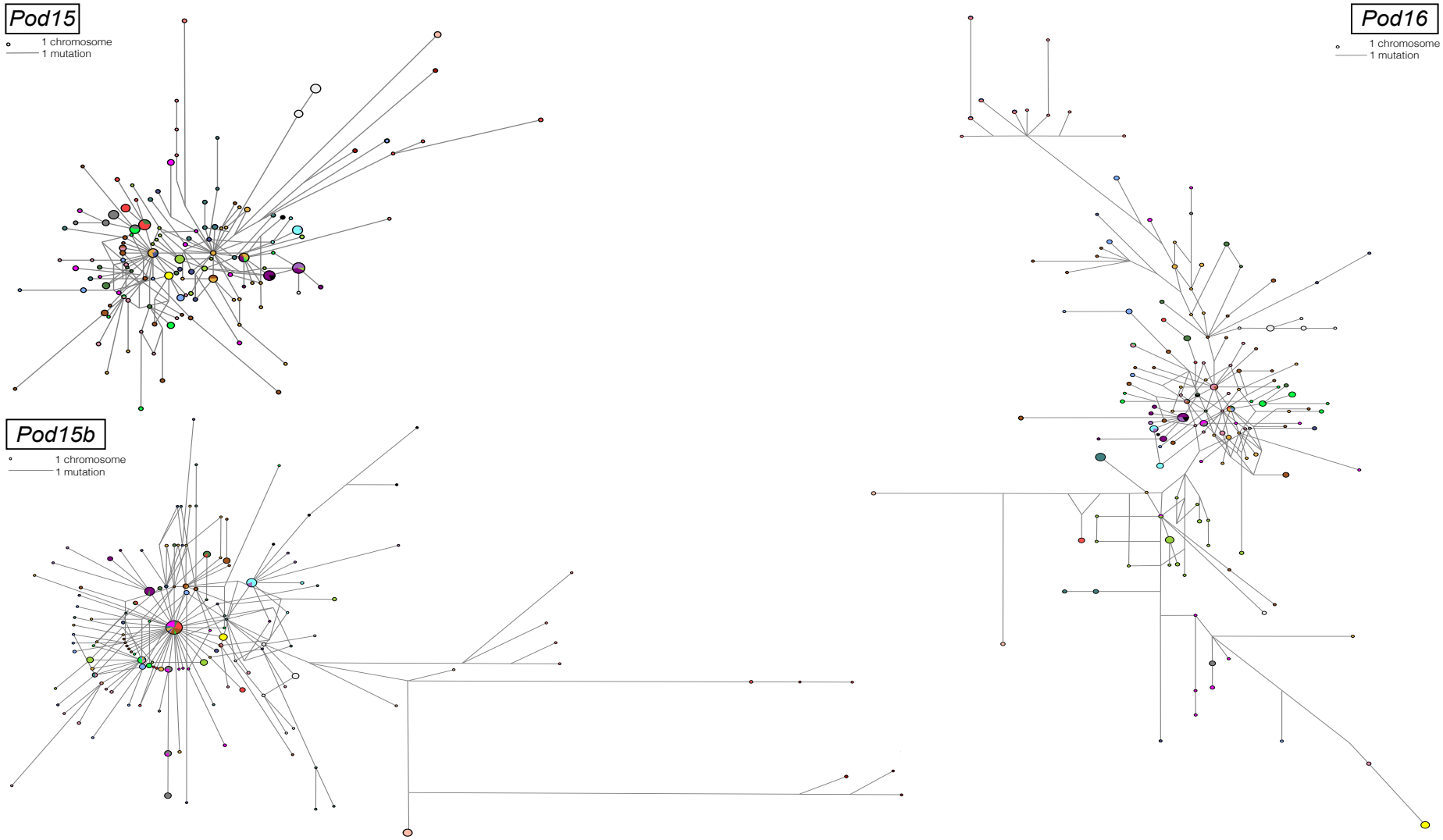


Figure A1. (Continuation)

