Integrating genetic variation and species distribution models to better understand species' evolutionary history and improve redistribution projections under climate changes

Pedro Victor Poli da Silva Pestre

Thèse de Doctorat

Mention Sciences Ecologiques
Spécialité Ecologie, Evolution, Biodiversité

présentée à *l'Ecole Doctorale en Sciences Technologie et Santé (ED 585)*

de l'Université de Picardie Jules Verne

par

Pedro POLI

pour obtenir le grade de Docteur de l'Université de Picardie Jules Verne

Combiner la variation génétique et les modèles de distribution des espèces pour mieux comprendre leur histoire évolutive et améliorer les projections de redistribution en contexte de réchauffement climatique

Soutenue le 29 Janvier 2021 après avis des rapporteurs, devant le jury d'examen :

Didier CASANE, Professeur, Université de Paris	**Président**
Laurence DESPRES, Professeur, Université Grenoble Alpes	**Rapporteur**
Antoine GUISAN, Professeur, Université de Lausanne, Suisse	**Rapporteur**
Karine HUBER, Chargée de Recherche, INRAE	**Examinateur**
Ronan MARREC, Maître de Conférences, UPJV	**Examinateur**
Annie GUILLER, Professeur, UPJV	**Directeur de thèse**
Jonathan LENOIR, Chargé de Recherche, CNRS	**Co-encadrant**

Acknowledgements

I am deeply grateful to my committee members, **Laurence Després**, **Antoine Guisan**, **Didier Casane**, **Karine Huber,** and **Ronan Marrec** for agreeing to read the manuscript and to participate in the defense of this thesis.

Among the many people that helped me along this journey, my PhD supervisor had the greatest importance on the fulfilment of this project: **Annie Guiller** and **Jonathan Lenoir**, I am very grateful to have worked with you.

Annie Guiller supported me in all stages of this work. I cannot begin to express how grateful I am for her guidance, patience, and commitment not only to the thesis itself but also to my personal development as an evolutionary ecologist. She always gave me constant encouragement and advice despite her busy agenda. I could never thank Annie enough for all she invested in this thesis.

My PhD co-supervisor **Jonathan Lenoir** is the second pillar of this thesis. His insights in species distribution modelling, statistics, and scientific writing were fundamental to this dissertation. Not to forget his keen eyes to keep-up with my many writing mistakes. He should be an editor in an important scientific review. Oh, wait! He already is.

The completion of this dissertation would not have been possible without the invaluable support of Professor **Guillaume Decocq**. From putting me in touch with the great FLEUR research network to helping with all kinds of difficulties, his contribution was invaluable.

I would like to express my deepest appreciation to the PhD monitoring committee, Professor **Guillaume Decocq**, Research Director **Eric Petit**, and Research Director **Olivier Plantard** for their insights and advice.

Monsieur **Olivier Plantard** deserves a paragraph all to himself. This research could have never been possible without his support and the free access he kindly conceded to his personal tick collection. I hope we can continue to collaborate in the future!

I would like to extend my sincere thanks to all the staff of the Centre Régional des Ressources en Biologie Moléculaire, le cher CRRBM. I am very grateful for the technical support of **Gaëlle Mongelard** and **Stéphanie Guénin**, who devoted time and energy to help with the laboratory work. I'd also like to acknowledge the support of **Hervé Demailly** and **Christophe Pineau**, and the commitment of Mr. **Laurent Gutierrez**.

Stéphane Dreano is another person whose inestimable help should not be forgotten since it traces back to my under graduation studies. My deepest thanks for your samples, but most of all for all the late nights of migration and sequencing.

I am extremely grateful to my PhD colleague and dear friend **Gauthier Delvoye**. His mathematical and statistical insights helped me a lot to understand the many methods for data analysis in population genetics. His friendship also helped me to discovering that some people are not good at board games, they are just incredibly lucky.

This work would not have been possible without the invaluable contribution from Dr. **Donatella Magri** and Dr. Remy Petit, who shared their data and knowledge on the past distribution of *F. Sylvatica.*

I'd like to acknowledge the assistance of the many colleagues that helped with the 'treasure hunt' of *Oxalis* and *Geum* samples: **Adele Mennerat**, **Alexandra Ha'nova**, **Alexandre Fruleux**, **Eddie McCormack, Jaan Lira**, **Jan Plue**, **Jenny Hagenblad**, **Jonas Lembrechts**, **Kristoffer Highlander**, **Magni Olsen Kyrkjeeide**, **Martin Dieckman**, **Martin Večeřa**, **Nina Roth**, **Pekka O. Niittynen**, **Pierre-Marie Leffénaff**, **Ronja Wedegaertner**, **Serena A. Sébastien**, **Siri Lie Olsen**, **Sylvia Haider**, **Thilo Heinch** and **Tiit Hallikma**.

I also wish to thank my Edysan colleagues **Thomas Kichey** and **Fabien Spicher** for their help in the field and in the lab; **Marie Boley**, **Fréderique Berquin**, and **Franck Krawczyk** for their superb organisation skills; **Françoise Dubois** for helping with almost everything, from material labs to a nice office table so I could keep working in this crazy pandemic times.

I could not go by this list without stressing my gratitude to all the team from the ST2A Master Course: **Pascale Bruxelle**, **Amelie Chabot**, and, most of all, **Jérôme Lacoux** and **Manuella Catterou** for the great opportunity the offered me of teaching for such a nice program. And, of course, my statistics students, from who I have learned so much.

I would also like to thank the internship students I co-supervised during this PhD: **Nicolas Duhamel**, **Hugo Bonelle**, and **Adeline Jalquin** for not burning down the lab. Seriously, I learned a lot from you, so thank you!

Des bisous to my fellow PhD students **Thomas Denoirjean** and **Romain Ulmer**, with who I shared an office. Hope you don't miss me too much, guys! Un coucou aussi to **Ali Almoussawi**, **Eva Gril**, and **Marion Casati**. And, of course, to all my colleagues at Edysan, especially those from BIPE with whom I spent so many hours those last three years. Not to forget **Diane Lacoste**, **Boris Brasseur**, and **Arnaud Ameline**, and **Vincent Le Roux** for all the time spent outside the lab, either around a beer or around a big cup of coffee. I was really lucky to work in a laboratory with such nice people.

To all my friends, thank you so very much for being such good listeners. Darwin knows this work could not have been possible without you.

All my love and respect to **Alexandra Elbakyan** for her work on democratising the scientific knowledge. Your legacy was of great help, especially during those mandatory home office times.

At last but in no way least, I thank **René Pestre**, my dear father that always supported this crazy idea of doing a PhD, my stepmother, my brothers, and my beloved sister for their support and love. And a special thank you to my mother, who will always be in my heart. **Obrigado, família**!

This thesis was founded by the **European Regional Development Found** and the **Région Hauts-de-France**.

If we shadows have offended,
Think but this, and all is mended,
That you have but slumbered here
While these visions did appear.
And this weak and idle theme,
No more yielding but a dream,
Gentles, do not reprehend:
If you pardon, we will mend:
And, as I am an honest Puck,
If we have unearned luck
Now to 'scape the serpent's tongue,
We will make amends ere long;
Else the Puck a liar call;
So, good night unto you all.
Give me your hands, if we be friends,
And Robin shall restore amends.

Robin Goodfellow, In: A Midsummer Night's Dream

(Shakespeare, 1595)

Chapitre 1 General Introduction **12**
Impacts of climate changes on species distribution: a glance at the past, a gaze at the present and a glimpse at the future **12**
A glance at the past **12**
A gaze at the present and a glimpse at the future **15**
Linking Species Distribution Models and Populations Genetics **18**
Studied Species **22**
Structure and aims of the thesis **29**
Chapitre 2 Population Genetics of three forest-dwelling species **32**
Introduction **33**
Investigating loci variation in *Oxalis acetosella* **35**
Sampling **35**
Existing genetic markers variability **36**
Single Nucleotide Polymorphism Characterisation of *Oxalis acetosella* **39**
Population genetic structure of *Geum urbanum* **45**
Material and Methods **45**
Sampling **45**
DNA extraction and genotyping **46**
Clustering analysis and population differentiation **47**
Results **48**
Discussion **55**
Published article: Strong genetic structure among populations of the tick *Ixodes ricinus* across its range **58**
Chapitre 3 Towards better understanding population genetic structure: contribution of species distribution models **68**
Introduction **69**
Material and Methods **71**
Species Distribution Models **71**
Present and past projections for Ixodes ricinus **71**
Present and past projections for Geum urbanum **73**
Influence of climate change on the spatial genetic structure of *Ixodes ricinus* and *Geum urbanum* **75**
Identifying candidate loci under selection **76**
Allele frequency, bioclimatic variables, habitat suitability and habitat suitability changes **79**
Influence of habitat suitability on the spatial autocorrelation of alleles of *Ixodes ricinus* and *Geum urbanum* **81**
Habitat suitability and genetic structure **83**
Results **84**

Species Distribution Models 84
Candidate loci under selection 87
Allele frequencies as a function of habitat suitability and habitat suitability changes 87
Spatial autocorrelation and influence of habitat suitability 90
Genetic structure as a function of habitat suitability 95
Hierarchical clustering analysis 96
Discussion **105**
Chapitre 4 Chapter 4: Genetically-informed Species Distribution Models **110**
Presentation **111**
Submitted article: How does incorporating genetic information improve species distribution models? **112**
Introduction **112**
Material and Methods **114**
Study species and genetic data **114**
Distribution data for *Fagus sylvatica* **115**
Distribution data for *Ixodes ricinus* **117**
Bioclimatic variables **117**
Model calibration during present-day climate **118**
Model validation and comparison during present-day climate **120**
Model evaluation and comparison during the Mid-Holocene period **122**
Results **123**
Model comparison and validation during the present-day climate (*Fagus sylvatica* & *Ixodes ricinus*) **123**
Model evaluation and comparison during the Mid-Holocene period for *Fagus sylvatica* **124**
Discussion **129**
Independent model evaluation for *Fagus sylvatica* **131**
Implications for forecasting species redistribution under future climate change **132**
Limits of the genetically-informed SDM approach **133**
References **134**
Chapitre 5 General Discussion **138**
Contributions of species distribution models to understand the evolutionary history of species **139**
Applying SDMs to test hypotheses of gene flow **140**
Perspectives for integrating population genetics into species distribution models **143**
Modelling invasive species and range expansion **145**
General limits **147**
Conclusion **148**

References **150**

Figures Index **161**

Table Index **166**

Résumé détaillé en Français **167**

Appendix 1: Supplementary Information for Poli et al., 2020. ***Ticks and Tick-Borne Dis.*** **180**

Appendix 2: Supplementary Information for the submitted article: How does incorporating genetic information improve species distribution models? **208**

List of Abbreviations

DAPC: Discriminant Analysis of Principal Components.

FDR: False Discovery Rate.

FIS: Inbreeding coefficient.

FST: Wright's Fixation Index.

GCM: General Climate Model.

IBD: Isolation by Distance.

LGM: Last Glacial Maximum.

MCMC: Monte Carlo Markov Chain.

PCoA: Principal Coordinates Analysis.

RCP: Representative Concentration Pathway

SDM: Species Distribution Models.

SNP: Single Nucleotide Polymorphisms.

WGA: Whole Genome Amplification.

Chapitre 1 General Introduction

Impacts of climate changes on species distribution: a glance at the past, a gaze at the present and a glimpse at the future

A glance at the past

Climate is one of the most important forces determining species distribution (Woodward 1987, Davis and Shaw 2001, Dawe and Boutin 2016). As the Earth experienced many cycles of warming and cooling events (i.e. glacial-interglacial cycles), some species have gone extinct, while others have gone throw similar cycles of range expansion and contraction (Bateman *et al.* 2016, Nogués-Bravo *et al.* 2018, Williams and Blois 2018) on both land (e.g. (Svenning *et al.* 2008, Diniz-Filho *et al.* 2016, Palma *et al.* 2017) and in the oceans (Bellwood *et al.* 2017). It has been suggested that Milankovitch climate oscillations are the main cause for these cyclic changes in species distribution, coined "orbitally forced species' range dynamics" (ORD) by Dynesius & Janssen (2000), which repetitions led to the geographical patterns in species diversity, species' range sizes, polyploidy, and the degree of specialization and dispersability of organisms we observe on Earth (Janssen & Dynesius 2001).

Patterns of range shift and migration following past climate changes varied among taxa. Range shifts in terrestrial species in response to climate warming are intuitively expected to take place poleward in latitude and upward along elevation gradients, following a gradient of temperature changes. Although it seems to have been the case for many (if not most of the) species (e.g. (Dawe and Boutin 2016, Williams and Blois 2018), some have shown multidirectional range shifts (Bateman *et al.* 2016). Since the last glacial maximum (LGM) event (from around 26,000 ybp to 19,000 ybp – (Clark *et al.* 2009), some species from the Norther Hemisphere have experimented a complete range shift (e.g. *Picea* spp. – (Davis and Shaw 2001), with the retraction of the south range border and the expansion of the north border, literally migrating their range distribution northward. Other species have radiated from glacial refugia or cryptic refugia, almost only expanding their range (e.g., *Ixodes ricinus* and *Fagus*

sylvatica – (Svenning *et al.* 2008, Porretta *et al.* 2013). The location of those refugia is also variable according to the species in question. Particularly in Europe, it has been long accepted that southern zones on the continent were important glacial refugia for a multitude of taxa (Taberlet *et al.* 1998, Hewitt 1999, Tzedakis *et al.* 2013), but northern refugia seem to have also played an important role in the maintenance of viable populations of certain species (Stewart and Lister 2001, Svenning *et al.* 2008, Schmitt and Varga 2012, Kühne *et al.* 2017, Quinzin *et al.* 2017) (**Figure 1.1**). To make matters even more interesting, some species have also survived in 'refugia within refugia', where populations were isolated within one or more of those refugial zones (Gómez and Lunt 2007, Abellán and Svenning 2014).

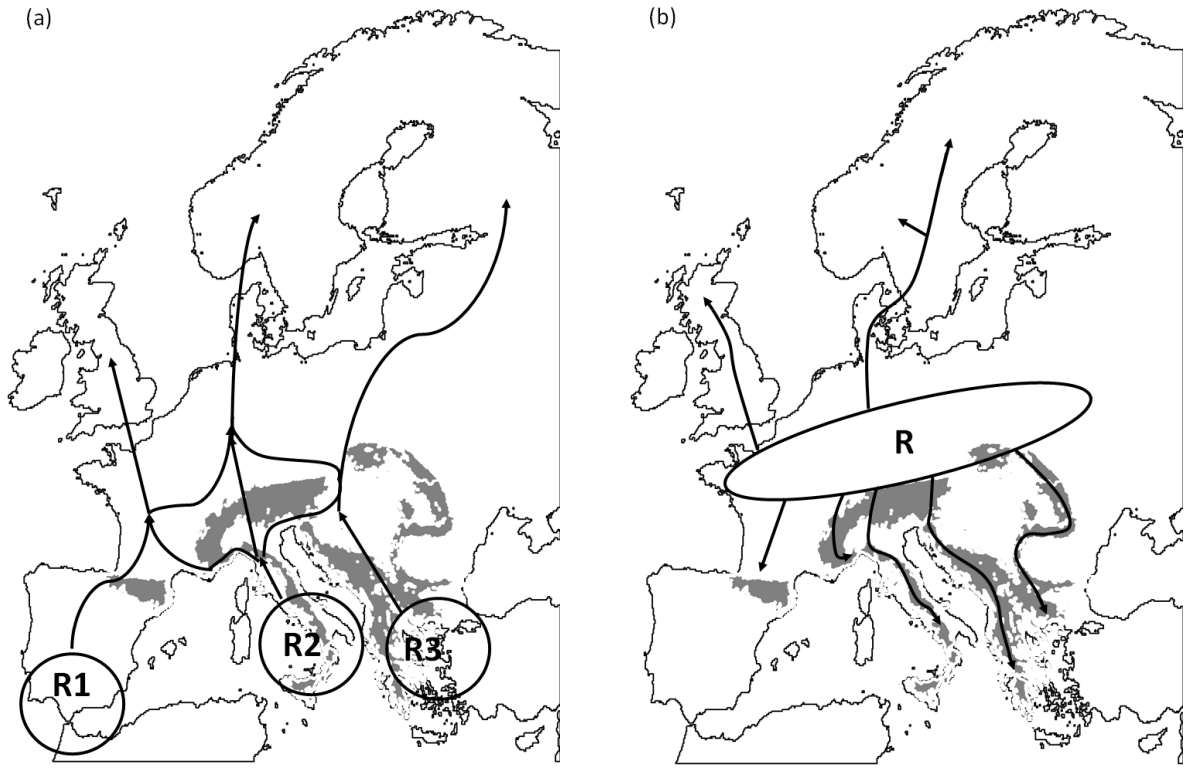


Figure 1.1. Two complementary hypothesised refugia ("R") in Europe and post glaciation range shifts (arrows): (a) the classical (Hewitt, 1999) Mediterranean refugia *versus* (b) the northern refugia hypothesis. Modified from Schmitt & Varga (2012).

Past range shifts have had important consequences for the evolution of species and the current biodiversity patterns (Dynesius and Jansson 2000, Jansson and Dynesius 2002). Rapid changes in the distribution of species, like those observed in the post-Pleistocene expansion, has led to changes in

population demography and the contact between previously isolated populations, which in turn has left genetic imprints on those populations. The scientific literature is full of examples of present population genetic structure and phylogeography highly influenced by post-Pleistocene range dynamics (Schmitt 2007). Phylogeographic investigations of current widely distributed species in Europe have shown that many of those species have survived the LGM in isolated populations distributed over the three main Mediterranean peninsulas: the so-called main glacial refugia of the Iberian, Italian and Balkan peninsulas (Hewitt 1999, Tzedakis *et al.* 2013). Brito (2005), for example, reconstructed the phylogeny of the tawny owl (*Strix aluco*) based on mitochondrial DNA (mtDNA) and has found three well-supported lineages in Europe that coincides with the Iberian, Italian and Balkans peninsulas. The species is a non-migratory bird, with highly sedentary adults and juveniles that disperse only within a few kilometres of the natal nesting sites (Coles & Petty, 1997 *in* Brito, 2005). Other species, however, have maintained connected populations during the last glacial period. That seems to be the case for the castor bean tick *I. ricinus* (Porreta et al., 2013), a very generalist ectoparasite. In this study, genetic analysis based on nuclear and mtDNA sequences coupled with Species Distribution Models (SDMs) and fossil records of the host species suggests that *I. ricinus* persisted in interconnected populations that did not experience prolonged isolation, leaving negligible genetic traces of range expansion. Those two examples are extremes cases where the biology of the focal species had either make it very susceptible to isolation due to extreme climate conditions (and leading to the strong differentiation of lineages), or seemingly not or too little affected by the climate conditions during and after the LGM. Some species still may have experienced complex dynamics during the LGM that can also be traced by phylogeographical investigations. García *et al.* (2020) have identified two endemic Mediterranean lineages and three related continental lineages of the vole species *Microtus arvalis*, suggesting that the source of the northward range expansion of this species was not the Mediterranean populations but probably more continental and northern populations outside the Mediterranean zone. In any case, it is clear that past climate fluctuations have influenced the contemporary genetic structure of species.

A gaze at the present and a glimpse at the future

The global mean annual surface temperature has increased by about 0.72°C during the 20th century and is projected to rise from 1.0°C to 3.7°C on average according to the different Representative Concentration Pathway (RCP) scenario used by the Intergovernmental Panel on Climate Change (IPCC, 2013 technical summary of the Working Group I – Stocker et al., 2014). This mean annual temperature rise will be accompanied by changes in other climate variables (such as the mean annual precipitation) as well as more frequent and severe extreme climatic events such as repeated and prolonged heatwaves and droughts, not mentioning the expected loss of coastal areas due to sea level rise. Those projections of future climatic changes in such a small window of time will lead to the redistribution of the species able to keep track (i.e. migrate) of those changes. Depending on the species tolerance to such changes, species will have to move, adapt, or face extinction (Berg *et al.* 2010). Species range shifts are already observed and have been documented across many taxa (Pecl *et al.* 2017, Lenoir *et al.* 2020), such as: birds (Virkkala and Lehikoinen 2017, Freeman *et al.* 2018); mammals (Mallory and Boyce 2018); insects (Marshall *et al.* 2020); acarians (Jore *et al.* 2011); plants (Lenoir *et al.* 2008, Lamprecht *et al.* 2018, Geppert *et al.* 2020); diverse marine species (Pinsky *et al.* 2013); and many others (see Lenoir et al. 2020 for a quantitative synthesis). Noteworthy, data shows that marine species are better at tracking isotherm shifts under contemporary climate change, and move towards the pole six times faster than terrestrial species (Lenoir et al. 2020).

Changes in species distribution may also induce complex changes throughout the range with potential range disjunctions. For instance, Kuhn *et al.* (2016) projected the future distribution of 25 submountainous forest plants under future climate change and argued that range disjunction is a likely consequence of changes in the species distribution as climate warms up. This is not a surprising result considering what is known of species redistribution during the LGM, a period during which some species had discontinuous and isolated populations across the Mediterranean refugia. Following this parallel to species past distribution changes, it is also expected that some populations will persist in

refugia or microrefugia with favourable microclimate under future macroclimate change (Kuhn *et al.* 2016, Quinzin *et al.* 2017, García *et al.* 2020).

Whatever the scenario, species range shifts have also the potential to impact species evolution (Garnier and Lewis 2016). A species' range expansion dynamic is well known for changing the frequencies of genes among populations at the leading edge (Edmonds *et al.* 2004, Excoffier *et al.* 2009). The rapid population growth observed in expanding populations at the leading edge of a shifting range can lead to high frequencies of previous rare alleles and the fixation of a new combination of alleles or new mutations (Slatkin 1996, Excoffier and Ray 2008). For instance, Klopfstein *et al.* (2006) coined the term 'surfing' to describe this process since those rare alleles or new mutations would 'surf' on the wave of the expanding range. Although not in a climate-changing context, the same general principle was also investigated by Carson & Templeton (1984) and Slatkin (1996) as the funder-flush theory (or effect). This process can lead to a reduction of the genetic diversity at the leading edge (Excoffier *et al.* 2009, Neve *et al.* 2009) and a consequent differentiation of the expanding population from the founder one. (Garnier and Lewis 2016) simulated a population experiencing a range shift under different scenarii of climate change velocity. They concluded that low velocities of climate changes do not reduce the genetic diversity at the leading edge. Since the expansion is not rapid enough, the range boundaries between expansion waves still allow the spatial mix of genetic components within the populations (**Figure 1.2**). On the other hand, a high climate change velocity, and consequently a rapid range expansion dynamic, can reduce gene diversity. This response was even more pronounced when the reduction in range size due to range retraction dynamics at the trailing edge was taken into account (i.e., local extinction or extirpation events at the trailing edge due to deteriorating conditions in macroclimate). The disjunction of a species' ranges under climate change and the persistence of isolated populations within microclimatically suitable habitats is another potentially important factor for the evolutionary processes at play (**Figure 1.3**). Isolated populations are more vulnerable to stochastic environmental and demographic effects (Baguette *et al.* 2013), inbreeding depression (Charlesworth & Willis 2009), and genetic drift. This can be particularly

important to species that are currently narrowly distributed, such as mountain endemics. Overall, changes in species' distribution due to climate changes are almost certain to impact the species genetic structure and evolution.

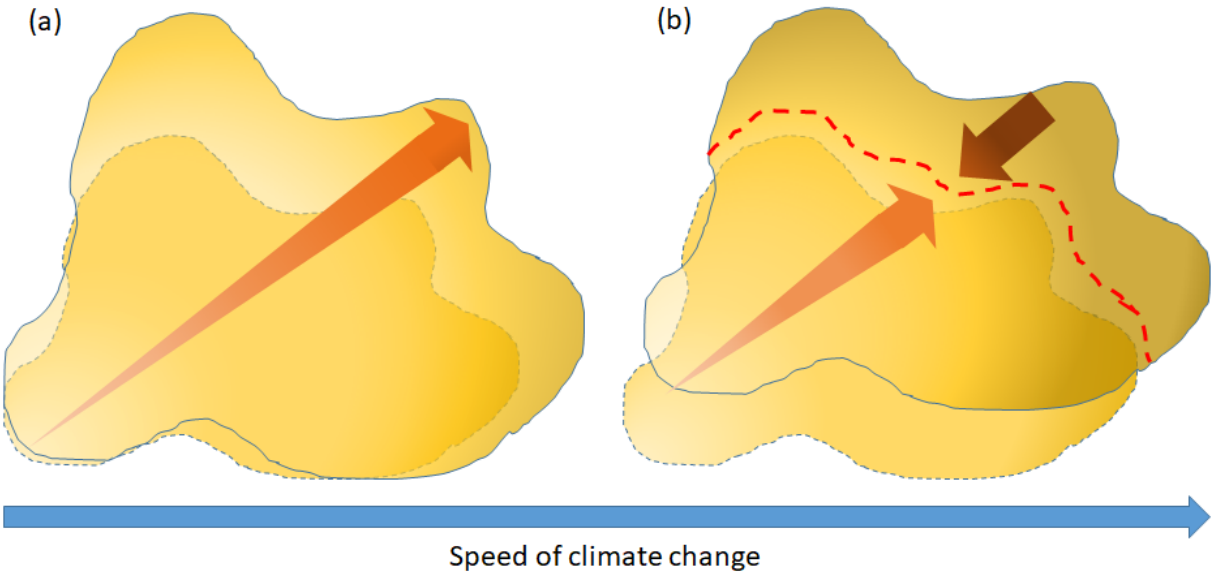


Figure 1.2. Schematic representation of the models from Garnier & Lewis (2016) of population differentiation in response to range shifts. Gradual changes in climate and consequent gradual expansion of the species' range (a) would not have important impacts in the gene diversity across populations on the leading edge. When climate is changing with high speed in comparison to the spreading speed of the species (b), an erosion of the genetic diversity will be observed on the leading edge (red broken line) compared to populations in the core of the ancient distribution. This effect will be more intense when coupled with a reduction of the trailing edge.

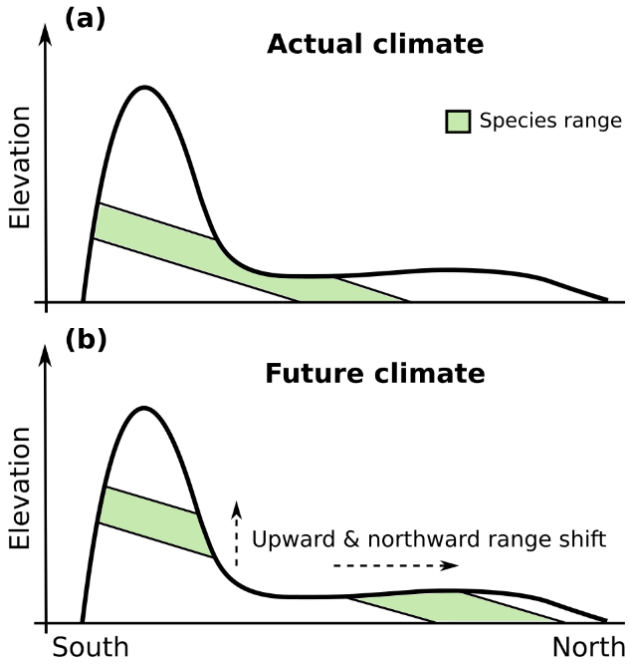


Figure 1.3. Schematic representation of range disjunction as a consequence of climate warming. In the present climate conditions (a), the species has a continuous range. Climate warming makes lowland and central areas unsuitable to the species and the range shift in both latitude and altitude causes a disjunction on the species range (b), possibly leading to the isolation of certain populations. Extracted from Kuhn et al. (2016).

Linking Species Distribution Models and Populations Genetics

Species distribution models (SDMs) are widely implemented in ecology, biogeography, and conservation biology (Guisan and Thuiller 2005, Grimmett *et al.* 2020). Most SDMs' applications include one or a combination of the following: (**i**) supporting conservation planning (Hannah *et al.* 2007, Guisan *et al.* 2013); (**ii**) forecasting or hindcasting species distribution changes in response to climate change (Thuiller *et al.* 2008, Alkishe *et al.* 2017); (**iii**) simulating the spread of invasive alien species (Guisan *et al.* 2014, Hattab *et al.* 2017); and (**iv**) investigating biogeographic and evolutionary hypotheses (Svenning *et al.* 2008, Ives and Helmus 2011). While a vast set of modelling techniques are available, most rely on correlative approaches between species presence (and absence or pseudo-absence) data and contemporary environmental variables (Guisan et al., 2017).

SDMs have long been employed to understand shifts in species distribution in response to climate changes. The general approach is to calibrate a model using current climatic conditions as

predictor variables of the contemporary data on species distribution (occurrence data, presence-absence observations or abundance data). The set of predictor variables is supposed to capture the bioclimatic realized niche (Meier et al., 2011; Maiorano et al. 2012) of the species so that one can use it to interpolate the species distribution across space and potentially over time when time series are available. By comparing the projected distributions among different time periods, a researcher can estimate how a species' range has shifted in the past or will shift in the future (Svenning *et al.* 2008, Alkishe *et al.* 2017). Some major assumptions underlying SDMs need to be observed: i. the climatic niche of the focal species remains stable over the investigated time period (i.e. no adaptation to new conditions); ii. the species is already in equilibrium with the current conditions across its range (i.e. no range filling under stable conditions), and; iii. all populations belonging to the same species share the same climatic niche and will respond identically to climate change (Wiens *et al.* 2009). Recently, different methods were conceived to deal with those assumptions.

Given the correlative nature of SDMs, it's hard to deal with the first assumption of niche stability over time, especially if time series on the focal species distribution are missing. The second assumption could be dealt with by carefully choosing the data used to calibrate the model. Hattab et al. (2017) have proposed a method to differentiate environmental from dispersal-limited absences within the spatial extent of the occupied range. Their method was conceived to deal with the non-equilibrium distribution of invasive alien species with their environment within the invaded range, especially so during the early introduction phase. Although developed for invasive alien species, nothing in the described methods prevents it from been extrapolated to contexts other than biological invasion. Although not relaxing the assumption of equilibrium, by separating environmental from dispersal-limited absences, it would be possible to better predict the potential niche of the focal species without the 'noise' of absences that relate directly to dispersal limitations and which tend to blur the signal in SDMs if ones aim at capturing the potential niche and not the realized niche.

Concerning the third assumption, even though local adaptation to local climatic conditions has already been demonstrated (Pelini *et al.* 2009, Wasof *et al.* 2013, Peterson *et al.* 2018), it is still open to debate how local adaptation may impact SDMs' performances (Pearman *et al.* 2010, Chardon *et al.* 2020, Collart *et al.* 2020). Theoretically, if distinct genetic groups within the same species respond differently to the same climatic variables, modelling those genetic units independently could provide more reliable predictions of the species response as a whole (Pearman *et al.* 2010, Peterson *et al.* 2018, Smith *et al.* 2018). In this context, some recent studies have already explored the contribution to model performance when incorporating intra-specific genetic variation into SDMs (Palma *et al.* 2017, Smith *et al.* 2018, Lecocq *et al.* 2019, Boyer *et al.* 2020, Chardon *et al.* 2020, Collart *et al.* 2020), with different results and conclusions. Most of those recent studies followed the same general three steps procedure: i. occurrence data from a given species was split according to an intra-specific level of genetic organisation (normally a phylogeographic lineage); ii. an SDM model was calibrated for each genetic unit, and; iii. the resulting models were assembled and performances were compared to the (traditional) species-level SDM. For example, Palma et al. (2017) compared SDMs calibrated at the lineage levels of two rodent species in the Andes to models of the species as a whole, finding that the assembled lineage models had a similar to better performance than the species-level model. By contrast, Lecocq et al. (2019) applied a similar approach to bumblebee species in the West-Palearctic but found that those lineage-based models did not increase model performance. Collart et al. (2020) went a step further by estimating model performances of SDMs calibrated at the intraspecific level with a small number of occurrences. This is particularly important because genetically-based SDMs normally has a reduced number of occurrences in comparison with the species level (whole) model. They observed that the combined intra-specific models predicted larger ranges than the species-level model. The authors finally suggest that, given the difficulties of model calibration and evaluation from small datasets, models should be performed at the species level unless niche divergence between intra-specific units is observed.

Going in another (complementary) direction, some studies have applied SDMs to test whether intraspecific levels of organisation (lineages or populations) display different fundamental niches (e.g. (Cooper *et al.* 2010, Schulte *et al.* 2012, Wasof *et al.* 2015, Gutiérrez-Rodríguez *et al.* 2017, Meynard *et al.* 2017). Those studies followed a general protocol similar to the three steps one mentioned described above. First, the occurrences of a given species or assemblage of species were split according to the geographic location of those occurrences or the genetic structure and phylogeography of the focal species. Next, the climatic niche of those different groups was modelled and compared. According to the observed differences in the climatic niche, the authors of the different above-mentioned studies concluded for the species niche conservation or divergence between geographic locations or genetic groups. This is a valid approach to observe differences in the climatic response of populations, but it is hard to make inferences on the evolutionary bases of those differences. If niche divergence has its bases in the evolutionary history of the species, then it is expected that at least some of the alleles in those populations are under selection from climatic variables, i.e. it should not be the consequence of plasticity or a simple geographic coincidence (occurrences from different geographic zones will most probably show different response curves to the same climatic variables). True niche divergence (i.e. adaptation) can be tested quantitatively by different means, like transplantation (Wright *et al.* 2006, Pelini *et al.* 2009, Yousefi *et al.* 2017) or controlled laboratory experiments (Wei *et al.* 2017, Sandoval-Castillo *et al.* 2020). There are also different methods to identify candidate genes under selection and to directly correlate allele frequencies to environment variables, such as Bayescan (Foll and Gaggiotti 2008), Bayenv (Coop *et al.* 2010, Günther and Coop 2013), and pcadapt (Privé *et al.* 2020), among others. However, very few studies that applied SDMs to test niche divergence went this step further to test whether the observed differences really correspond to selection and adaptation: a very challenging task. Diniz-Filho et al. (2016) applied a linear regressive approach to test the influence of changes in the climatic environment on the spatial dependency of allele frequencies from microsatellite loci of the tree *Eugenia dysenterica* in the Brazilian 'Cerrado'. Although they found significant results, no further investigation was conducted to verify whether this significance was due

to selective pressures and whether it represents true niche differentiation between populations. Yet, coupling SDMs, investigation of alleles under selection, and population genetic structure can be a simple indirect measure of selective pressures and niche divergence.

Considering the conclusions and perspectives of the different studies working on the inclusion of genetic information into SDMs, it is clear that more investigations are needed to get a better understanding of the intra-specific variability and how this can be used to better inform SDMs and thus refine future predictions on biodiversity redistribution. This is one of the main objectives of this thesis, which will be detailed at the end of this chapter.

Studied Species

Four model species were studied: the two forest herbs *Geum urbanum* (wood avens) and *Oxalis acetosella* (wood sorrel); the tree *Fagus sylvatica* (European beech); and the tick *I. ricinus* (castor bean tick). Those are forest-dwelling species widely distributed in Europe, but with contrasting levels of forest specialization and different reproduction strategies.

Geum urbanum is an ancient hexaploid (2n = 42) (Jordan *et al.* 2018), perennial, and generalist herbaceous plant species, commonly occurring in gardens and disturbed habitats (Endels *et al.* 2004) as well as in forests, especially along forest roads and tracks. The species is native to Europe, North Africa and West Asia, and is considered invasive in parts of North America, and East Asia (**Figure 1.4**). *Geum urbanum* has hermaphrodite flowers that can be pollinated by insects. Yet, the species is considered to be chiefly self-pollinated (Arens *et al.* 2004, Schmidt *et al.* 2009). Due to its common self-pollination abilities, it was supposed by Arens et al. (2004) that gene flow will depend mainly on seed dispersal. Fruits of *G. urbanum* have burrs adapted to adhesive dispersion (epizoochory) and so could be dispersed by small mammals and birds. Over most of its European range, *G. urbanum* occupies the same geographical area as *G. rivale*, although the two species are normally partially separated by ecological optimum, flowering times, and pollinators (Taylor 1997). Nonetheless, the two species may hybridise and produce fertile descendants (Taylor, 1997).

Few previous studies have aimed to understand the influence of landscape features on the spatial genetic structure of *G. urbanum* across relatively small spatial extents (Vandepitte *et al.* 2007, Schmidt *et al.* 2009) but, as far as I am aware, no study to date has tried to disentangle the species' population genetic structure across its entire geographical range. Recently, our research team (EDYSAN, UMR CNRS 7058) has been involved in the Woodnet BiodivERsA project (BiodivERsA 2016), a landscape genetic study comparing the influence of the history of land uses in contrasting landscape windows in French Britany and Picardy. Within that research framework, I co-supervised the 2019's Master 2 Internship of Nicolas Duhamel. Previous studies showed weak (Schmidt *et al.* 2009) to no (Vandepitte et al., 2007) correlation between landscape features and genetic diversity between *G. urbanum* populations. In our current project, we have found a weak genetic structure inside each of the studied landscape windows and no correlation between individual genetic distance and the presence of hedgerows. On the other hand, this study has demonstrated that, on a localised geographic scale, past land uses are more important for inter-patch gene flow than the current landscape structure (unpublished results, but see chapter 5: General Discussion). From those results, it seems that populations of *G. urbanum* are mainly structured by the history of those populations and not much influenced by present-day gene flow.

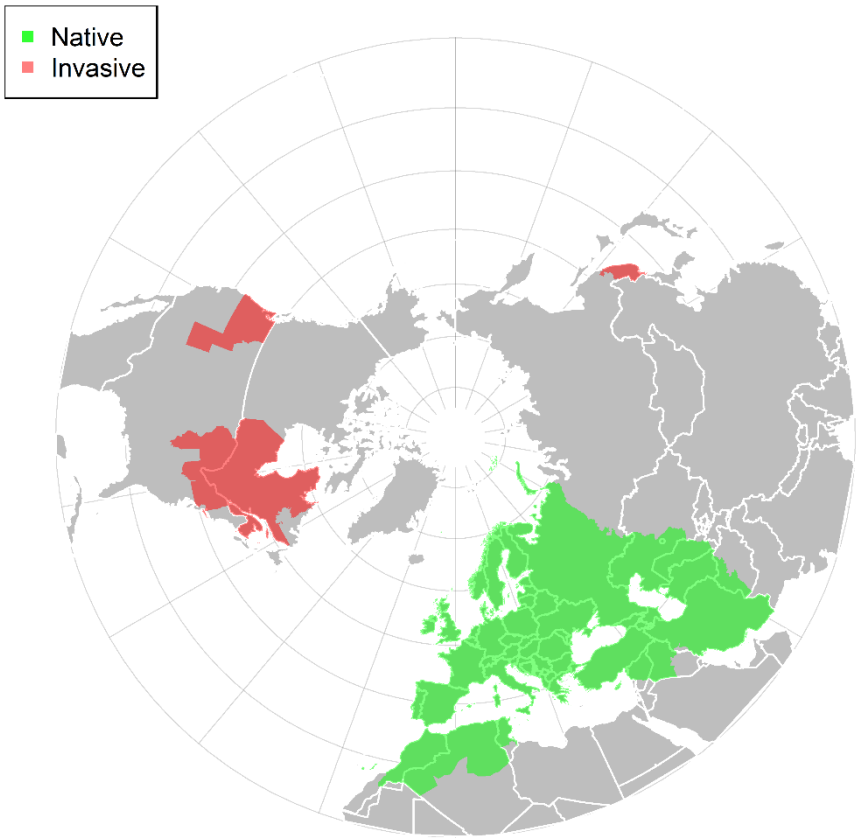


Figure 1.4. The distribution range of *Geum urbanum* including its native Eurasian range (green) and the invaded zones in North America and East Asia (red). Adapted from Plants of the World online, Royal Botanic Garden, UK (http://www.plantsoftheworldonline.org/about).

Oxalis acetosella is a diploid (2n = 22), perennial and cosmopolitan species, widely distributed across Eurasia (Marks 1956). The species has a broad geographical distribution (**Figure 1.5**) but is a more specialized species than *G. urbanum*, chiefly occurring in undisturbed forest habitats and less commonly found in disturbed systems. The species has a pronounced clonal growth, forming local patches in the forest understorey. Flowers are solitary and hermaphrodites, and can be either chasmogamous (i.e., open flowers that potentially allow cross-pollination) or cleistogamous (i.e., closed flowers where self-pollination occurs) (Berg and Redbo-torstensson 1998, Berg 2000). Fruits from both flower types are capsules containing a variable number of seeds (Redbo-Torstensson and Berg 1995, Berg 2000), which are dispersed by autochory when the fruit explodes. One recent study has investigated the phylogeny of the genus *Oxalis* as a whole (Aoki *et al.* 2019), which recognizes the Eurasian population of *O. acetosella* as a monophyletic group closely related to the East-Asian species:

O. griffithii, occurring in China, India, Bhutan, and Nepal, and *O. nipponica* occurring in Japan. Hitherto, no other work has investigated the population genetics nor the phylogeography of this species.

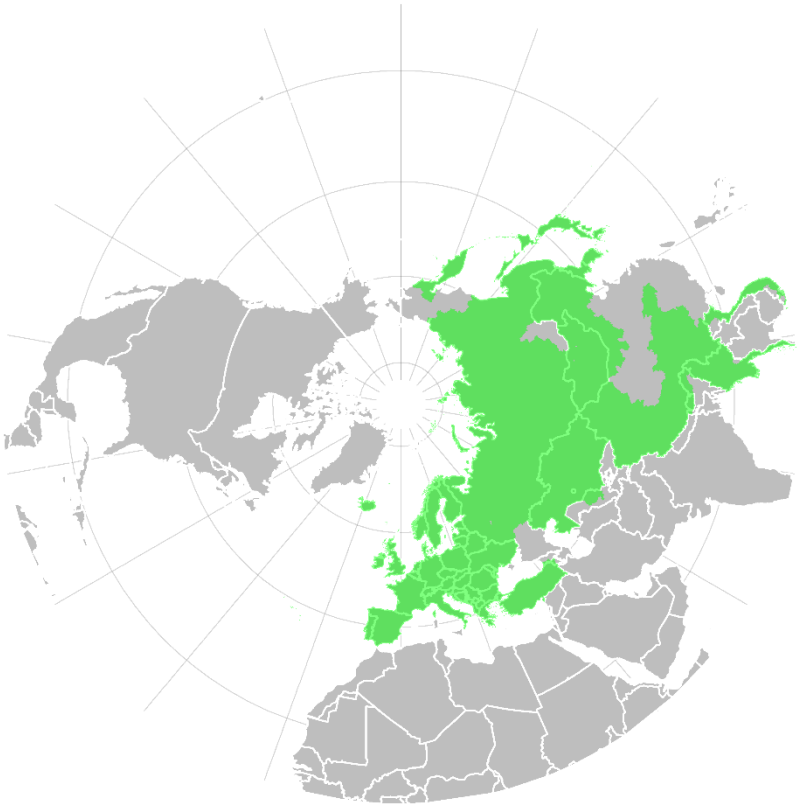


Figure 1.5. The distribution range of *Oxalis acetosella* in its native range (Eurasia). Adapted from Plants of the World online, Royal Botanic Garden, UK (http://www.plantsoftheworldonline.org/about).

Fagus sylvatica is one of the most thoroughly investigated forest trees in Europe (Demesure *et al.* 1996, Magri 2008, Svenning *et al.* 2008) because of its wide distribution in Europe (**Figure 1.6**) and most likely its economic importance (Durrant *et al.* 2016). Because of its economic value, *F. sylvatica* has been introduced outside its natural continuous range, where the populations are now considered as naturalised (Caudullo *et al.* 2017), such as in Scotland, Ireland, along the coasts of Norway, and along the coasts of the Baltic sea (see **Figure 1.6**). It is a monoecious species with a life span of about 300 years and a late reproduction stage (40-50 years old) (Packham *et al.* 2012). *Fagus sylvatica* is a wind-pollinated species and seeds are dispersed by vertebrates like squirrels, woodpigeons, and woodpeckers (Durrant et al., 2016). In most parts of the south-eastern species range, it occurs in

sympatry with the oriental beech (*F. orientalis*), a closely related taxon within the *Fagus* genre. There has been some debate whether European beech and oriental beech are two subspecies of *F. sylvatica* (*F. sylvatica* spp. *sylvatica* and *F. sylvatica* spp. *orientalis*) or two separate species (*F. sylvatica* and *F. orientalis*) (Papageorgiou *et al.* 2008, Packham *et al.* 2012). In regions where the two groups overlap, there is frequent hybridization and some intermediate types have been documented (Dorren *et al.* 2005). The species phylogeny based on morphology and molecular variation of two sequences of the nuclear internal transcribed spacer (ITS1 and ITS2) has indicated that there is only one species of *Fagus* (i.e. *Fagus sylvatica*) in Europe and Asia Minor (Denk *et al.* 2002). Since it is not the objective of the present thesis to investigate the intraspecific phylogeny of the genus *Fagus*, I will accept that all *Fagus* in the studied zone are from the same species, as suggested by Packham et al. (2012).