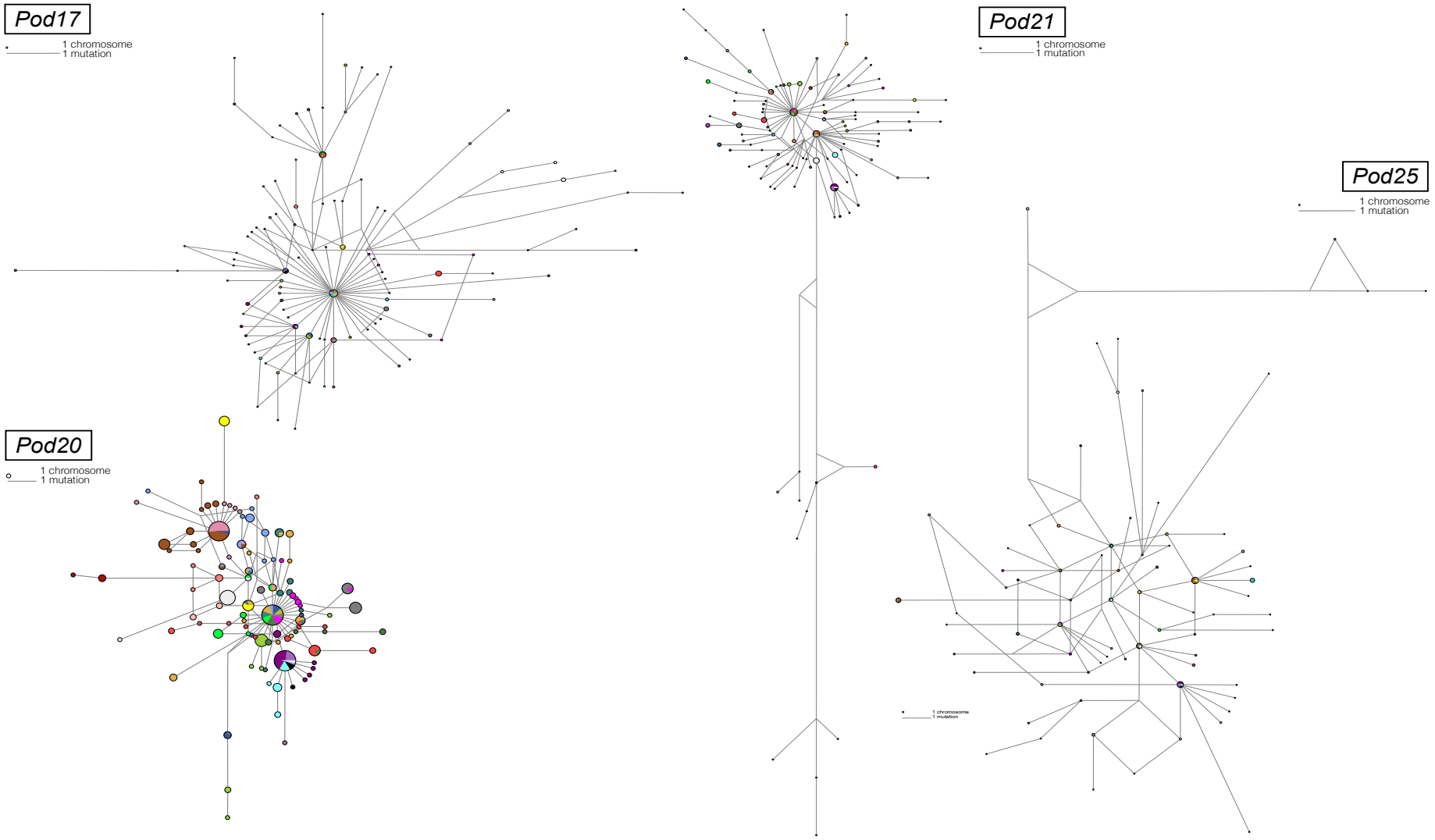
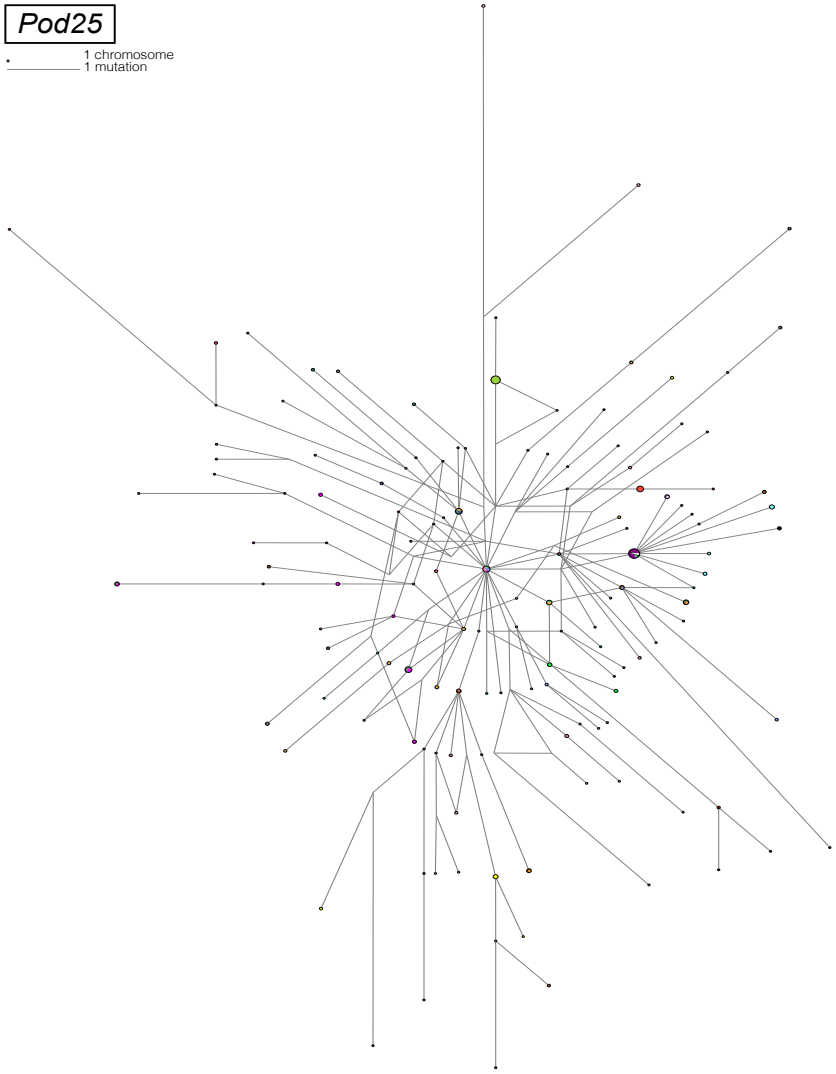