The genetic structure of *F. sylvatica* is well studied (Demesure *et al.* 1996, Denk *et al.* 2002, Magri *et al.* 2006). Demesure et al. (1996) have identified that most of the populations on the centre of the species distribution zone share the same haplotype of 10 chloroplast genetic markers, with other 10 haplotypes dispersed around the edges of the species range. Denk et al. (2002) have observed a much more structured pattern, with clades frequently corresponding to subtypes restricted to certain geographic zones. Finally, Magri et al. (2006) identified 9 lineages across the species range based on isozymes. The lineages identified by Magri et al. (2006) are somewhat coincident with the geographic distribution of haplotypes from Demesure et al. (1996), with a dominant lineage (#1) occupying almost exclusively the centre of the species distribution, and other lineages occurring in the peripheric regions generally situated at the edge of the species range (**Figure 1.7**).

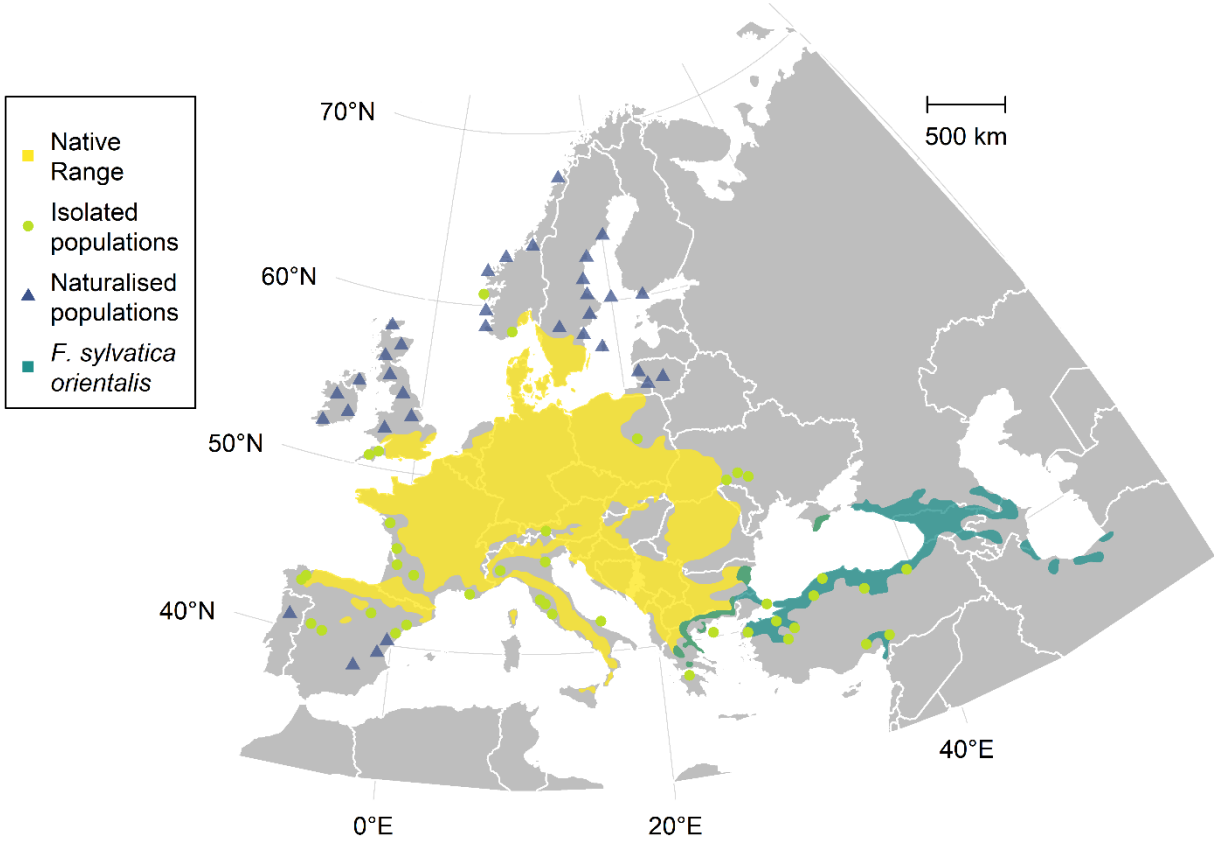


Figure 1.6. The distribution range of *Fagus sylvatica* spp *sylvatica* (yellow zones) and *Fagus sylvatica* spp *orientalis* (dark green zones). The light green dots represent isolated populations of *F. sylvatica* and blue triangles populations introduced and naturalised according to Euforgen (http://www.euforgen.org/species/fagus-sylvatica/ - Caudullo et al., 2017).

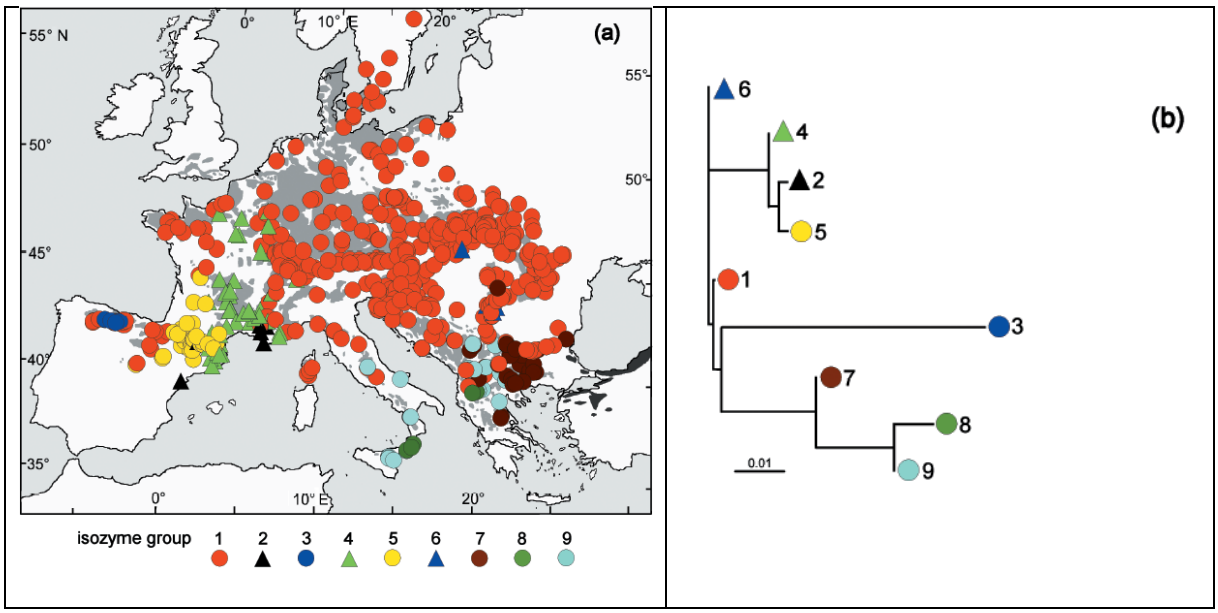


Figure 1.7. The distribution of the nine lineages of *Fagus sylvatica* based on isozymes. Modified from Figure 6 of Magri et al. (2006).

Finally, *I. ricinus* is a generalist ectoparasite species and the most widespread tick species in Europe (**Figure 1.8**), easily found in and near forested areas. It is an important vector of multiple tick-borne diseases, including: i**.** the bacteria *Borrelia burgdorferi* sensu lato, responsible for the Lyme borreliosis, which is the most prevalent tick-borne disease in temperate Europe (ECDC, 2015); ii**.** arboviruses (genus *Flavivirus*) causing tick-borne encephalitis (TBE) and louping-ill disease (LI); iii**.** the protozoan *Babesia microti*, responsible for the babesiosis; and iv**.** the bacterium *Candidatus* Neoehrlichia mikurensis, responsible for neoehrlichiosis, an emerging tick-borne pathogen (Welinder-Olsson *et al.* 2010, Portillo *et al.* 2018). Recent studies have demonstrated that the range of *I. ricinus* is already shifting northward and to higher elevations (Lindgren and Gustafson 2001a, Jore *et al.* 2011, Hvidsten *et al.* 2020) and those shifts are expected to continue in the near future (Medlock *et al.* 2013, Alkishe *et al.* 2017).

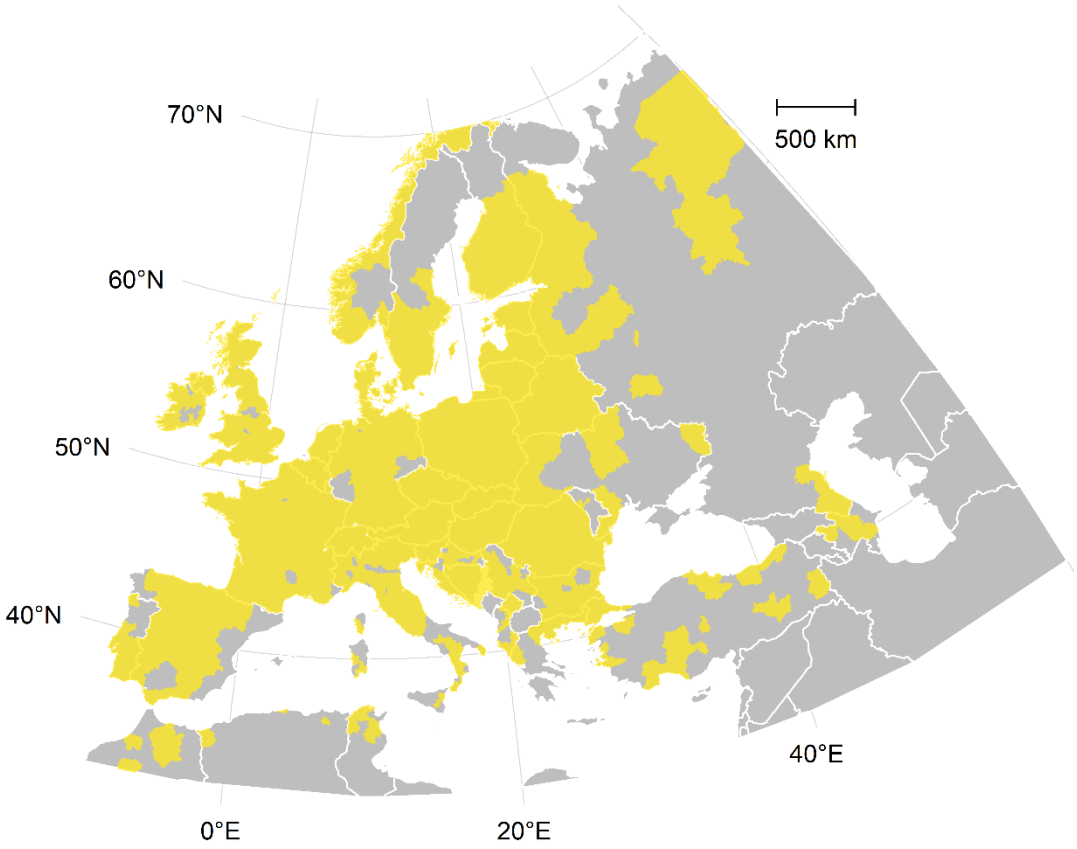


Figure 1.8. The distribution range of *Ixodes ricinus* in its native range (Eurasia). Adapted from the European Centre for Disease Prevention and Control – ECDC (January 2020).

Structure and aims of the thesis

The main objective of this thesis is to investigate some of the intersections between the research fields of SDMs and the spatial distribution of genetic variation in a climate change context. From this main objective, three more specific aims were established:

I. Determining the population structure and phylogeography of the study species for which the phylogeographic structure is unknown at the European scale, i.e. for three of the model species: *Oxalis acetosella*, *Geum urbanum*, and *Ixodes ricinus*;

II. Investigating some of the ways by which SDMs may help to understand the current genetic structure of the model species;

III. Investigating the usefulness and effectiveness of incorporating genetic information into SDMs.

Considering that the chosen model species do not have the same level of available genetic information, the starting point of the analysis varied according to each model species. For *O. acetosella*, the first step was to identify candidate genetic markers for the population genetics analysis. Investigations of *G. urbanum* and *I. ricinus* started at the population genetics point. *Fagus sylvatica* already has an established phylogeographic structure (Magri et al., 2006), and the geographic distribution of lineages was directly used to validate the incorporation of intraspecific genetic information into SDMs. At the beginning of each chapter, I will present a schematic representation of the different levels of analysis and the model species being used (**Figure 1.9**).

The thesis consists of four additional chapters after this introduction chapter as follows:

Chapter 2: Investigation of candidate loci for *O. acetosella* and population genetics analysis of *G. urbanum* and *I. ricinus*. For *I. ricinus*, the study has been published in the journal *Ticks and Tick-borne Diseases* in 2020, and the published version is thus provided in the thesis manuscript.

Chapter 3: Attempt to go beyond the mere comparison of projected distributions from SDMs and the spatial genetic structure of species. I will apply correlative and regressive methods to investigate the role of past climate changes moulding the observed present genetic structure and allele frequencies across populations of *G. urbanum* and *I. ricinus*.

Chapter 4: Assessment of the putative benefits, in terms of model performance, to add genetic information into SDMs. For this goal, I will compare the results of genetically-informed SDMs versus traditional SDMs for *I. ricinus* and *F. sylvatica* during both contemporary and past climate conditions. This chapter corresponds to a standalone manuscript submitted to *Global Ecology and Biogeography* (October 28th 2020) and which is currently under review in this scientific journal. The manuscript is presented here, in this fourth chapter. Beside changing the placement and cross-references of figures and tables to better reading, the manuscript is presented as submitted. Since data from *G. urbanum* was not available by the time of the draft production, the species is not included in this study.

Chapter 5: Attempt to synthesise the main results of the thesis and discuss the limitations and perspectives in face of those results.

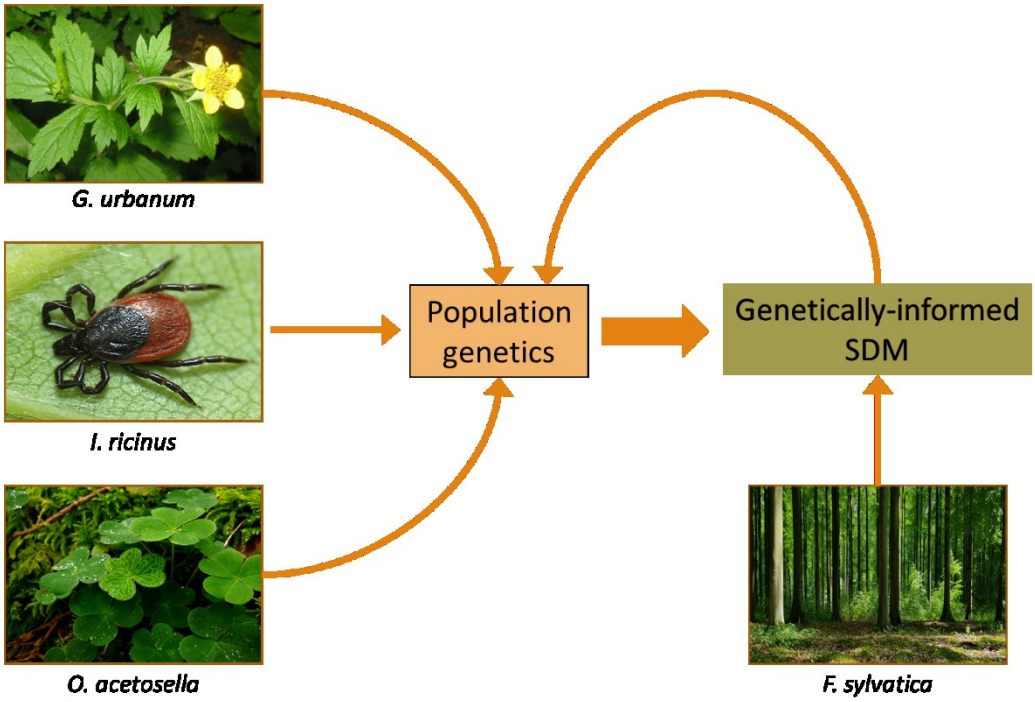
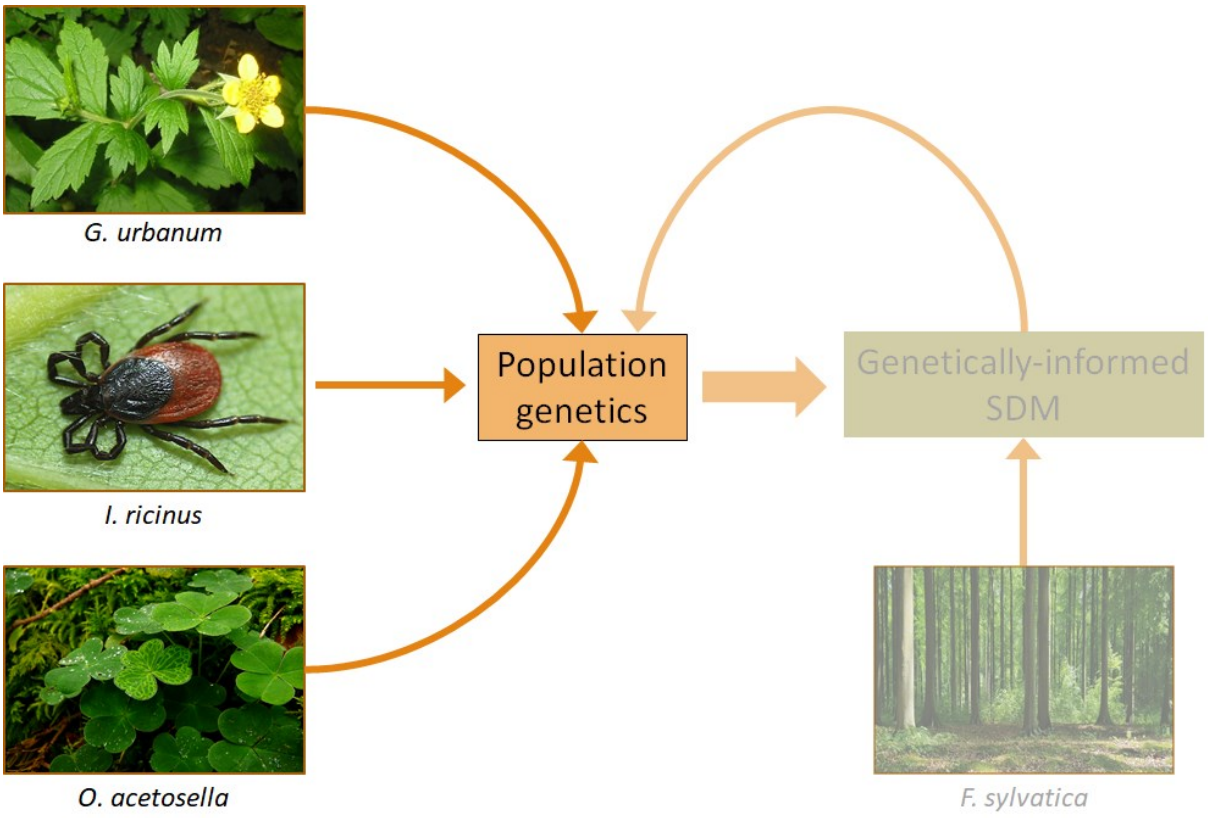


Figure 1.9. Schematic representation of the thesis. In the second chapter, I will present the genetics analyses for *Oxalis acetosella*, *Ixodes ricinus*, and *Geum urbanum*. In the third chapter, I will apply species distribution models (SDMs) to better understand the observed genetic structure of the model species. In the fourth chapter, I will investigate the gain in SDMs' performances by the inclusion of genetic information. In this fourth chapter, I will compare genetically-informed SDMs against traditional (whole-species) SDMs for *I. ricinus* and *Fagus sylvatica*. Source of photos: *I. ricinus*: https://alchetron.com/Ixodes-ricinus; *G. urbanum*: https://en.wikipedia.org/wiki/Geum_urbanum; *O. acetosella* and *F. sylvatica*: personal collection from Dr. Jonathan Lenoir.

Chapitre 2 Population Genetics of three forest-dwelling species



Introduction

Population genetics may help to understand and describe the present and past ecological dynamics such as (but not limited to) migration and gene flow (Rousset 1997), cryptic genetic variation (Vial *et al.* 2006), and local adaptation (McCoy 2008). In case of disease ecology, population genetics may also help to inform predictions about future risks of pathogen spread between populations (Kozakiewicz *et al.* 2018). Population genetics approaches such as individual genetic clustering and assignment methods help identify genetic units and genetic discontinuities that could be linked to the geographical distribution of those populations. Considering the main objectives of this PhD thesis (*see Chapter 1: General Introduction*), understanding the genetic structure at the spatial extent of the entire species range is the first step towards incorporating the genetic information into species distribution models (SDMs).

Three of the four model species were investigated in terms of population genetics: the two forest herbs *Geum urbanum* and *Oxalis acetosella* as well as the tick *Ixodes ricinus*. Those are forest-dwelling species widely distributed across Europe, but with contrasting levels of forest specialization (*see Chapter 1: General Introduction*).

The main objective of this chapter was to characterize the genetic structure of the three model species and infer the intraspecific genetic dynamics. I hypothesized that the two herbaceous plant species occurring in temperate forests in Europe should show a contrasting population structure, as a result of their contrasting levels of habitat specialization and dispersal abilities, with a stronger genetic structure observed between populations of *O. acetosella* across Europe as opposed to *G. urbanum* for which I assumed a lower genetic structure. Concerning the genetic structure of *I. ricinus,* besides its environmental exigence and its occurrence mainly in forested areas, I expected that a generalist ectoparasite should show a week to no genetic structure.

In this first chapter, I will first explore the results found for the two herbaceous plant species (*O. acetosella* and *G. urbanum*), before presenting the results for *I. ricinus* which have been recently

published in the scientific journal *Ticks and Tick-borne Diseases* and under the title "*Strong genetic structure among populations of the tick Ixodes ricinus across its range"* (Poli *et al.* 2020). All Supplementary Informations cited in the article are presented in **Appendix 1**.

Investigating loci variation in *Oxalis acetosella*

Considering that no genetic marker was available for *O. acetosella*, the first step was to identify loci that would allow characterising the population genetic structure and the phylogeography of this model species. Three different types of traditional genetic markers were investigated during my PhD thesis.