Figure A1. (Continuation)



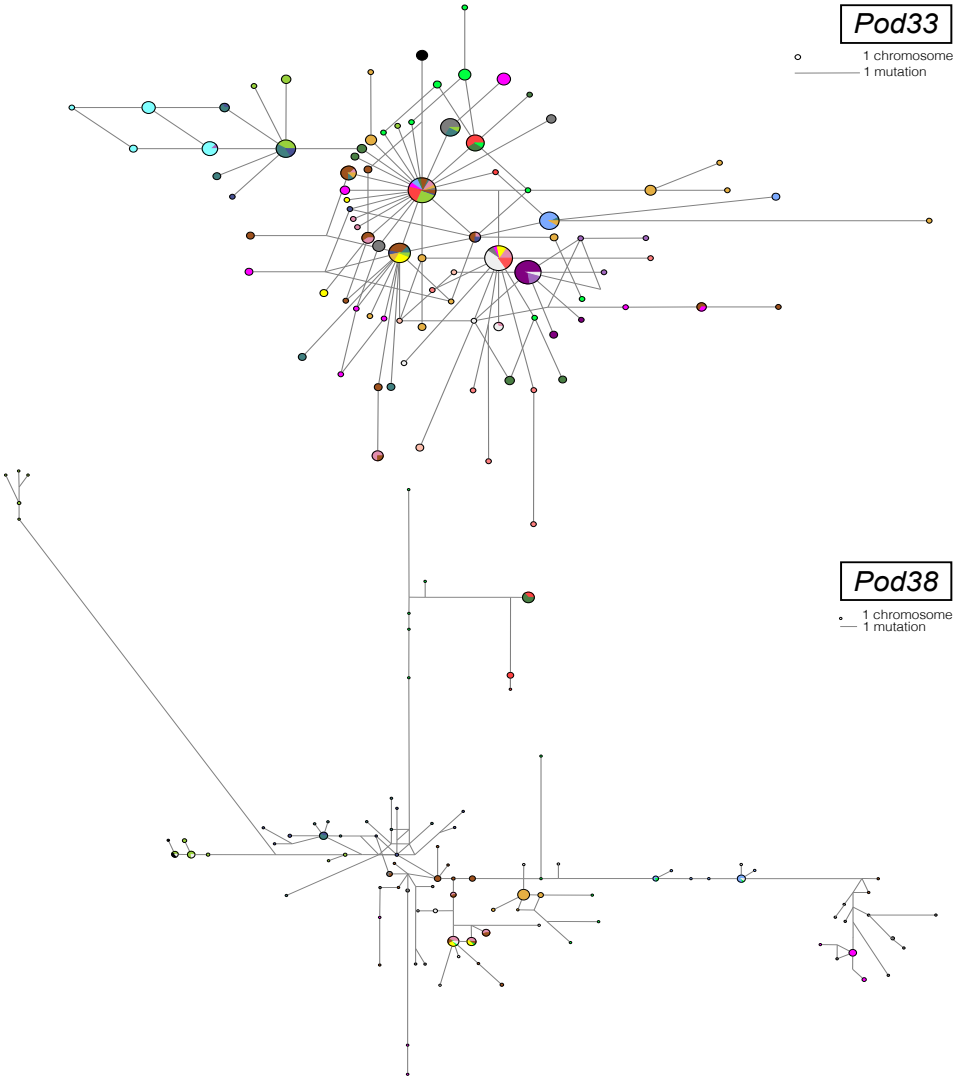


Figure A1. (Continuation)