Sampling

Samples were collected from most of the species' European distribution range (**Figure 2.1**). Sampling took place during two consecutive years, in 2017 (green triangles in **Figure 2.1**) and in 2018 (black dots in **Figure 2.1**). The 2017's samples were used to test the variability of existing genetic markers and to characterize Single Nucleotide Polymorphisms, while 2018's samples were intended for the population genetics analysis *per se*. Sampling during the year 2017 followed a flexible protocol, with the only constraint being that samples had to belong to different patches of *O. acetosella* to avoid clones. The 2018 sampling campaign followed a more constrained protocol. First, all sampled locations were collected at a minimum of five meters from each other. From each individual in a given location, between three to five leaves were collected, and only "healthy" leaves were used in the following analysis (no signs of infection by fungus and no signs of herbivory). All sampled locations were georeferenced by GPS. Leaves from each sample were stored in individual paper envelopes (one envelope for each a group of leaves from one individual) and air-dried. Samples were collected by colleagues around Europe after a call on Twitter, by using mailing lists of colleagues and throughout scientific meetings during 2018. All volunteers received a detailed and illustrated sampling protocol (https://jonathanlenoir.wordpress.com/2018/05/). After receiving samples, all individual leaves were verified before storage. In the year 2018, 348 individuals were sampled across 30 different populations (from 5 to 22 individuals sampled per population, mean = 11.6).

DNA extraction

The DNA extraction methods used differed between the two sampling years. For the 2017's samples, DNA was extracted with the E.Z.N.A Plant DNA Kit (Omega Bio-tek), following the supplied protocol. For the 2018's samples, DNA was extracted with the DNeasy 96 Plant Kit (Qiagen). As usual for DNA extraction from plants, DNA was always treated with RNAse in the first steps of the extraction protocol.

Existing genetic markers variability

Microsatellites variability

First, I searched the scientific literature for variable loci from related species. A set of twelve polymorphic microsatellites loci has been developed for a North American species from the *Oxalis* genus, namely *O. montana* (Tsyusko *et al.* 2007). However, based on results from colleagues from the Leibniz Centre of Agricultural Landscape, Müncheberg (Dr. Naaf, personal communication), those loci were not amplified in *O. acetosella.* (Weising and Gardner 1999) have characterized a set of 10 microsatellites from the tobacco chloroplast genome (CCMPs) and were successfully amplified in a vast diversity of non-related Angiosperm families, including representatives from both monocots and dicots (Actinidiaceae, Brassicaceae, Fabaceae, Rosaceae, Myrtaceae, Poaceae, and Agavaceae). I tested the application of those microsatellites for population genetics of *O. acetosella* in Europe by verifying the presence of variation within samples from North Germany, Finland, North and South France, Czech Republic, West Romania, and South Sweden. Seven out of ten of those microsatellites were amplified with the primers designed by Weising and Gardner (1990) (CCMP-2 to CCMP-7, and CCMP-10). The multiplex PCR followed the protocol of Weising and Gardner (1990). The PCR products were migrated in 2% Agarose gel for 45 min at 100V and in 40% Acrylamide gel for 40 min at 200V, followed by a 20 min migration at 135V (**Figure 2.2**a). The CCMP 4 locus was the only one that seemed to vary across individuals, and only in the Agarose gel (**Figure 2.2**b). After this preliminary analysis, it was decided not to follow up with those uninformative markers.

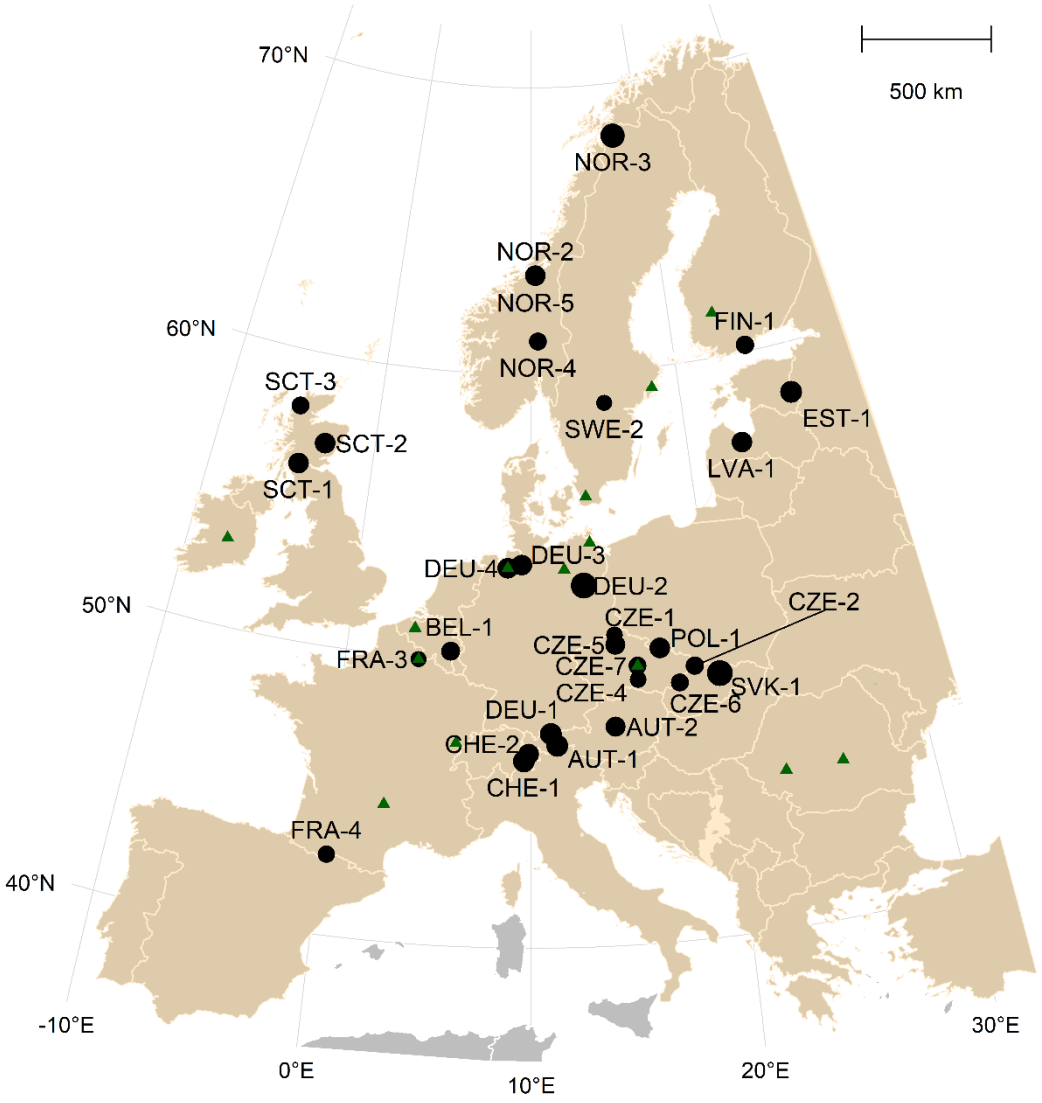


Figure 2.1. Sampling locations for *Oxalis acetosella* from 2017 (green triangles) and 2018 (black dots). Only IDs of the 2018's samples are shown. Differences in size of dots represent differences in the number of individuals sampled per site during the 2018 campaign (5-22). The light brown colour in the background represents the species distribution range across the study region (Kwescience, Plants of the World online, 2020 - http://www.plantsoftheworldonline.org/).

Nuclear and Chloroplast sequences variability

I investigated the variability of one nuclear and one chloroplast sequence among the same populations described in the previous section: the internal transcribed spacer (ITS) region of the nuclear sequence of the ribosomal RNA gene and the non-coding chloroplast sequence petA-psbJ. Although ribosomal sequences are normally applied to resolve phylogenies of higher than family taxonomic levels, the ITS

region is considered to evolve rapidly (White *et al.* 1990) and may be used at the species level (Grivet and Petit 2002, Rosselló *et al.* 2007, Jin *et al.* 2016). Non-coding chloroplast sequences have different levels of variation (Shaw *et al.* 2005) and have been frequently applied to phylogeography and population genetics at the intraspecific level (Yang *et al.* 2017, Sánchez-del Pino *et al.* 2020). We tested the variability of the ITS sequence described by White et al. (1990) and the chloroplast sequence petA-psbJ described by (Shaw *et al.* 2007) in analysing two individuals of each region for which we have data, i.e., North Germany, Finland, South France, Czech Republic, West Romania, and South Sweden. Amplification protocols followed White et al. (1990) and Shaw et al. (2007) for ITS and petA-psbJ, respectively. Sequencing was conducted in an Applied Biosystems 3130 XL 16 capillary sequencer at the "Institut Génétique & Développement de Rennes" (IGDR UMR 6290 CNRS – UR1). Base correction and sequences aligning were conducted in CodonCode Aligner version 8.0.1. (CodonCode Corporation).

Results for the two sequences were again discouraging. No variation was observed in the ITS locus. Aside from one sample from South France, no variation was observed for the petA-psbJ locus. We checked the Genbank for the sequence from the one varying sample and it coincides with deposed sequences from the close-related species, *O. montana.*

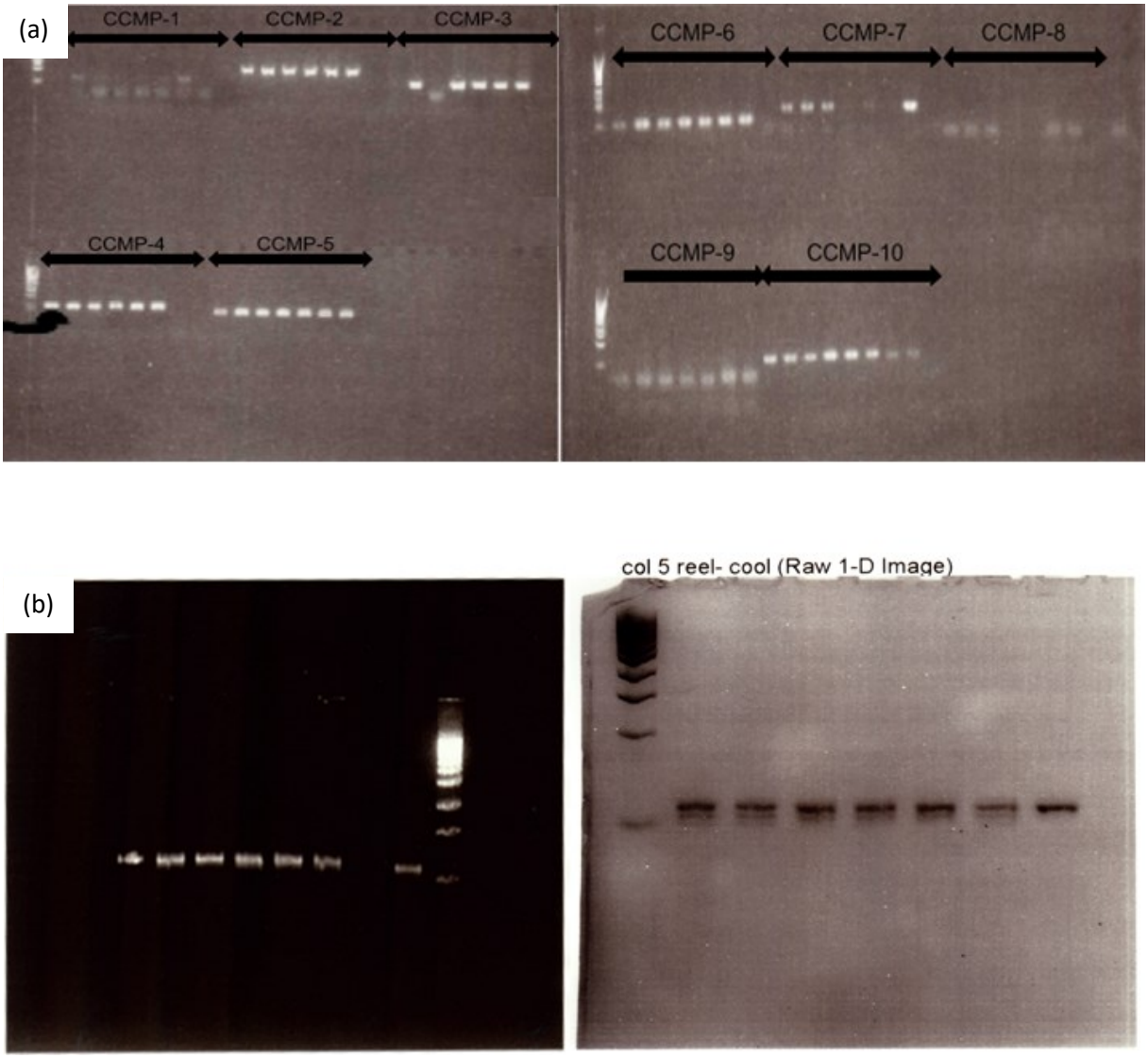


Figure 2.2. Genetic variation explored in *Oxalis acetosella*. Migration of the ten microsatellite loci characterized by Weising and Gardner (1990). In (a) the ten loci were migrated in Agarose gel. In (b) detail of the locus CCMP-4 in Acrylamide gel (left) and Agarose gel (right). The CCPM-4 locus showed no variation when PCR products migrated in Acrylamide gel but seemed to have at least two alleles when PCR products migrated in Agarose gel. Each well was loaded with the PCR product for one individual from North Germany, Finland, North and South France, Czech Republic, West Romania, and South Sweden.

Single Nucleotide Polymorphism Characterisation of *Oxalis acetosella*

Considering the results of the investigation on the variability of published markers for *O. acetosella*, we included the species in the joint project 'ASSETS: 2nd phase of BASC flagship project'. This project

includes the SNP characterisation of multiple species and is integrated by the UMR 'Ecologie et Santé des Ecosystèmes' (INRAE – Agrocampus Ouest), Université Paris Saclay and the INRAE institute of Grignon. The project was intended to identify SNP bi-allelic loci in multiple species, following the pipeline described by (Delord *et al.* 2018). This method is a multispecies approach that allows the characterisation of high-quality SNPs. The pipeline is similar to traditional RAD-seq and *de novo* assembly: pooled DNA from individuals from each species are digested with a chosen restriction enzyme, DNA fragments are tagged according to the species pool, and the final library is sequenced. Delord et al. (2017) proposed a bioinformatics pipeline using the Stacks software (Catchen *et al.* 2011) and Python scripts to further select SNPs based on: i. coverage, position and polymorphism (first filter); ii**.** specificities of flanking regions to favour primers design (second filter); and iii**.** an almost redundant third step of filtering reads based on the presence of polymorphism in the flanking regions to assure unique SNP candidates (third filter).

Library preparation and sequencing

For the library preparation, 10 samples from different populations from the 2017 campaign were pooled. To be included in the pool, a sample should have a minimum DNA concentration and quality. Only samples with DNA concentration superior to 20 ng/µl were selected. DNA quality, as measured from the absorbance ratios of 260nm/280nm and 260nm/230nm, was frequently low. Only samples that had a 260/280 ratio superior to 1.5 and a 260/230 ratio superior to 1.2 were selected. Absorbance was measured on a multimode lector Infinite® M1000 (Tecan) at the Regional Resource Centre of Molecular Biology (CRRBM, Amiens France). After this filter, two samples from North Germany, Romania, Finland and Sweden as well as one sample from South France and Belgium were selected. Although some of the selected samples did not rank on the top DNA concentration and/or quality, this was thought to be the best compromise between population representativeness (and, by consequence, DNA variability) and DNA concentration and quality. Individual samples were

concentrated in a speed vacuum and concentration was normalized to 46 ng/µl, which was the final concentration of the pool, with 260/280 and 260/230 ratios of 1.7 and 1.45, respectively.

The pool of samples was then digested with the *PstI* restriction enzyme. The enzyme recognises the sequence 5'CTGCAG and cleaves the guanine-adenine bond. For the Delord pipeline, an excessive number of digested fragments from one particular species can, theoretically, impact the read depth of the other species. *Oxalis acetosella* has a genome size of 6.2 Gb (Grime and Mowforth 1982) and the GC (Guanine and Cytosine) content, albeit unknown, was estimated to reach ~40% based on the results of (Biswas and Sarkar 1970) for *O. corniculata*. The restriction enzyme *PstI* was than chosen because of the intermediate number of fragments (Parchman *et al.* 2018). In addition to *O. acetosella*, the final species pool was composed of six other species, including three invertebrates and three vertebrates.

All the digestion steps, multispecies library preparation, and sequencing were conducted at the MGX sequencing platform (Montpellier, France). Sequencing was conducted by SBS (Sequencing By Synthesis), which consists of the sequential incorporation and detection of nucleotides. All sequencing steps were conducted on NovaSeq 6000 sequencer. A preliminary bioinformatic step was conducted by MGX consisting of filtering off reads with low quality and demultilplexing. At the end of this step, more than 14,000,000 reads of *O. acetosella* were kept for further analysis, representing 4% of the total reads for the seven species in the species pool.

Bioinformatics treatment and candidate SNP selection

Bioinformatics treatments were conducted by the UMR Ecologie et Santé des Ecosystèmes (INRAE-Agrocampus Ouest). *De novo* assembling was conducted in Stacks and SNP candidates were filtered according to the bash/Python script from Delord et al. (2017), at which point 1224 candidates bi-allelic sequences were selected and received as a .fasta file. All sequences where the polymorphism was situated at less than 52 bases from the extremities were excluded from further analyses. Next BLAST was queried for similarities and any sequence with a query cover of 100% were also excluded. From the 524 remaining sequences, 192 were randomly selected for validation.

SNP genotyping

Genotyping was also conducted at the MGX sequencing platform. All of the 348 samples from 2018 (**Figure 2.1** black dots) were sent for genotyping of the 192 SNPs selected in the previous step. DNA concentration from our samples was very variable, from 10 ng/µl to more than 200 ng/µl. Samples with less than 40 ng/µl were first amplified by Whole Genome Amplification (WGA), and all samples were then normalised to that concentration. Genotyping was conducted in a Biomark HD System (Fluidigm) and KASPar assays. The KASPar method is a KBiosciences competitive allele-specific PCR amplification. A PCR mix containing two allele-specific forward primers and one common reverse primer was carried out. Each forward primer had a 5' tail sequence homologous to universal secondary oligos labelled with a fluorophore (FAM or HEX). If a particular locus is homozygous, only one fluorescent signal is generated. Bi-allelic loci generate both fluorescent signals. The system allows the genotyping of 96 samples × 96 markers in a unique run. For each group of 96 candidate SNPs, 95 (plus a negative control) were genotyped. For this first investigation, two runs of 96 loci were tested over 95 samples.

The results were as disappointing as it could be. **Figures 2.3** and **2.4** show the allele calls for the two aforementioned runs. Each column is a candidate SNP and each row is a sample. When a sample is homozygote for a particular locus, it will show as a red or green dot (depending on the allele for which it is homozygote). If a sample is heterozygote for a locus, it will show a blue dot. Most of the loci from the first run (**Figure 2.3**) showed no variation, while most of the loci of the second run (**Figure 2.4**) were not even amplified. Concerning this second run, the few amplified loci are most probably artefacts. There are two main explanations for those results. First, DNA quality could be the main source of errors. The 2018 samples were stored in paper envelopes at room temperature for around a year before extraction, which probably contributed to DNA degradation. It was, nonetheless, possible to amplify the petA-psbJ sequence from a selection of those samples following the same protocol applied to the samples from 2017. Another possible reason for the results is if something went wrong

during the bioinformatics step, although for the moment this possibility could not be verified. The next step is to try to validate those SNPs with a small number of the best quality samples, but those results will not be available soon.

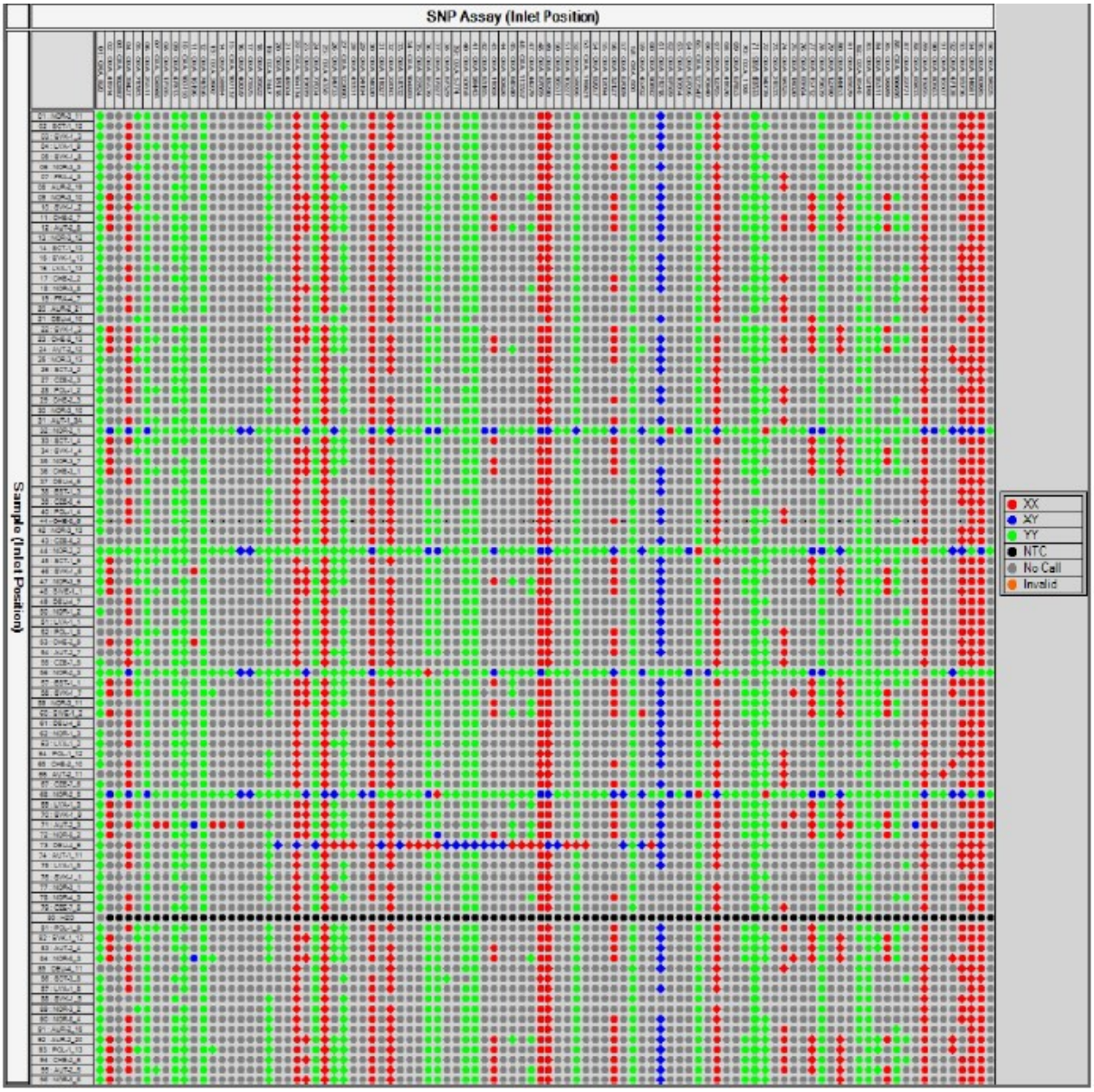


Figure 2.3. Allele call for the first run of 96 SNP loci for 95 samples of *O. acetosella* and one negative control (black dots). For a particular locus, samples may be homozygote for one allele (red or green dots) or heterozygote (blue dots). Most of the loci showed no variation across samples, with very few loci exhibiting one to three of the rare alleles.

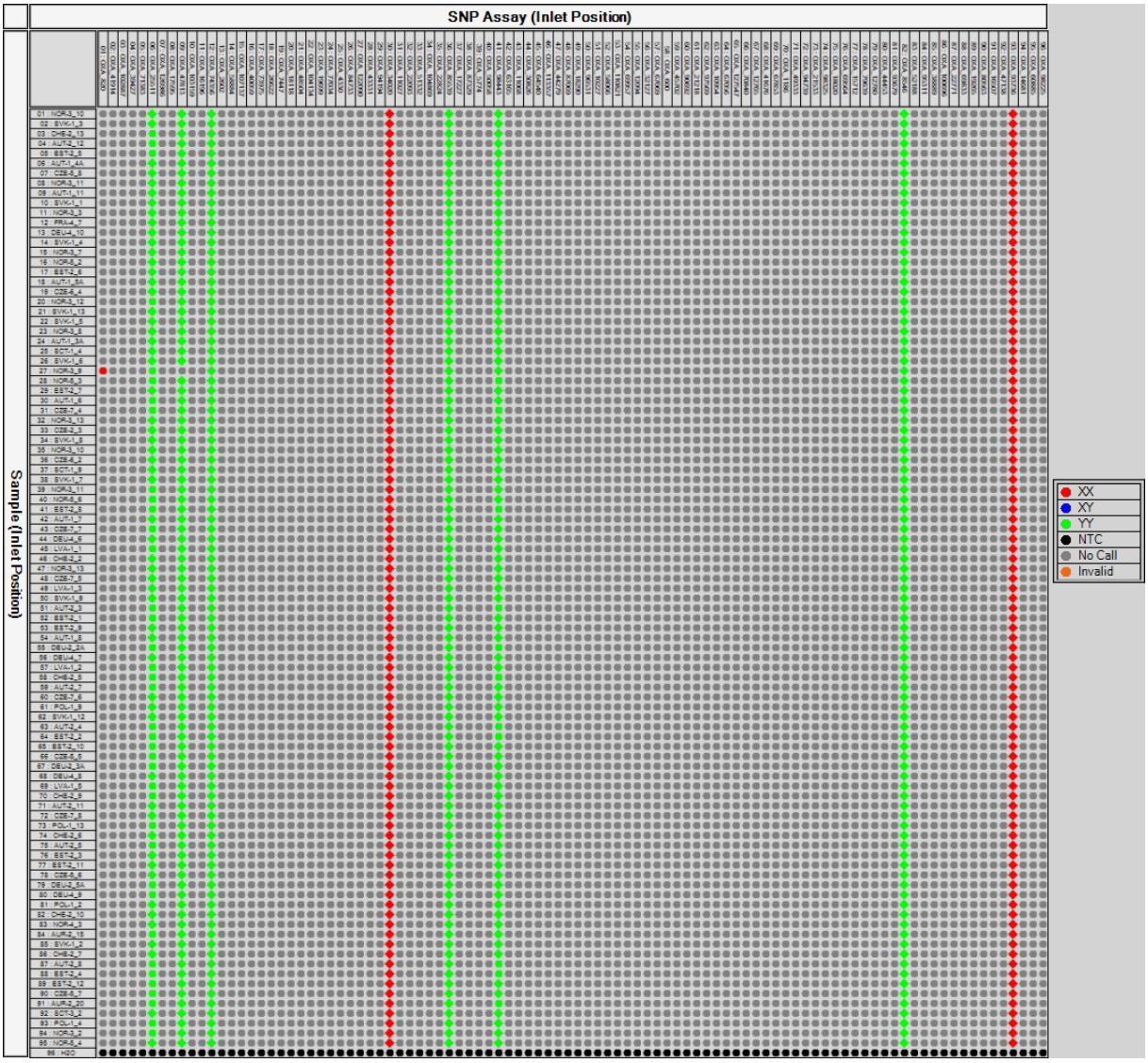


Figure 2.4. Allele call for the second run of 96 SNP loci for 95 samples of *O. acetosella* and one negative control (black dots). For a particular locus, samples may be homozygote for one allele (red or green dots) or heterozygote (blue dots). Almost no loci were amplified, and the few loci that seem to have been amplified are most probably artefacts and it seems that they are homozygote for the same allele in all samples.

Population genetic structure of *Geum urbanum*

Material and Methods

Sampling

Analysed samples were collected from most of the species' European distribution range (**Figure 2.5**) during 2018. All sampled individuals were collected a minimum of five meters from each other. From each individual in a given location between three to five leaves were collected, and only "healthy" leaves were used in the following analysis (no signs of infection by fungus and no signs of herbivory). All sampled locations were georeferenced with a GPS. Leaves from each sample were stored in individual paper envelopes (one envelope for each a group of leaves from one individual) and air-dried. Samples were collected by colleagues around Europe after a call on Twitter, by using mailing lists of colleagues and throughout scientific meetings during 2018. All volunteers received a detailed and illustrated sampling protocol (https://jonathanlenoir.wordpress.com/2018/05/). After receiving samples, all individual leaves were verified before storage.

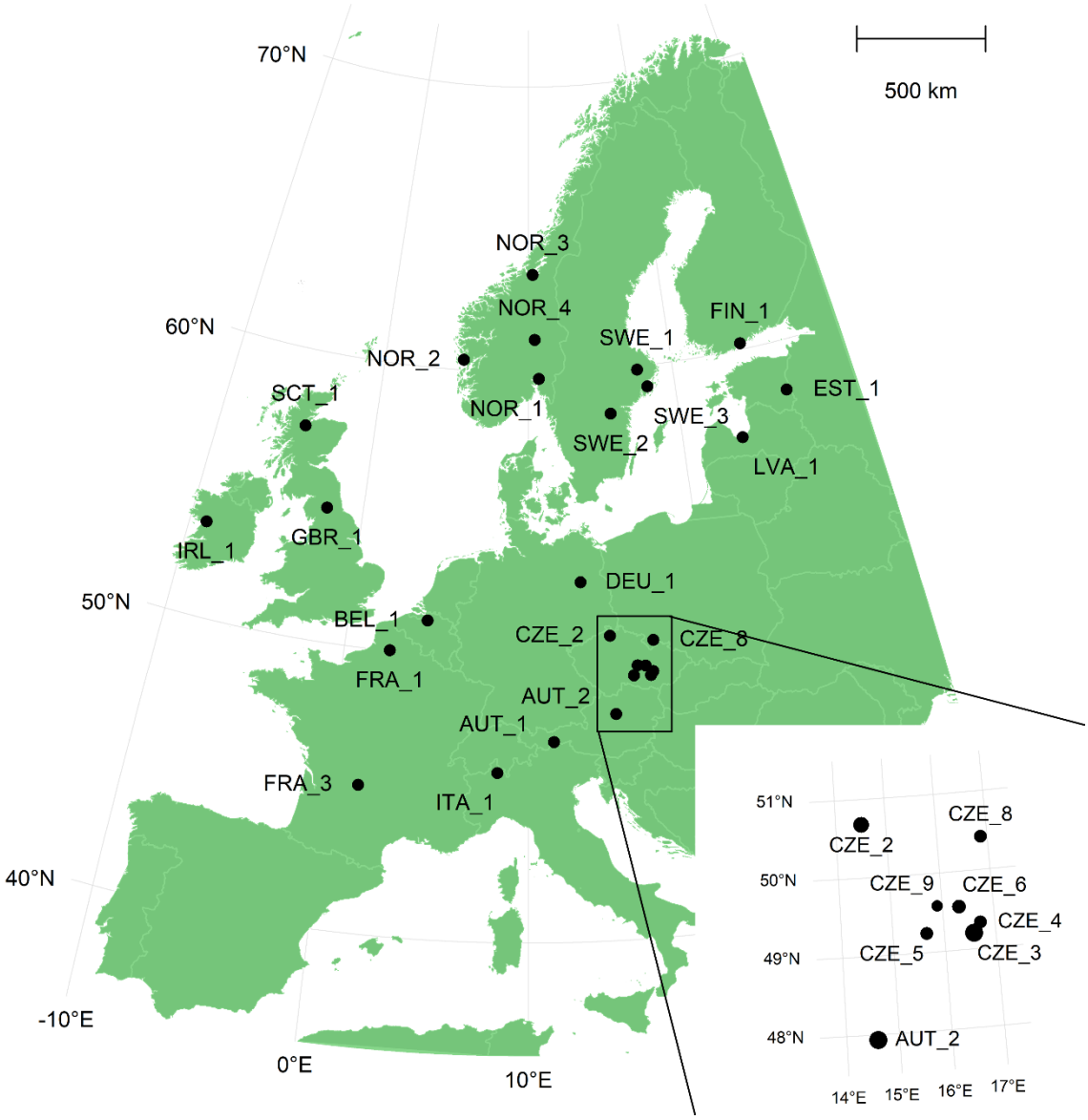


Figure 2.5. Sampling locations and IDs for *Geum. urbanum*. Differences in size of the points represents differences in the number of individuals per site (5-17). In green, the distribution range of the species, that cover all of the study area (Kwescience, Plants of the World online, 2020 - http://www.plantsoftheworldonline.org/).

DNA extraction and genotyping

DNA was extracted at Genoscreen (Lille, France) following the protocol from the kit NucleoSpin Plant II (Macherey-Nagel). Six microsatellites loci were genotyped in two multiplex, with fluorescent dyes added to the forward primer: i. WGU2-28 (HEX), WGU6-5 (FAM), and WGU6-7 (NED); ii. WGU2-10 (FAM), WGU2-48 (HEX), and WGU6-1 (NED) (Arens *et al.* 2004). PCR was performed in 20 µl reaction volumes containing 10 µl of mix FastStart Taq DNA polymerase (Roche), 0.8 µl of each pair of primers,

and 1.5 µl of DNA extract. The PCR program was the same as described in Arens et al. (2004): 95 °C for 15 min; followed by 35 cycles of 94 °C for 30 s, 57 °C for 90 s, and 72 °C for 90 s; followed by a final elongation step at 72 °C for 10 min. Fragments were separated in an Applied Biosystems 3130 XL 16 capillary sequencer at the 'Institut Génétique & Développement de Rennes' (IGDR UMR 6290 CNRS – UR1). The scoring of microsatellites was performed in a semi-automated way using GENEMAPPER v4.0. First, a panel was created for each locus. Next, all chromatographs were loaded in GENEMAPPER and an automatic analysis was performed using default parameters. Samples with good quality scores were selected to automatically generate bins for each locus. Another analysis was performed and bins were adjusted based on the results. Then, all picks were manually verified and, whenever needed, allele sizes were manually corrected. Finally, all samples with ambiguous pick sizes or with no amplification for at least one loci were excluded from further analysis. The final data set was composed of 302 samples from 27 populations without any missing data.

Clustering analysis and population differentiation

A Bayesian analysis was performed in STRUCTURE (Pritchard *et al.* 2000), with parameter *K* (i.e., the optimal number of genetic clusters) varying from 1 to 15. Ten repetitions of an admixture model with the population of origin as prior and 80,000 MCMC iterations with the burning of 20,000 iterations were run for each value of *K*. The results were analysed with Structure Harvester (Earl and vonHoldt 2012). The best *K* value for the optimal number of clusters was identified by comparing the estimates of log probabilities of the data (i.e. $\ln[\Pr(X|K)]$) for each *K* value as well as Evanno's delta *K* method (Evanno *et al.* 2005). Janes *et al.* (2017) have pointed out that this method frequently identifies a $K = 2$ when data has 'a representative number of markers'. This is not the case in this analysis, since only six microsatellites loci were presented in the data. Nonetheless, besides being one of the most widely applied methods to analyse results from STRUCTURE, results from Evanno's method should be regarded with care. The final value of *K* was chosed based on the combination of Evanno's method and qualitative analysis of the $\ln[\Pr(X|K)]$ distribution and the genetic structure for all values of *K* from 1 to 15.

For each population, the number of alleles, allelic richness (Ar), observed heterozygosity (Ho), gene diversity (He), and the inbreeding coefficient (Fis) were estimated in the diveRsity R package (Keenan et al., 2013). Allelic richness is calculated in the diveRsity package using 1000 re-sampling with a n equal to the smallest sample size across population (n = 5 in this case).

Isolation by distance (IBD) was tested at a global level and inside of each main genetic clusters identified by STRUCTURE. Pairwise F_{ST} values were estimated with the package 'hierfstat' (Goudet & Jombart, 2018) in R (R Core Team, 2019) as the Weir and Cockerham unbiased parameter θ (Weir and Cockerham 1984). Since the 27 sampled locations are distributed across a large continental extent, pairwise geographical distances were calculated with the 'geosphere' package (Hijmans, 2017) in R (R Core Team, 2019) to account for the curvature of the Earth. The strength of the IDB was evaluated as the relationship between $\theta/(1-\theta)$ and the natural logarithm of the geographic distance as described by Rousset (1997). The significance of the IBD pattern was assessed by Mantel tests as implemented in the 'ade4' package (Dray and Dufour 2007) in R (R Core Team, 2019), with default parameters.

Spatial patterns of population differentiation were also analysed as a spatial variation of the expected heterozygosity and by a principal coordinate analysis (PcoA) (Legendre and Legendre 1998) on a pairwise distance matrix of θ values (and not $\theta/(1-\theta)$). The spatial pattern of θ as measured by the first axis of the PCoA was compared to that of the expected heterozygosity by means of a Spearman rank correlation test.

Results

Clustering analysis with STRUCTURE was not conclusive about the value of *K*. The Evanno's delta *K* method and the probabilities ln P(*X*|*K*) suggest two genetic clusters (**Figure 2.6**). Based on the probabilities of assignment for a *K* = 2 (**Figure 2.7**), the STRUCTURE results indicates the genetic proximity between populations from the same regions, notably the similar probabilities of the Southern cluster composed of most of the Czech Republic, the Austrian populations, Italy and Southern France populations that differentiate itself from the Northern cluster composed of all Scandinavian

populations and those from the United Kingdom and Ireland. This pattern can also be clearly identified when considering a *K* from 3 to 4 (**Figure 2.8**). When considering a *K* = 5 or greater, this differentiation between the two groups is less clear and the pattern is no more perceivable after a *K* = 6 or greater (**Figure 2.8**). Interestingly, the Estonian population (EST_1) and one of the Czech Republican populations (CZE_9) were most of the time placed close to the geographically more distant clusters (**Figure 2.7** and **Figure 2.8**). As the value of *K* becomes more important, the genetic similarity between the two Alpine populations (ITA_1 and AUT_1) becomes evident (**Figure 2.8**).

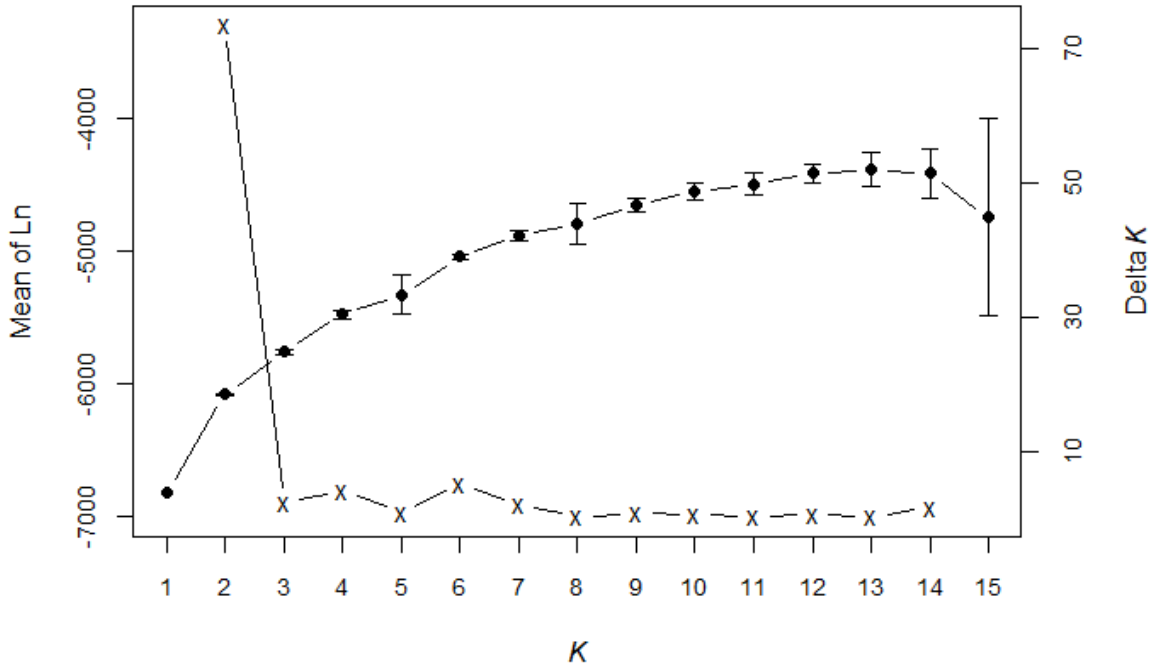


Figure 2.6. Probabilities ln P(*X*|*K*) (points) and delta *K* (crosses, Evanno et al., 2005) for each value of *K*.

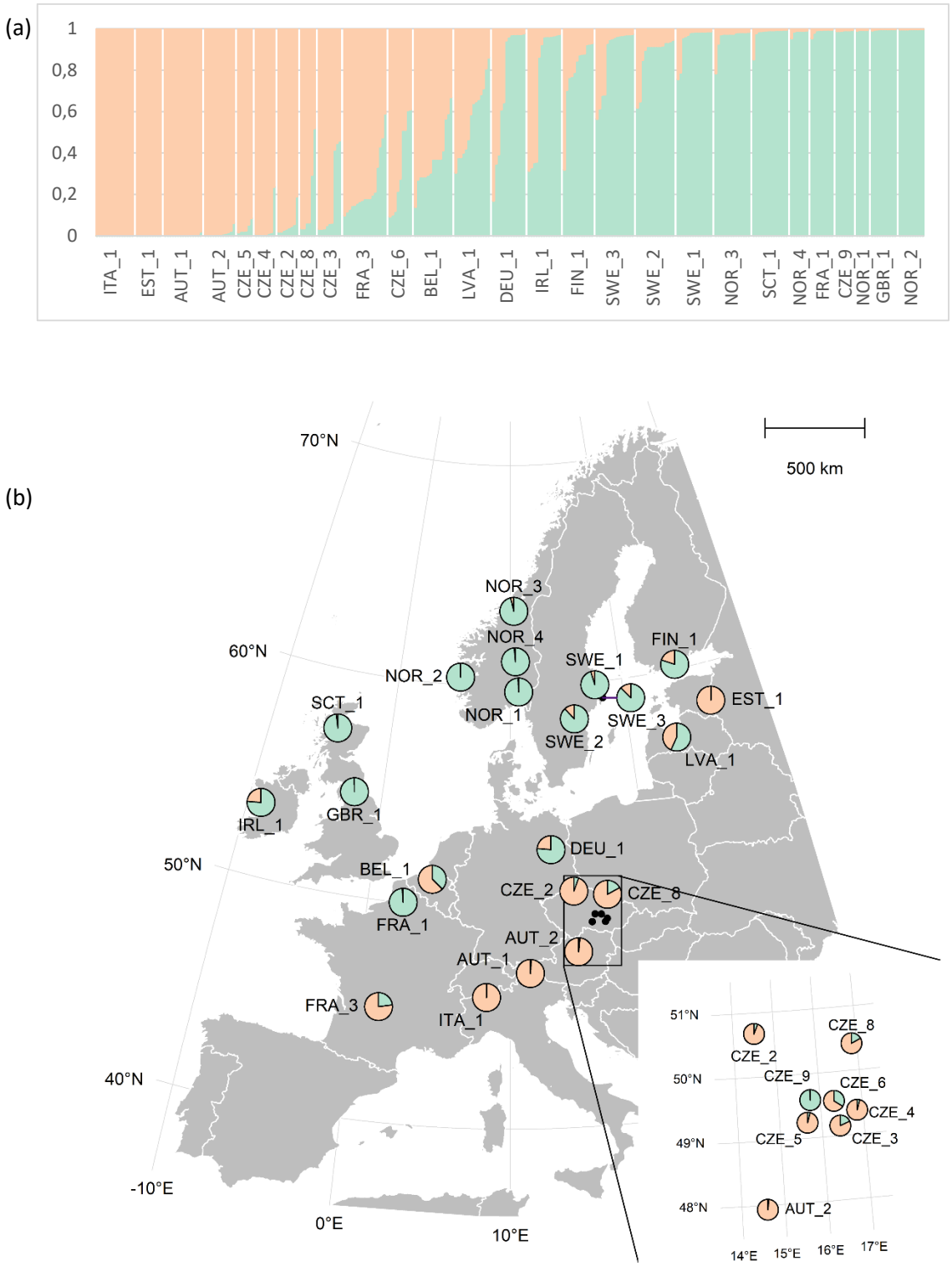


Figure 2.7. Probabilities of assignment to each of the two genetic clusters (*K* = 2) inferred from STRUCTURE for *Geum urbanum* in European populations. (a) Individual probabilities of assignment. (b) Geographic distribution. The pie charts in (b) in each population represents the overall probabilities for that population to be assigned to one of the two clusters (light green: northern cluster; light orange: southern cluster).

Table 2-1. Basic statistics for each population of *Geum. urbanum*. N, number of samples. Ar, Allelic richness. Ho, observed heterozygosity. He, gene diversity. Fis, inbreeding coefficient.