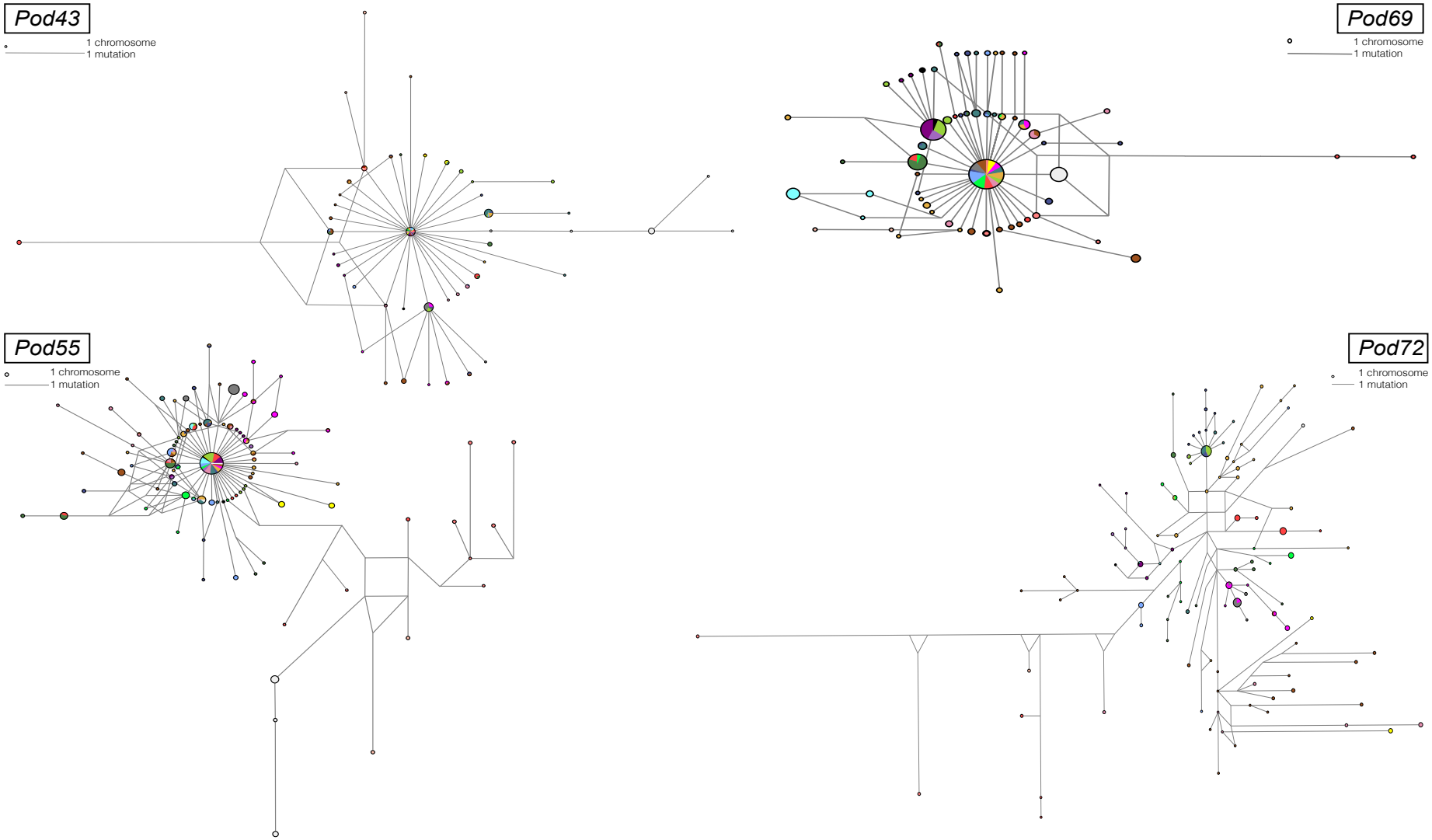


Figure A1. (Continuation

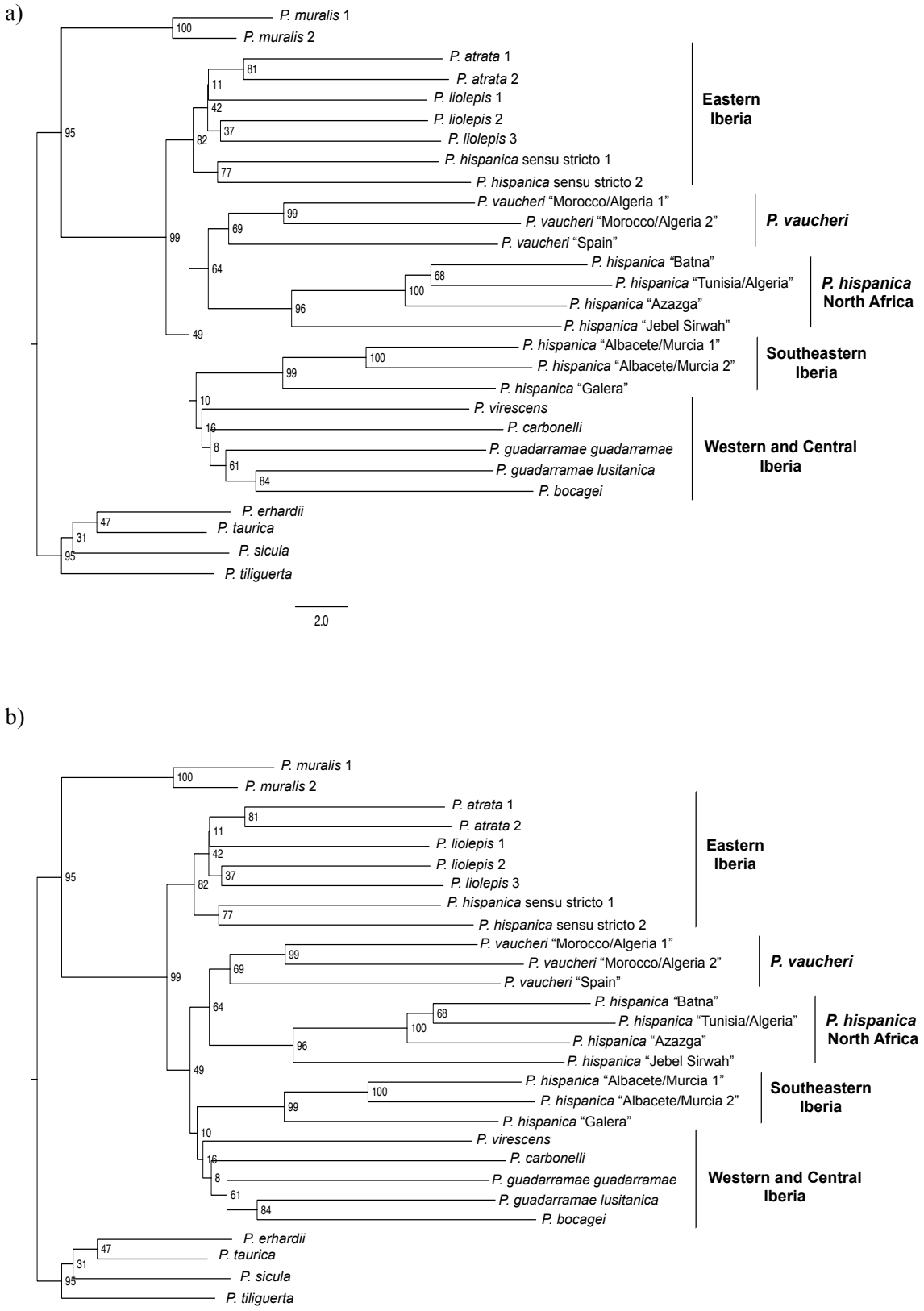


Figure A2. Phylogeny of Iberian and North African *Podarcis* species as estimated with the NJst method using each gene ML tree topology from the "full" dataset; a) "best" NJ species-tree; b) NJst consensus consensus (branch lengths are transformed and do not reflect any amount of evolution). Trees were inferred unrooted and rooted for visualization in the branch leading to *P. erhardii*, *P. taurica*, *P. sicula* and *P. tiliguerta*. The numbers on the nodes indicate multilocus bootstrap support values for branches calculated following Seo, 2008.

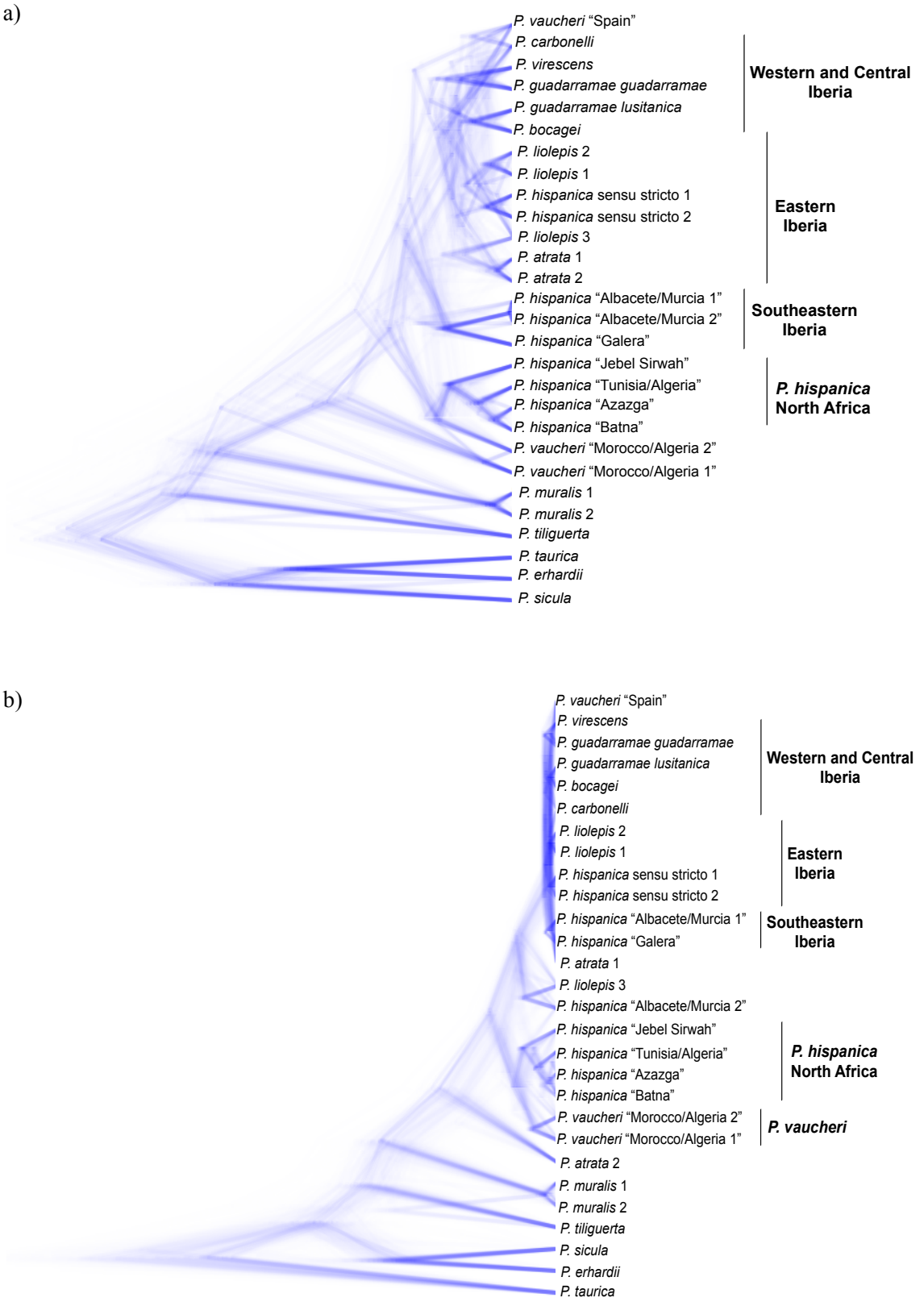


Figure A3. Posterior density of consensus species-trees (cloudogram) from *BEAST analyses for the Iberian and North African *Podarcis* species for 21 loci, for a chain length/tree prior of a) 605M/Yule; b) 168M/coalescent. Consensus for every topology with branch length calculated as the average of the branch length for all trees with the same topology.