	N	Ar	Ho	He	Fis
ITA_1	15	2.12	0.26	0.42	0.3867
EST_1	10	1.5	0.22	0.18	-0.2037
AUT_1	15	2.23	0.36	0.42	0.1519
AUT_2	12	3.21	0.33	0.6	0.4462
CZE_5	6	3.01	0.36	0.56	0.358
CZE_4	8	2.45	0.15	0.43	0.6637
CZE_2	8	2.91	0.29	0.61	0.5254
CZE_8	6	3.03	0.22	0.63	0.6471
CZE_3	9	3.34	0.39	0.63	0.3874
FRA_3	17	3.04	0.36	0.59	0.3866
CZE_6	9	2.78	0.35	0.54	0.3436
BEL_1	15	2.27	0.24	0.39	0.372
LVA_1	14	3.67	0.51	0.64	0.2037
DEU_1	13	2.66	0.26	0.53	0.5149
IRL_1	13	2.89	0.36	0.6	0.4052
FIN_1	12	2.66	0.44	0.51	0.1213
SWE_3	15	3.26	0.32	0.66	0.5129
SWE_2	15	2.33	0.4	0.44	0.0985
SWE_1	14	2.86	0.2	0.59	0.6566
NOR_3	14	1.98	0.19	0.3	0.3554
SCT_1	14	2.87	0.35	0.52	0.3311
NOR_4	7	2.16	0.21	0.37	0.4247
FRA_1	9	2.24	0.28	0.38	0.2722
CZE_9	7	2.2	0.29	0.43	0.328
NOR_1	5	1.88	0.23	0.29	0.2045
GBR_1	10	2.09	0.28	0.39	0.2672
NOR_2	10	1.38	0.13	0.18	0.2727

This pattern of general differentiation between Northern and Southern populations can also be observed on the first principal plan of the PCoA (**Figure 2.9** and **Figure 2.10**a). The first axis of the PCoA based on the distance matrix of pairwise θ follows a somewhat north-south gradient, sub-grouping the 27 populations in a way similar, even though not identical, to that of STRUCTURE. However, this spatial pattern was not reproduced by the distribution of the expected heterozygosity, although the Czech Republic cluster can yet be identified (**Figure 2.10**b). The two spatial patterns were actually negatively correlated (ρ = -0.4027, p = 0.0373).

With few exceptions, values of He and F_{IS} were relatively high across populations (**Table 2-1**). Concerning the last, a notable exception was the Estonian population EST_1, the only population to exhibit important negative values of F_{IS}. Allelic richness was generally lower in the populations assigned to the Northern cluster, besides having more samples per population, on average.

A pattern of IBD is observed in the global analyse across all sampled populations (Mantel r = 0.1964, p = 0.006). A relatively strong signature of IBD was observed in the Southern group (K=2, Mantel r = 0.4659, p = 0.022), but no such pattern was observed in the Northern group (Mantel r = -0.1113, p = 0.740). This intra-group analysis was repeated with the exclusion of the two geographically isolated populations in each group (EST_1 in the Southern group, and CZE_9 in the Northern group). This time, no significance was observed in neither of the two groups (Southern: Mantel r = 0.2840, p = 0.122; Northern: Mantel r = -0.1144, p = 0.744). As an exercise to better understand the equilibrium/drift dynamics across the European continent, the Northern group was again tested against IBD, this time including the Estonian population EST_1 considering its geographic proximity to the remaining populations within this genetic cluster. Again, no significance was observed, but the slope of the regression became slightly positive (Mantel r = 0.0606, p = 0.391).

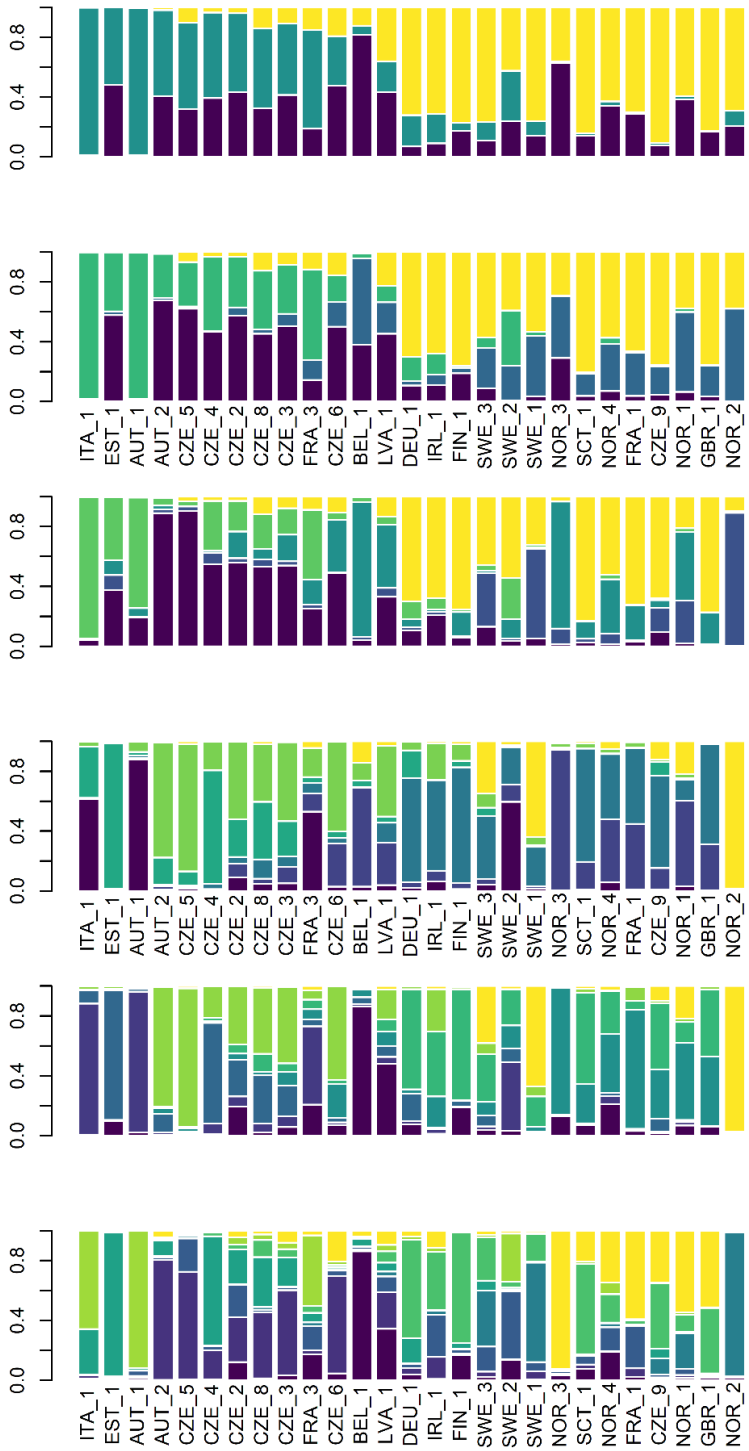


Figure 2.8. Genetic structure of *Geum urbanum* populations: population probabilities of assignment for each value of *K* from 3 (upper graphic) to 8 (lower graphic). Until *K* = 5, there is a general differentiation between the Northern populations (all populations on the left of LVA_1) and Southern populations (on the right of LVA_1). This pattern becomes less clear with *K* values of 6 or greater. For all values of *K*, populations from Czech Republic seem to form a somewhat concise group (with the exception of CZE_9), as do populations from Scandinavia, United Kingdom and Ireland. As the value of *K* becomes more important, the genetic similarity between the two Alpine populations (ITA_1 and AUT_1) becomes evident.

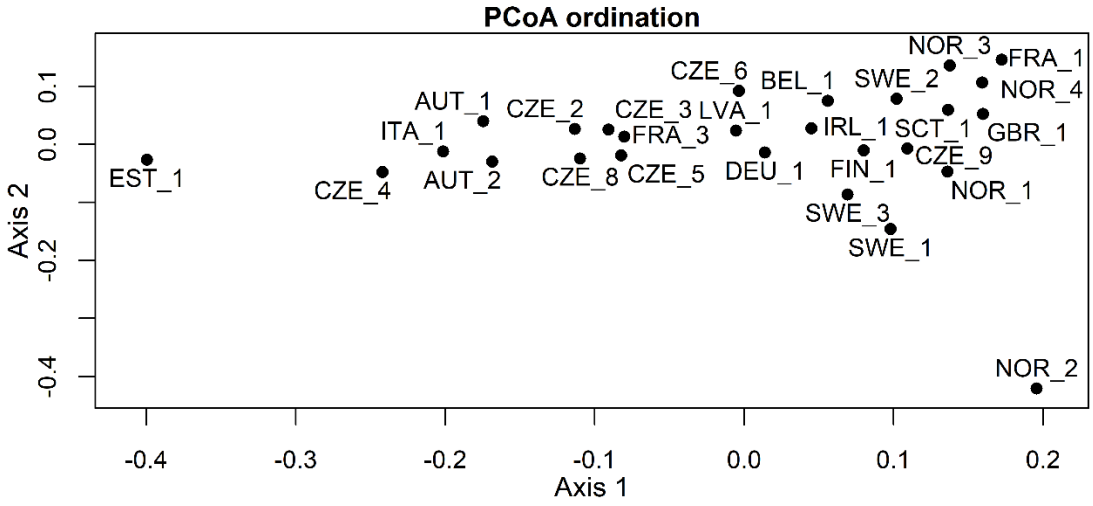


Figure 2.9. Biplot of the two first coordinates of the PCoA. The first coordinates differentiate Northern (right side) from Southern populations (left side) of *Geum. urbanum*. The results of the PCoA are very similar to those from STRUCTURE (Figs. 3 and 4). As for the STRUCTURE results, populations EST_1 and CZE_9, situated at the geographic North and South of the study area, respectively, are grouped with populations from the opposite geographic zone.

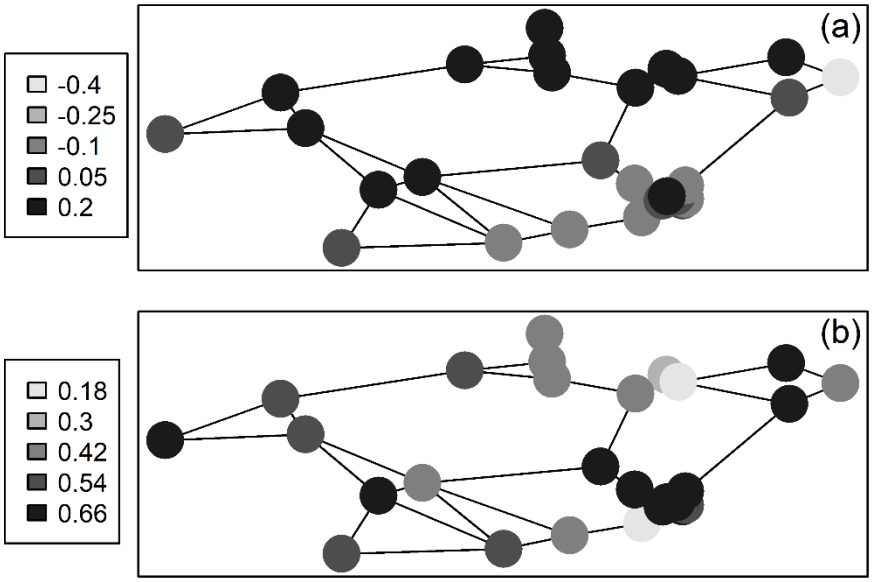


Figure 2.10. Spatial patterns of the first principal coordinates of the PCoA based on the pairwise values of θ (a) and the expected heterozygosity calculated for each population of *G. urbanum* (b). The PCoA pattern in (a) closely reproduces the results from STRUCTURE (note the high and low values of CZE_9 and EST_1, respectively), while expected heterozygosity has no clear spatial pattern and is inversely correlated to the first PCoA coordinate (ρ = -0.4027, p = 0.0373). A Gabriel graph was draw to better represent the two-dimensions geographic relationship between populations.

Discussion

Despite being a generalist species with a high dispersal ability, *G. urbanum* shows a marked population structure across its European range. Two spatially explicit genetic groups were identified both by Bayesian clustering approach (STRUCTURE) and multidimensional scaling (PCoA), which generally coincides with a geographic north/south gradient. A finer genetic structure becomes more evident when we analyse in more details the results obtained when increasing the values of possible *K* genetic clusters and the continuous variation of the first PCoA axis.

Within the Northern cluster, populations from the United Kingdom and Scandinavia – particularly those from Norway – seem to be closely related besides the presence of the North Sea. A possible reason for this proximity is that birds migrating from one region to the other could carry seeds attached to their feathers. Migratory birds are well known to disperse a vast range of taxon across long distances (Coughlan et al., 2017), including crustaceans (Rachalewski et al., 2013), acarians (Røed et al., 2016), snails (van Leeuwen & van der Velde, 2012), and, of course, plants (Mellado and Zamora 2014, Green 2016). Seeds of *G. urbanum* are reported to be highly viable, with rates of germination higher than 70% in variable temperature and radiation conditions in laboratory (Taylor 1997), and the periodic exchange of few seeds between the two regions could be enough to ensure the observed genetic proximity.

The Italian and the Southernmost Austrian populations (AUT_1) were particularly close to each other, which was brought to light by both STRUCTURE and the PCoA analysis. Both populations are situated in the Alpine regions, and it is possible that the mountain chain represents a somewhat important barrier to the gene flow with the two other closest populations, notably the one in South France (FRA_3) and the one in North Austria (AUT_2).

In face of the results of the global genetic structure of the European populations of *G. urbanum*, it was surprising that one of the Czech Republic and the Estonian populations (CZE_9 and EST_1, respectively) were grouped together with populations from distant geographic regions. The two

populations were amplified on the same plate but were genotyped in different runs. During the pick assignment step of the microsatellite genotyping, no particularity was observed, as picks were in general clearly identifiable. This leads to the conclusion that the observed genetic pattern is not a methodological artefact during the lab analyses. Analysed sample sizes were of n = 9 (CZE_9) and n = 12 (EST_1). It is possible, although unlikely, that the observed pattern is simply due to chance. Another possible explanation is the human factor: seeds could be transported from one region to another by human activities. But in that case and considering the geographical extent of the study area, a mixed pattern would most probably be observed between other populations. At this point, it is hard to find a clear explanation for the genetic relationship of those two populations within the global genetic structure across the studied region.

Although populations close to each other in the geographic space tended to show similar values of gene diversity, no global clear pattern was identified across all populations. In general, gene diversity was high, with a mean of 0.4752 and a median of 0.5100. Relatively high values were also observed by Schmidt et al. (2009) in Germany and Switzerland (not investigated in the present study), although the values they found were smaller than the one we observed here. Again, the Estonian population (EST_1) was a remarkable exception, where gene diversity was the lowest (0.19), which was also identified by Schmidt et al. (2009). The relatively high levels of F_{IS} suggests a high level of inbreeding, which might be linked to the common self-pollinating strategy of *G. urbanum*.

An IBD pattern was observed across Europe, but this pattern tends to disappear when analysed separately within each of the two identified genetic groups. At first glance, this global pattern suggests that, at the European scale, the populations are in equilibrium under dispersal and genetic drift (Slatkin, 1993), but investigating the within group patterns of IBD may tell a different history. No IBD pattern was observed within the Northern cluster and the weak IDB pattern found for the Southern cluster was not statistically significant when removing the Estonian population (EST_1). The within cluster absence of IBD could be explained by the species preferred self-pollination strategy. Self-

fertilizing species tend to experience a weaker selection pressure due to its reduced Effective Population Size *Ne* (Hartfield *et al.* 2017). In this sense, local selective pressures would be less important to population differentiation. Also, the genetic drift tends to be more important in self-fertilizing species (Nordborg and Donnelly 1997). Although many ecological and evolutionary processes may influence spatial genetic patterns, if the increase in the genetic drift is strong enough to overcome gene flow, it is expected that no IBD pattern will be observed (Slatkin, 1993). The within cluster patterns also suggest that some populations are somewhat isolated from other populations from within the same genetic cluster. That seems to be the case, for example, for the population NOR_2 in the Northern group. The second coordinate from the PCoA analysis places this population as very isolated from the remaining populations in the cluster. When considering multiple values of *K*, the STRUCTURE analysis also differentiates this population. It is also the second population with the lowest level of gene diversity, behind the Estonian population. Those results suggest that geography (physical barriers), ecology (impeding abiotic conditions in-between populations), or history (colonisation time or even contribution of hybrids of *G. urbanum X rivale* on the overall local population genetics) could contribute to explain the isolation of certain populations that can't be acknowledged with the methods presented here. The ecological factors potentially contributing to the observed genetic structure will be further explored in the next chapter.

Finally, since there is no evident physical geographic barrier between the Northern and Southern clusters we identified for *G. urbanum*, the continental genetic structure in two distinct clusters could be explained by a post-Pleistocene northward expansion. Founder-flush theory (Slatkin, 1996) preconizes that a rapidly growing founding population would experience a relaxed selection due to reduced ecological pressure, and so genetic drift and recombination could fix different alleles or new combination of alleles from the original population. Founder events have been proposed to decrease the genetic diversity of populations (Nei *et al.* 1975), and allelic richness was generally lower in the Northern populations. Again, this hypothesis of a post-Pleistocene expansion explaining the continental genetic structure of *G. urbanum* will be further explored in the next chapter.

Ticks and Tick-borne Diseases 11 (2020) 101509

Contents lists available at ScienceDirect

Ticks and Tick-borne Diseases

journal homepage: www.elsevier.com/locate/ttbdis

Original article

Strong genetic structure among populations of the tick *Ixodes ricinus* across its range

Pedro Poli[a,*], Jonathan Lenoir[a], Olivier Plantard[b], Steffen Ehrmann[c], Knut H. Røed[d], Hans Petter Leinaas[e], Marcus Panning[f], Annie Guiller[a]

[a] Université de Picardie Jules Verne, UMR « Ecologie et Dynamique des Systèmes Anthropisés » (EDYSAN, UMR 7058 CNRS), 33 Rue Saint Leu, 80000 Amiens CEDEX 1, France
[b] BIOEPAR, INRAE, Oniris, 44307, Nantes, France
[c] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany
[d] Department of Basic Sciences and Aquatic Medicine, Norwegian University of Life Sciences, N-0033, Oslo, Norway
[e] Department of Biosciences, University of Oslo, Box 1066 Blindern, N-0316 Oslo, Norway
[f] Institute of Virology, Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, Hermann-Herder-Str, 11 79104, Freiburg, Germany

ARTICLE INFO

Keywords:
Gene flow
Infection risks
Range shift

ABSTRACT

Ixodes ricinus is the most common and widely distributed tick species in Europe, responsible for several zoonotic diseases, including Lyme borreliosis. Population genetics of disease vectors is a useful tool for understanding the spread of pathogens and infection risks. Despite the threat to the public health due to the climate-driven distribution changes of *I. ricinus*, the genetic structure of tick populations, though essential for understanding epidemiology, remains unclear. Previous studies have demonstrated weak to no apparent spatial pattern of genetic differentiation between European populations. Here, we analysed the population genetic structure of 497 individuals from 28 tick populations sampled from 20 countries across Europe, the Middle-East, and northern Africa. We analysed 125 SNPs loci after quality control. We ran Bayesian and multivariate hierarchical clustering analyses to identify and describe clusters of genetically related individuals. Both clustering methods support the identification of three spatially-structured clusters. Individuals from the south and north-western parts of Eurasia form a separated cluster from northern European populations, while central European populations are a mix between the two groups. Our findings have important implications for understanding the dispersal processes that shape the spread of zoonotic diseases under anthropogenic global changes.

1. Introduction

Ixodes ricinus (Acari, Ixodidae) is the most widespread tick species occurring across Europe and an important vector of multiple tick-borne diseases, both to humans and livestock. Commonly reported pathogens transmitted by *I. ricinus* include: bacterium *Borrelia burgdorferi*sensu lato, responsible for the Lyme borreliosis, which is the most prevalent tick-borne disease in temperate Europe (ECDC, 2015); arboviruses (genus *Flavivirus*) causing tick-borne encephalitis (TBE) and louping-ill (LI); the protozoan *Babesia microti*, responsible for the babesiosis; and the bacterium *Neoehrlichia mikurensis*, responsible for neoehrlichiosis, an emerging tick-borne pathogen (Portillo et al., 2018; Welinder-Olsson et al., 2010).

The current climate-driven redistribution of hematophagous arthropods such as ticks and mosquitoes may lead to severe challenges to public health and husbandry, by carrying a wide range of vector-borne diseases to new areas (Dantas-Torres, 2015; Pecl et al., 2017). For instance, many studies have demonstrated that the range of *I. ricinus*is already shifting northward and to higher elevations (e.g. Hvidsten et al., 2020; Jore et al., 2011; Lindgren and Gustafson, 2001) and those shifts are expected to continue in the future (Alkishe et al., 2017; Medlock et al., 2013).

Despite the threats of emerging infectious diseases following the redistribution of *I. ricinus*, little is known about the genetic structure of tick populations across the entire species range. Population genetic differentiation and spatial structuring can, however, impact the vector fitness and distribution, and therefore disease transmission (Blanchong et al., 2016; Wonham et al., 2006). Population genetics approaches such as individual genetic clustering and assignment methods enable inference on migrants (exchange of genes between populations) and the

* Corresponding author.
E-mail addresses: pvpoli@gmail.com (P. Poli), annie.guiller@u-picardie.fr (A. Guiller).

https://doi.org/10.1016/j.ttbdis.2020.101509
Received 16 April 2020; Received in revised form 3 July 2020; Accepted 6 July 2020
Available online 07 July 2020

risk of pathogen spread between populations (Kozakiewicz et al., 2018). For example, Lang and Blanchong (2012) applied clustering and distance-based methods to assess gene flow and disease spread risk between populations of white-tailed deer in the USA. Similarly, Van Zee et al. (2015) identified different genetic clusters between the southern and northern range of the tick *Ixodes scapularis* while the prevalence of borreliosis is known to be lower in the southern range. The authors suggest that this pattern of spatial genetic structure might be linked to differences in questing behaviour as ticks from the northern range would be more likely to bite humans. Differences in several life history traits of *I. ricinus* – such as the temperature at which nymphs begin to quest – have been reported along a latitudinal gradient (Gilbert et al., 2014), suggesting a spatially explicit phenotypic plasticity or adaptation. Yet, such basic knowledge about the distribution of genetic variation in *I. ricinus* and the migration processes involved in disease transmission remain largely unknown, albeit being essential to design better vector control strategies (Araya-Anchetta et al., 2015; Gooding, 1996; Tabachnick and Black, 1995).

The genetic structure of parasites' populations is known to be influenced by the distribution of the hosts (Kempf et al., 2009; Wessels et al., 2019). In general, it is assumed that generalist parasites relying on a wide range of hosts tend to show weak or no genetic structure, as shown in many studies on various parasite species (e.g. Archie and Ezenwa, 2011; Wessels et al., 2019). The tick species *I. ricinus* is a generalist ectoparasite infesting a wide range of hosts, such as reptiles, mammals, and birds (Casati et al., 2008; Norte et al., 2012). It has been proposed that tick abundance and population genetic structure are dependent on the species' biology (such as reproduction strategies and life cycle), but also on the host distribution and behaviour (Kempf et al., 2011; McCoy et al., 2001; Rizzoli et al., 2009; Norte et al., 2012). Large ungulates, such as deer, bovidae, and wild boar may be highly efficient carriers of ticks for long distances, as long as there are no severe barriers to their migration (Handeland et al., 2013; Kriz et al., 2014). By contrast, transportation of ticks by migrating birds seems to be less efficient across contiguous landmasses (Hasle et al., 2009; Røed et al., 2016). Based on these findings, it is expected that *I. ricinus* populations should show a weak spatial genetic structure.

Regarding previous works on population structure and dispersal of *I. ricinus*, Noureddine et al. (2011) found a clear differentiation between European and African populations using sequences from three nuclear and three mitochondrial markers. Regarding the results from that study, it was later suggested by Estrada-Peña et al. (2014) that those northern African samples could correspond to *Ixodes inopinatus*, a sibling species of the *I. ricinus* complex within the *Ixodes* subgenus. Considering only European populations, some studies showed weak to no differentiation, but an extensive genetic diversity was observed within each local population (Casati et al., 2008; Noureddine et al., 2011; Porretta et al., 2013; Carpi et al., 2016). Other investigations analysing the frequency of mitochondrial haplotypes showed a marked phylogeographical structure in northern Europe, notably when considering populations from the north of the UK (Scotland) and Scandinavia (Al-Khafaji et al., 2019; Dinnis et al., 2014; Røed et al., 2016). Although none of the mitochondrial haplotypes was exclusive to any of those populations, their frequencies varied significantly between populations from different regions. Interestingly, the British clade identified by Røed et al. (2016) coincides with the occurrence of a particular subtype of the louping-ill virus, which is closely related to other Irish and Spanish subtypes. Other studies focusing on the genetic structure of *I. ricinus* populations were based on microsatellite loci (Kempf et al., 2009, 2011). Microsatellite variations have led to the identification of significant levels of genetic structure at different spatial scales, deviation from panmixia in *I. ricinus* populations likely due to assortative mating and patterns of host use (see Araya-Anchetta et al., 2015 for a review). However, those studies have also assessed patterns of genetic variation from localised samples that cover only a subset of the species range and thus likely do not capture the entire species genetic structure at the continental level.

Here, we aim to elucidate the population genetic structure of the tick *I. ricinus* based on single nucleotide polymorphisms (SNPs). To the best of our knowledge, no other study on the population genetic structure of *I. ricinus* throughout the Eurasian continent was based on the variation detected by this type of marker. Although generally having a weaker mutation rate than microsatellites, SNPs offers the possibility of building a larger range of markers and have been suggested to be more reliable markers for population genetic studies (Helyar et al., 2011; Smouse, 2010). Our main objective is to describe the genetic structure of *I. ricinus* populations to infer the geographical and environmental factors shaping this structure. Particularly we hypothesized that (i) *I. ricinus* from the western parts of Europe might have genetic similarities to the Great Britain lineage (Røed et al., 2016) while (ii) there should be a pronounced genetic differentiation between ticks south and north of the extensive mountain areas covering central Europe (i.e., the Eastern Alps, the Western Alps, the Carpathian Mountains, and the Balkan Mountains).

2. Materials and methods

2.1. Sampling

A total of 28 tick populations from 20 countries were sampled covering most of the species' range, including populations close to the northern (Norwegian, Sweden, Ireland, and England) and southern (Iran, Spain, and northern Africa) range limit of *I. ricinus* (Fig. 1). Samples were collected by flagging inside or near forest fragments from the ground vegetation and were preserved in alcohol. A significant subset of the sampled populations we used, covering 8 regions across Europe (southern and northern France; Belgium; western and eastern German; southern and central Sweden; and northern Estonia), originated from a single project (smallFOREST, BiodivERsA 2010–2011 Joint call: https://www.biodiversa.org/491/download) and was sampled by the same person during the same year 2013 (See Ehrmann et al., 2017 for details). The remaining samples were collected for different projects (for details on those projects see Røed et al., 2016 for the Norwegian samples and Noureddine et al., 2011 for the remaining samples). The coordinates of the sampled populations are provided in Table S1 (see Supporting information). Aside from smallFOREST samples, sampling dates varied among the sampled populations (Table S1).

Ticks sampled for those projects were identified at the laboratory using standard morphological keys provided in Babos (1964); Hillyard (1996), or Pérez-Eid (2007). As most samples we used were identified before the description of *I. inopinatus* (Estrada-Peña et al., 2014) and considering that it was impossible to re-evaluate the identification of samples based on morphological features, we conducted an *a posteriori* evaluation of the potential presence of *I. inopinatus*among our samples. To fulfil this aim, northern African *I. inopinatus*samples analysed by Noureddine et al. (2011) were included in the present study.

2.2. DNA extraction and SNP genotyping

Since ticks and DNA samples analysed in this study had different origins and therefore different storage methods, three different methods were used to ensure DNA extraction. Ticks were either: (i) frozen and crushed with a pestle in individual tubes before extracting DNA using DNeasyTM Tissue Kit (Qiagen); (ii) disrupted using a Tissue Lyser (Qiagen) before DNA extraction using the Wizard Genomics DNA Purification Kit (Promega, USA); or (iii) crushed with Lysing matrix H (MP Biomedicals, Santa Ana, USA) before extracting DNA with MagNA Pure LC Total Nucleic Acid Isolation Kit (Roche, Basel, Switzerland).

We genotyped 192 SNPs as described by Quillery et al. (2014). The list of SNPs, variant basis, and primers are presented in Table S2. All samples were amplified by whole genome amplification (WGA) before genotyping. The PEP-PCR WGA kit (LGC-Biosearch Technologies) was

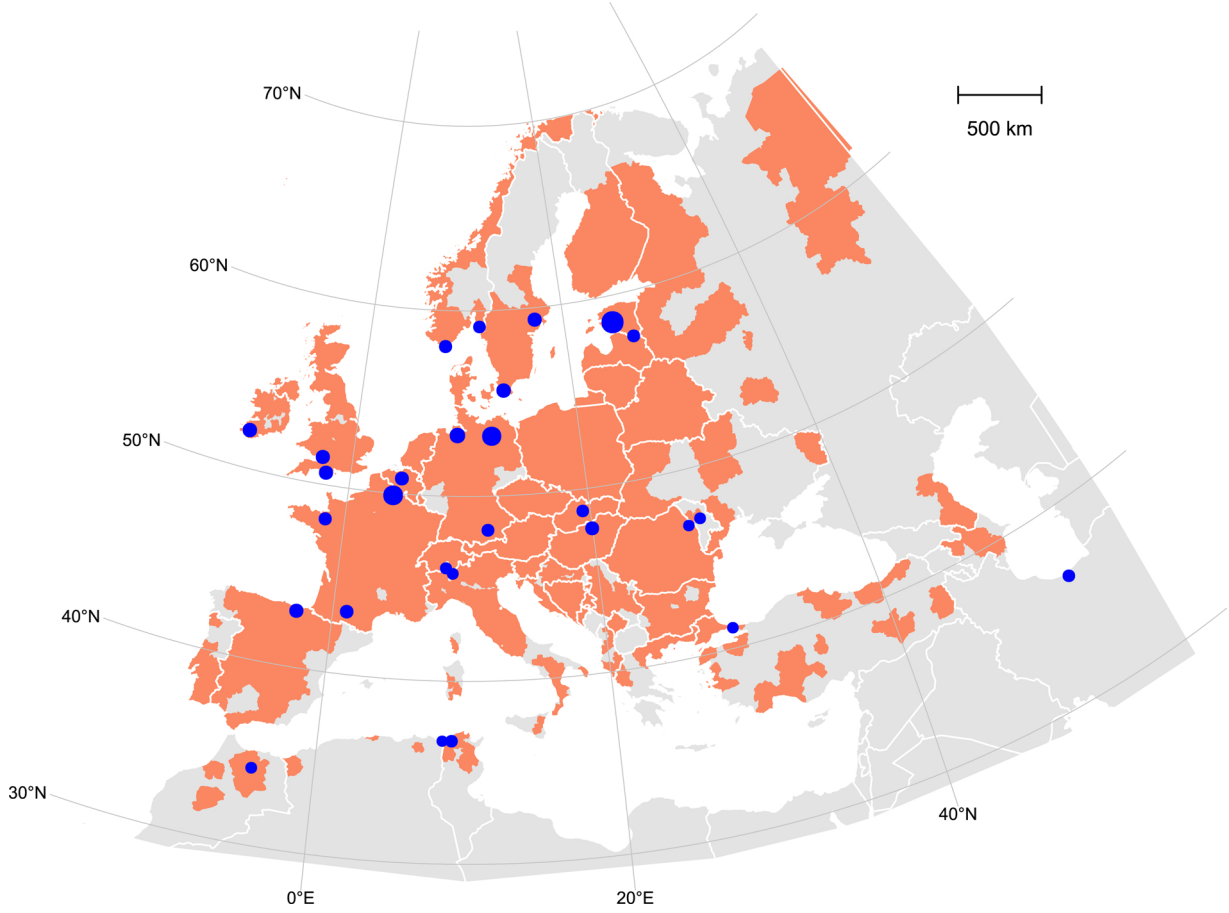


Fig. 1. Distribution of the sampled populations of *Ixodes ricinus* across its putative range. The range of *I. ricinus* is displayed in dark orange on the map and was adapted from the European Centre for Disease Prevention and Control – ECDC (January 2019). The size of each blue dot on the map is proportional to the sample size of each sampled population. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

used for whole genome amplification of each sample. The WGA protocol associated with KASP genotyping has already been tested by Quillery et al. (2014) and showed a reduced number of "no-call" data (missing values) during genotyping. The WGA and genotyping steps were subcontracted by the GENTYANE platform (INRA, Clermont-Ferrand, France: http://gentyane.clermont.inra.fr/). The GENTYANE platform is an INRAE (French National Research Institute for Agriculture, Food and Environment) research facility located in Clermont-Ferrand (France) which offers sequencing and genotyping services. Genotyping was conducted in a Biomark HD System (Fluidigm) and KASPar assays. The KASPar method is a KBiosciences competitive allele-specific PCR amplification. A PCR mix containing two allele-specific forward primers and one common reverse primer was carried out. Each forward primer had a 5′ tail sequence homologous to universal secondary oligos labelled with a fluorophore (FAM or HEX). If a particular locus is homozygous, only one fluorescent signal is generated. Bi-allelic loci generate both fluorescent signals.

2.3. Quality control

Data was filtered after genotyping and before statistical analysis. First, all invariant SNPs were removed. After this first filtering step, all individuals with more than 20 % of non-amplified sites (missing data) were removed. Finally, all remaining SNPs with more than 20 % missing data were also removed. The remaining dataset consisted of both individuals and SNPs with less than 20 % missing data. After quality control steps, 125 SNP loci and 497 individuals were kept for further analyses.

2.4. Cluster analysis and genetic structure

Two complementary clustering methods were used to access the genetic structure of *I. ricinus* populations. First, we investigated the genetic clustering by performing a discriminant analysis of principal components (DAPC, Jombart et al., 2010) with the package 'adegenet' (Jombart, 2008) in R (R Core Team, 2019). The optimal k number of clusters was identified by the k-means algorithm using the find.cluster() function based on BIC values. A maximum of 28 clusters was allowed, i.e. the total number of sampled populations. Next we performed a Bayesian analysis in STRUCTURE (Pritchard et al., 2000) with the parameter K, i.e. the optimal number of clusters, varying from 1 to 10, according to the results from the DAPC. We used a non-admixture model with the sampling locations as prior. Twenty repetitions of 80,000 MCMC iterations with a burning length of 20,000 iterations were run for each value of K. The results were analysed with Structure Harvester (Earl and vonHoldt, 2012). The best K value for the optimal number of clusters was identified by comparing the estimates of log probabilities of the data (i.e. ln[Pr(X|K)]) for each K value as well as Evanno's delta K method (Evanno et al., 2005). Pritchard et al. (2007) suggested aiming for the smallest value of K that captures most of the genetic structure in the data. Assigning probabilities for individuals and populations across repetitions were then averaged in CLUMPP (Jakobsson and Rosenberg, 2007). We applied a hierarchical clustering analysis (e.g. Vähä et al., 2007) in each identified cluster to detect more refined patterns of genetic structure. Hierarchical analysis in STRUCTURE was realised with ten repetitions and the same other parameters as the first round of analysis. We realised a similar analysis for each cluster identified by DAPC.

To test our data for isolation by distance (IBD), pairwise F_{ST} values were estimated with the package 'hierfstat' (Goudet and Jombart, 2018) in R (R Core Team, 2019) as Weir and Cockerham unbiased parameter θ (Weir and Cockerham, 1984). The IBD pattern was first tested across all pairs of Eurasian samples and second only between pairs of samples collected during the same year to avoid potential biases due to temporal variability in dispersal and genetic structure. Those corresponded to samples from southern and northern France, Belgium, western and eastern German, northern Estonia, southern and central Sweden, a total of 8 samples (28 pairs). Since the 25 Eurasian samples are distributed across a large continental extent, pairwise geographical distances were calculated with the 'geosphere' package (Hijmans, 2017) in R (R Core Team, 2019) to account for the curvature of the Earth. The strength of the IDB was evaluated as the relationship between $\theta/(1-\theta)$ and the natural logarithm of the geographic distance as described by Rousset (1997). In a two dimensions population, the slope parameter b of the linear regression $\theta/(1-\theta) = a + bD_{Geo}$ is inversely proportional to the average neighbourhood size $Nb = 1/b$, and $b = 1/(4D_e\pi\sigma^2)$, where D_e is the sub-population density and σ^2 is the averaged square axial distances between adults and their parents and σ is half the average adult-parent distance (Séré et al., 2017). In this case, a proxy of dispersal can be calculated as $\delta \approx 2*\sqrt{(4\pi Deb)}$ (Manangwa et al., 2019). The population density was calculated as $D_e = N_e/S\pi$, where S is the smallest distance between sites considered and included in the IBD analysis. We used NeEstimator version 2.1 to calculate effective population sizes (Ne) by applying two different methods, one based on linkage disequilibrium and another based on molecular co-ancestry (Do et al., 2014). We calculated the mean of Ne estimated with these two methods after the exclusion of 'infinity' results. The obtained mean value was weighted by the number of times one of the two methods generated a non-infinity value. The significance of the IBD pattern was assessed by Mantel tests as implemented in the 'vegan' package (Oksanen et al., 2019) in R (R Core Team, 2019).

2.5. Genetic diversity

For each locus, we estimated the observed heterozygosity (Ho), the gene diversity (Hs), and Wright's fixation indices F_{IS}, F_{ST}, and F_{IT}. Wright's statistics measure inbreeding in three levels of population structure: F_{IS} is the inbreeding coefficient of individuals relative to subpopulations; F_{ST} is the inbreeding coefficient of subpopulations relative to populations; and F_{IT} is a measure of the inbreeding of individuals relative to populations. All metrics were calculated with the package 'hierfstat' (Goudet and Jombart, 2018) in R (R Core Team, 2019). A Monte-Carlo permutation test (999 replicates) was conducted to test for the significance of the differences of mean gene diversity and F_{IS} values over loci between pairs of genetic clusters identified. For each replicate, individuals were randomly assigned to one genetic cluster and the simulated statistics were calculated. We ran the *randtest()* function from the 'ade4' package (Dray and Dufour, 2007) to access the significance of the observed differences.

To investigate null alleles and possible Wahlund effect on genotype frequencies, we followed the procedure proposed by De Meeûs (2018). According to that study, the presence of null alleles could be identified by a suit of comparisons of F_{IS}, F_{ST}, and the number of missing data. In case of null alleles, we would observe: (i) a high positive correlation between F_{IS} and F_{ST}; (ii) high variation of both F_{IS} and F_{ST} across loci; (iii) F_{IS} standard errors (StrdErrFIS) much bigger than F_{ST} standard errors (StrdErrFst); and (iv) F_{IS} values mainly explained by the presence of missing data. For the Wahlund effect, the correlation between F_{IS} and F_{ST} should approximate zero, a small variation of F_{ST} and a moderate variation of F_{IS} should be observed across loci, F_{IS} standard errors (StrdErrFIS) should be higher than F_{ST} standard errors (StrdErrFst) and no or rare missing data should be obtained. To test those relations, values of F_{IS}, F_{ST}, StrdErrFst, and StrdErrFIS were calculated in the FSTAT software version 2.9.4 (Goudet, 2003), the latter values calculated by Jackknife. The Spearman's rank correlation test was applied to test for correlations. Finally, De Meeûs (2018) suggested a linear regression between F_{IS} and missing data to quantify, using the R^2 value, the contribution of missing data in F_{IS} values. Because the Wahlund effect can produce between-locus dependencies, we also tested linkage disequilibrium for each pair of loci by using G-based tests implemented in FSTAT 2.9.4. Since p-values from each test are not independent, we applied the procedure described by Benjamini and Yekutieli (2001) to calculate the false discovery rate (FDR) and correct p-values.

3. Results

3.1. Clustering analysis, genetic differentiation and isolation by distance

The DAPC analysis identified two possibilities for the number of clusters, one suggesting three different genetic clusters and the other suggesting four genetic clusters (the BIC difference is 0.842 between $K = 3$ and $K = 4$, Fig. S1). Choosing $K = 4$ clusters created two overlapping groups, while $K = 3$ grouped individuals into 3 well-separated clusters (Fig. 2). Hence, we decided to set the number of clusters to $K = 3$ with the DAPC approach. Bayesian analysis performed with STRUCTURE also identified a $K = 3$ differentiated genetic clusters (Figs. 2b and S2) whose compositions are very similar to the three clusters retained with the DAPC approach. In both analyses, northern African (yellow colour in Figs. 2 and 3) and Eurasian populations (the other colours) were highly differentiated. Two main groups were identified within Eurasia, one corresponding mainly to northern and continental middle European populations (grey colour in Figs. 2 and 3), the other corresponding mainly to southern and western populations in Eurasia (blue colour in Figs. 2 and 3). The DAPC approach separated northern African populations from Eurasian ones along the first axis, while Eurasian clusters were mostly separated along the second axis (Fig. 2a). Regarding clustering analyses with STRUCTURE, individual probabilities of different K values ranging from 2 to 10, excepted for $K = 3$ which is already depicted in Fig. 2b, are presented in the Supporting information (see Fig. S3).

Finer genetic structure was identified from our hierarchical analyses (Figs. S4 and S5). These analyses, either carried out with DAPC (Fig. S4) or the STRUCTURE approach (Fig. S5), were able to isolate Iran and/or Turkey from the other sampled sites within the southern Eurasian cluster. Atlantic sites (Spain, southern and western France, Ireland, and England) were further isolated from the remaining sites in this group (Italy, Romania, Hungry, and Slovakia). The northern European sites showed a more admixture structure, and separation in further clusters varied between the DAPC and STRUCTURE approaches (see the 'Hierarchical analysis' section in the Supplementary Information for more details).

A pattern of isolation by distance (IBD) was observed across all sampled populations (Mantel r = 0.726, $p < 0.001$). Restricting the IBD analysis to the set of sites sampled during the same year, we found an even stronger pattern of IDB (Mantel r = 0.870, $p < 0.0001$, Fig. 4). In the latter case, the coefficient estimate of the slope parameter (b) in the regression was $b = 0.01$ with a 95 % confidence interval (CI) ranging from 0.007 to 0.013. Neighbourhood size (Nb) reached $Nb = 99$ individuals, on average (95 % CI = [71–140]), and immigration rate (N_em) was estimated to reach $N_em = 16$ (95 % CI = [11–22]) individuals per generation and subpopulation.

We found a mean effective population size of 62 individuals. The closest sampled sites were North France and Belgium, separated 119 km from one another. We found surface and population densities to reach, on average, $S^2 = 11.3\,km^2$ and $De = 5.4$ individuals/m^2, respectively. We found the dispersal rate to reach, on average, $\delta \approx 76$ km/generation (95 % CI = [65–90]).