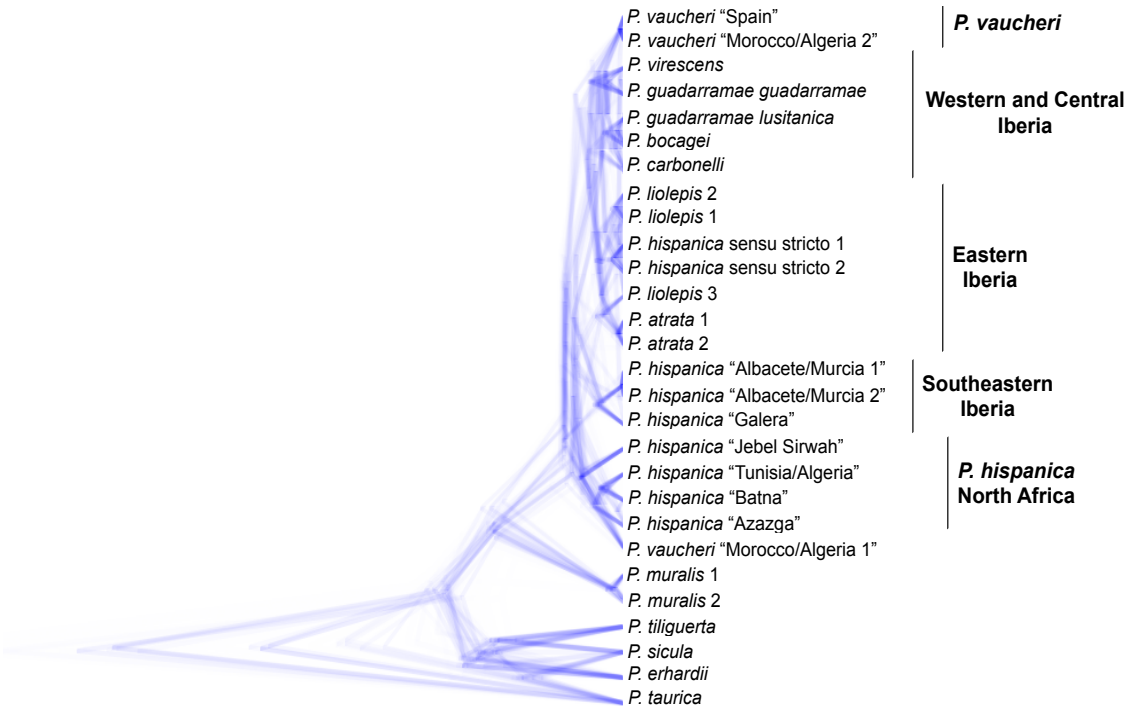


Figure A4. Posterior density of consensus species-trees (cloudogram) from *BEAST analyses for the Iberian and North African *Podarcis* species for all loci, for a chain length/tree prior of 341M/Yule. Consensus for every topology with branch length calculated as the average of the branch length for all trees with the same topology.