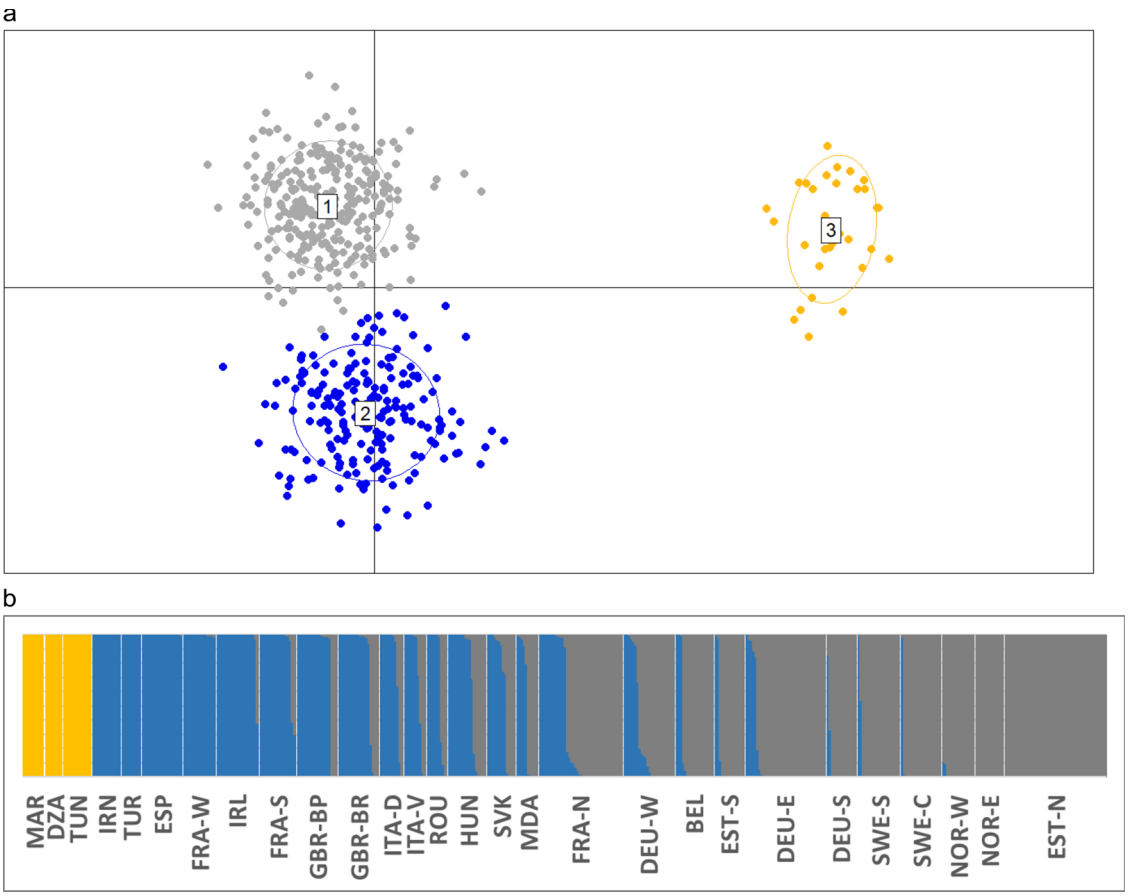


Fig. 2. Cluster assignment analysis results based on either the DAPC scatter plot of individual memberships for $K = 3$ (a) or the STRUCTURE individual membership probabilities for $K = 3$ as described by Evanno et al. (2005) (b).The sampled populations are coded as follows: MAR: Morocco; DZA: Algeria; TUN: Tunisia; ESP: Spain; IRN: Iran; TUR: Turkey; FRA-W: West France; IRL: Ireland; FRA-S: South France; GBR-BP: England Blue Pool; GBR-BR: England Bristol; ITA-D: Italy Domodossola; ITA-V: Italy Varese; ROU: Romania; HUN: Hungary; SVK: Slovakia; MDA: Moldavia; FRA-N: North France; DEU-W: West Germany; BEL: Belgium; EST-S: South Estonia: DEU-E: East Germany; DEU-S: South Germany; SWE-S: South Sweden; SWE-C: Central Sweden; NOR-So: Norway Søgne; NOR-Gr: Norway Grønnsundfjellet; EST-N: North Estonia. Coordinates of sampled populations are presented in Table S1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Genetic diversity

The observed heterozygosity (*Ho*), gene diversity (*Hs*), and F_{IS} were highly variable across loci (Table S3). The observed F_{ST} values were, however, more constant than F_{IS} ones. For most loci, gene diversity was higher than the observed heterozygosity. Consequently, the overall gene diversity across all loci was significantly higher than the observed heterozygosity (Wilcox Signed-Rank Test, V = 6959, $p < 0.0001$). The mean gene diversity per sampled population was still higher than the observed heterozygosity (Wilcox Signed-Rank Test, V = 406, $p < 0.0001$) and mean F_{IS} was always positive. Mean values of observed heterozygosity, gene diversity, and F_{IS} for each population are shown in Fig. S6 (Supporting information). The highest mean gene diversity and F_{IS} values over loci were identified in the southern Eurasian cluster (Hs = 0.355, $F_{IS} = 0.275$), followed by the northern European cluster (Hs = 0.340, $F_{IS} = 0.2708$) and the cluster from northern Africa (Hs = 0.171, $F_{IS} = 0.191$) (Fig. 5). The Monte-Carlo test showed a significant difference in gene diversity values for all pairs of clusters ($p = 0.001$ for all three comparisons), but none for F_{IS} values ($p = 0.199$ and 0.239 when comparing northern Africa to the northern European cluster and northern Africa to the southern Eurasian cluster, respectively; while $p = 0.644$ when comparing the southern Eurasian cluster to the northern European cluster). Populations from northern Africa showed a high deficit in heterozygosity, of which 71 out of 125 loci with *Hs* values of zero.

After *p*-value correction (Benjamini and Yekutieli, 2001), no pair of locus showed significance values of linkage disequilibrium. No correlation was found between F_{IS} and F_{ST} ($\rho = -0.0206$, $p = 0.8198$) and missing data were positively correlated to F_{IS} values ($\rho = 0.5804$, $p < 0.001$). The linear regression of F_{IS} against the number of missing data estimated an adjusted R^2 of 0.19, suggesting that around one-fifth of F_{IS} variance is explained by the number of missing data. Finally, StrdErrFIS was around 4 times bigger than StrdErrFst (0.033 and 0.008, respectively).

4. Discussion

We investigated the genetic structure of populations from the tick *I. ricinus* in much of its range, *i.e.* in Eurasia and in northern Africa. In addition to a strong and expected divergence between northern African and Eurasian populations, the two Eurasian genetic clusters described here showed clear spatial patterns. The isolation by distance patterns we found, either throughout the entire dataset or restricted to samples from the same period, suggest an association between the genetic structure of *I. ricinus* populations and the geographical location of these populations. Hierarchical analyses confirmed the genetic affinity between western European populations, from the UK and Ireland in the north to Spain in the south, supporting our first hypothesis regarding genetic similarities in western continental Europe and the British Isles. Also consistent with our second hypothesis stating a genetic signature of central European mountains, we found a clear differentiation between populations from southern Eurasia and populations from

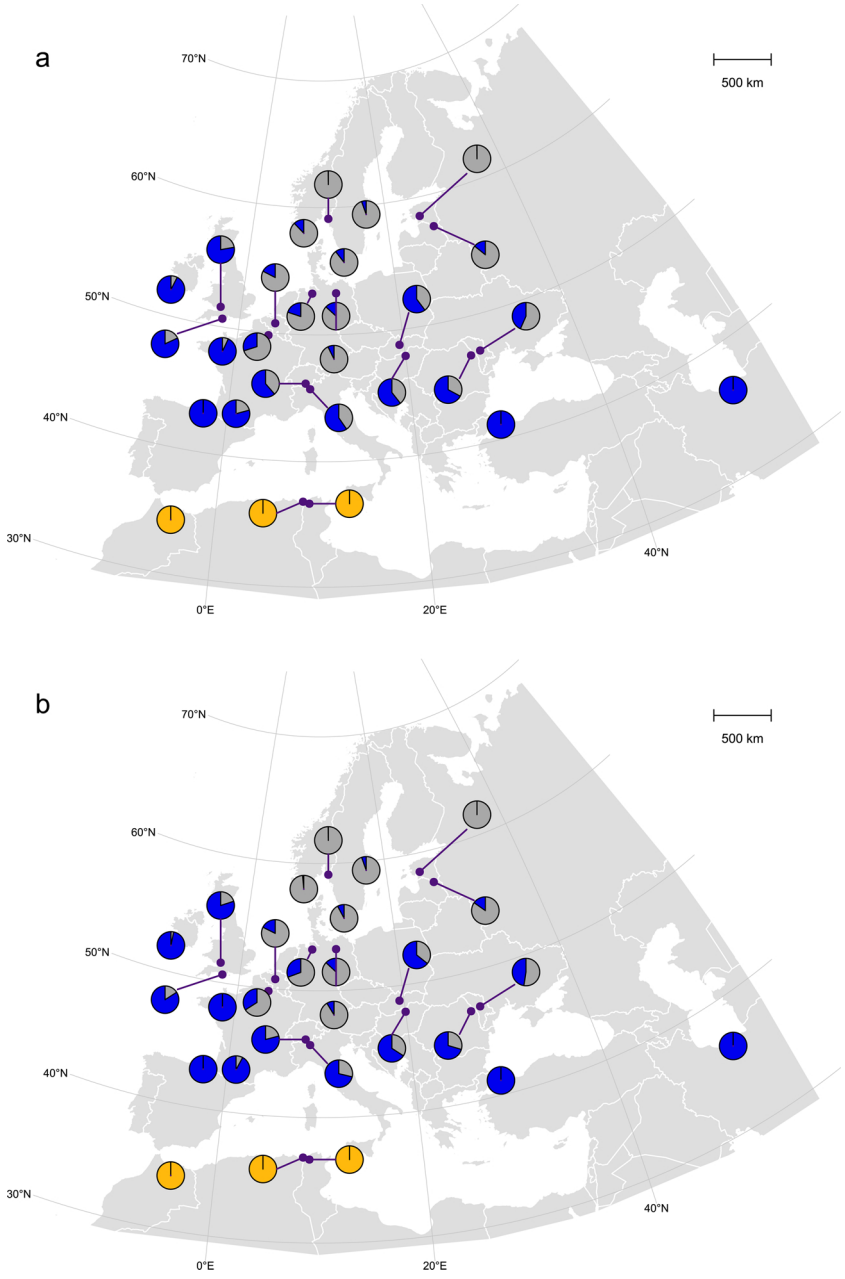


Fig. 3. Distribution of the relative importance of each cluster on each sampled population (see Fig. 2 for the groups which colors are matching). Results are provided for both the DAPC (a) and the STRUCTURE (b) analysis.

northern Europe. Indication of migration of individuals between the two clusters is suggested by the different degrees of affinity from central Europe with one cluster or another (e.g. in Romania, Hungary, Slovakia, and Moldova).

Ixodes ricinus and *I. inopinatus* have recently been suggested to be sympatric both in northern Africa (Younsi et al., 2020) and in Europe (Estrada-Peña et al., 2014; Chitimia-Dobler et al., 2018). Our results are clear concerning the genetic identity of northern African samples. According to the results from both the DAPC and STRUCTURE analysis, there is no possibility of any individuals from those populations to belong to any other genetic clusters. Also, no individual from Eurasia had any probability of identity with the northern African cluster. Converging results of both analyses indicate with a great deal of certitude that: (i) all samples from northern Africa belong to the same species and have the same ancestry; (ii) no sample from northern Africa share ancestry with those from Eurasia. Northern African samples were also a particular case as more than half loci were monomorphic across all three populations, which was not found in Eurasian populations. Again, it is important to note that individuals from the three northern African populations analysed here were identified before the description of *I. inopinatus* (Estrada-Peña et al., 2014). If *I. inopinatus* was present in the Eurasian samples, we would expect at least small probabilities of identity of Eurasian samples with the northern African cluster, which was not the case. The clear-cut genetic differentiation we obtained between Eurasian and northern African populations strongly suggests that all the individuals from the three northern African populations analysed here correspond to *I. inopinatus*. Those results also illustrate the potential of using some of the SNPs analysed here to differentiate at a molecular level the two *Ixodes* species.

Two previous studies covering a large spatial extent of *I. ricinus'* range (Noureddine et al., 2011; Porretta et al., 2013) did not find such a clear geographical structure between Eurasian populations. Several

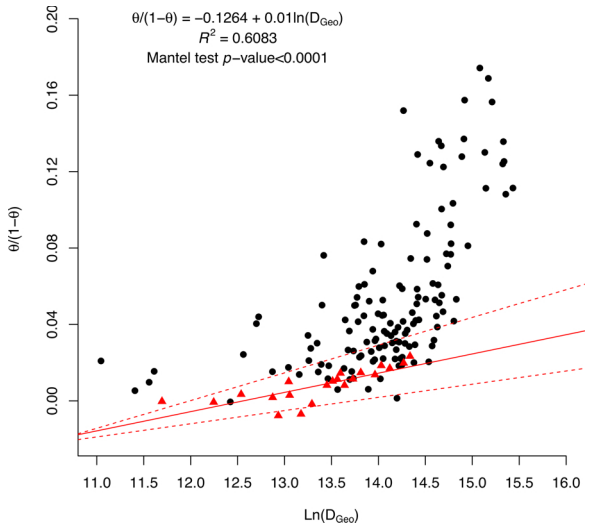


Fig. 4. Isolation by distance between all Eurasian samples. Red triangles represent the pair of samples from the same year: South and North France, Belgium, West and East German, North Estonia, South and Central Sweden. The regression line (plain line), 95 % confidence interval (CI) calculated by bootstrap (dashed lines), Mantel test significance and regression equation corresponds only to red triangles pairs of samples are also shown. Black points correspond to all other pairs of samples not used for further IBD analysis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

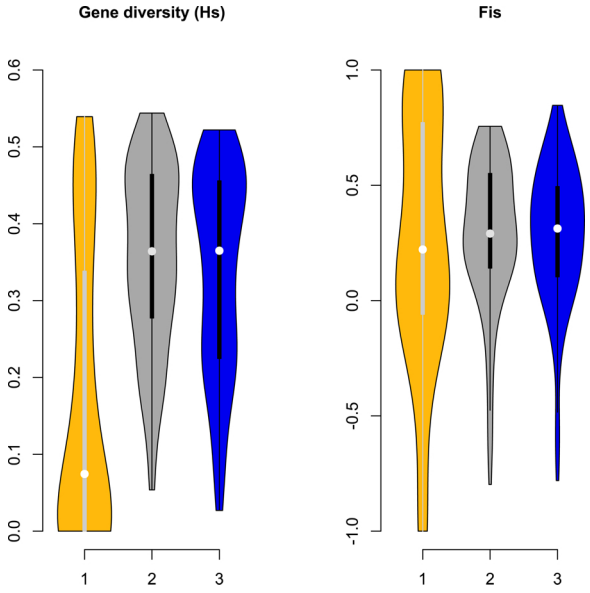


Fig. 5. Values of gene diversity (a) and F_{IS} (b) for each of the three genetic clusters identified by DAPC. Yellow: northern Africa cluster; Blue: southern Eurasia; Grey: northern Europe. Permutation test (Monte-Carlo test, 999 replicates) between all pairs of clusters was significant for gene diversity (p = 0.001) but no significance was identified for F_{IS}. Eurasian clusters show a more pronounced heterozygote excess than the northern African one. A variation of F_{IS} values across loci was observed in the three clusters, even though this variation was much larger in the northern African cluster. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

reasons may explain this difference. First, a somewhat reduced number of individuals per population (sometimes a single individual per population in Noureddine et al., 2011) may explain a lack of spatially structured signal in former studies. Second, those former studies were based on mitochondrial and nuclear sequences. This said, a marked genetic differentiation into two distinctive clades has already been reported (Dinnis et al., 2014; Røed et al., 2016), suggesting a split in *I. ricinus* populations between northern continental Europe and Great Britain. Our results confirm and extend this pattern to most of the Eurasian range of the species by suggesting that Scandinavian populations are genetically closer to the populations from the north-eastern continental parts of Europe. Although there is a certain degree of gene flow between the two clusters, the north vs. south-eastern exchange may be hampered by mountain areas in central Europe. This reinforces the argument that large animals efficiently maintain high gene flow between tick populations across contiguous and permeable landscapes, while intense transportation by birds, during spring and autumn migration across sea or mountains (Hasle et al., 2009; Røed et al., 2016), may not be as sufficient to break down boundaries between established genetic entities.

Surprisingly, we found a close genetic affinity between all Atlantic samples (i.e. Ireland, England, western and southern France, and Spain) and the geographically separated populations from Turkey and Iran. This genetic affinity among distant populations in Eurasia was supported by the two different clustering methods we used (DAPC and STRUCTURE). Besides these results, the refined hierarchical analyses isolated Iran and Turkey in their particular clusters in the first (DAPC) and second (STRUCTURE) round of hierarchical clustering analyses. This suggests that an east-west transport of ticks across southern Eurasia must be sufficient to maintain a genetically identifiable cluster across this extensive area. Interestingly, louping-Ill like viruses are also known from Greece and Turkey (Gao et al., 1993; Marin et al., 1995), which might further support our findings and a link between tick lineages and *Flavivirus*, although the causation is not known.

Since migratory birds carry *I. ricinus* across long distances, different migratory routes could also contribute to the north-south genetic differentiation we observed (Hasle et al., 2009; Røed et al., 2016). However, birds mainly carry larvae and nymphs. Since surviving rates between development states are low, the overall reproductive success of ticks transported by birds is likely smaller than that of adult ticks carried by large mammals. This may explain the maintenance of genetic differentiation e.g. between the UK and Norway despite massive transport of ticks larvae in both directions (Røed et al., 2016).

Regarding the population structure observed within samples, the deviation from Hardy-Weinberg equilibrium we found is in agreement with previous studies on population genetics of *I. ricinus* based on SNPs (Quillery et al., 2014) and microsatellites (Kempf et al., 2009, 2011; Røed et al., 2006), as well as other tick species (Dharmarajan et al., 2011). Possible causes of the observed deviation from the Hardy-Weinberg equilibrium are assortative mating (or assortative pairing), Wahlund effect, or errors in the genotyping. A tendency of mating between phenotypically or genetically similar individuals may effectively increase the inbreeding and thus heterozygote deficiency within populations (Jiang et al., 2013). Kempf et al. (2009) suggested that assortative mating might occur in *I. ricinus*, mostly via host selection (Kempf et al., 2011). Inbreeding in ticks could be a result of host infestation by related individuals, which leads to high breeding success of sibling groups (Araya-Anchetta et al., 2015). The highly aggregated egg masses in *I. ricinus* (1000–3000 eggs) and the limited active dispersal of larvae and nymphs may lead to a high likelihood of mating between related individuals and thus inbreeding. Finally, the parasite-host relationship specificities could also play an important role in establishing or maintaining population structure in *I. ricinus*. If different host populations are present locally and exhibit behaviours favouring mating within (and not between) each host population, this may induce a Wahlund effect and explains the heterozygote deficiency observed. The existence of such a host population behaviour has been characterized in *I. uriae*, a tick associated with sea birds (Mccoy et al., 2001) but also suggested in *I. ricinus* (Kempf et al., 2009, 2011). Even though we did not conceive this study to test for such a hypothesis, our results support at least partially non-random mating in *I. ricinus* populations and the

consequent Wahlund effect. Dharmarajan et al. (2011) facing a similar result for the American species *I. texanus* showed that subdivided breeding groups and high variance in individual reproductive success can correctly explain Hardy-Weinberg equilibrium deviation.

It is widely acknowledged that more or less isolated populations could develop particular adaptations in response to environmental differences between habitats. Nonetheless, very few studies to date have clearly observed phenotypic variations among *I. ricinus* populations from different geographical origins. In Estrada-Peña et al. (1996, 1998), differences in cuticular hydrocarbon composition among European populations of *I. ricinus* were observed according to the geographical origin of those populations. Interestingly, the multivariate phenotypic analysis presented in those studies showed a somewhat similar pattern to our hierarchical genetic clustering analysis, notably concerning what the authors call 'peripheral populations'. Aside from chemical differentiation, behavioural differences between ticks' populations have also been documented, such as mismatches in questing peaks (Schulz et al., 2014) and questing responses to temperature (Gilbert et al., 2014; Tomkins et al., 2014). In controlled conditions, Gilbert et al. (2014) and Tomkins et al. (2014) showed that *I. ricinus* nymphs from cooler climates begin questing at lower temperatures than nymphs from warmer climates. They also start questing sooner when the temperature was kept constant. In any case, local adaptations could impact the spatial redistribution of the species range in response to changes in abiotic conditions. In a global changing context, such consequences could be explored by environmental niche modelling to identify areas of potential future expansion. It remains to be investigated if the different clusters we identified here could pose different threats for human health and the potential risk of tick-borne disease transmission to humans.

Our findings on isolation by distance suggest small population densities and large dispersal distances among the sampled populations. The large dispersal distance is not a surprising result since ticks can parasitize highly mobile species. In a changing climate context, this result indicates that ticks could easily colonize new suitable habitats outside the current limits of the species geographical range in a few generations.

Despite being a generalist ectoparasite, our results highlight geographically distinct and genetically structured populations in *I. ricinus*. More research on host preference and dispersal capacity is needed to better understand those patterns. The differentiation of Eurasian populations into two geographically distinct clusters (northern Europe vs. southern Eurasia) could have important implications for the redistribution of *I. ricinus* in response to anthropogenic climate change. Ticks from a given genetic cluster could be more or less prone to increase in abundance in some regions. Combining tick and pathogen population genetics with knowledge on host distribution could help in the early detection of the spread of tick-borne diseases and thus improve the responsiveness of public authorities to limit major public health concerns.

CRediT authorship contribution statement

Pedro Poli: Investigation, Formal analysis, Writing - original draft, Visualization, Data curation. **Jonathan Lenoir:** Funding acquisition, Writing - review & editing. **Olivier Plantard:** Conceptualization, Funding acquisition, Writing - review & editing, Resources. **Steffen Ehrmann:** Resources, Writing - review & editing. **Knut H. Røed:** Resources, Writing - review & editing. **Hans Petter Leinaas:** Resources, Writing - review & editing. **Marcus Panning:** Conceptualization. **Annie Guiller:** Conceptualization, Funding acquisition, Writing - review & editing, Supervision, Formal analysis, Project administration.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgements

We thank all the invaluable contribution of Dr. Brigitte Degeilh (Laboratory of Parasitology and Mycology, Rennes University Hospital, France), Dr. Francesco Nazzi (Università Degli Studi di Udine), Dr. Mohammad Abdigoudarzi (Razi Institute, Iran), Dr Ali Bouattour (Institut Pasteur de Tunis, Tunisia), Dr Loubna Dib (Institut Vétérinaire, Centre Universitaire El Tarf, Algeria), M'hammed Sarih (Institut Pasteur du Maroc, Morocco), Dr. Irina Golovljova (National Institute for Health Development, Estonia), Nicole Voss (University of Erlangen-Nürnberg, Germany), Dr. Zati Vatansever (Kafkas University, Turkey), Dr. Eoin Healy (University College, Ireland), Dr. Davide Sassera (University of Pavia, Italy), Dr. Albert Agoulon (Oniris, France), Dr. Ionut Pavel (Regional Institute of Oncology, Romania), Dr. György Csóka (Hungary Forest Research Institut, Hungary), Dr. Elena Kocianová (Slovak Academy of Sciences, Slovakia), Dr. Alexandru Movila (Academy of Sciences of Moldova, Moldavia) for sharing some of the samples used in this study. We thank Dr. Michael Scherer-Lorenzen (Freiburg University, German) for the general smallFOREST study design, from which many samples were made available. The smallFOREST project was funded by the ERA-Net BiodivERsA, via the national funders ANR (France), FORMAS (Sweden), ETAG (Estonia), DFG (Germany), BELSPO (Belgium) and the European Union through the European Regional Development Fund (the EcolChange Centre of Excellence). We thank the Centre de Ressources Régionales en Biologie Moléculaire (CRRBM) from the Université de Picardie Jules Verne. This work was also supported by the Région Hauts-de-France, the Centre National de la Recherche Scientifique (CNRS, INEE) and the European Regional Development Fund.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ttbdis.2020.101509.

References

Al-Khafaji, A.M., Clegg, S.R., Pinder, A.C., Luu, L., Hansford, K.M., Seelig, F., Dinnis, R.E., Margos, G., Medlock, J.M., Feil, E.J., Darby, A.C., McGarry, J.W., Gilbert, L., Plantard, O., Sassera, D., Makepeace, B.L., 2019. Multi-locus sequence typing of *Ixodes ricinus* and its symbiont *Candidatus* Midichloria mitochondrii across Europe reveals evidence of local co-cladogenesis in Scotland. Ticks Tick-Borne Dis. 10, 52–62. https://doi.org/10.1016/j.ttbdis.2018.08.016.

Alkishe, A.A., Peterson, A.T., Samy, A.M., 2017. Climate change influences on the potential geographic distribution of the disease vector tick *Ixodes ricinus*. PLoS One 12, e0189092. https://doi.org/10.1371/journal.pone.0189092.

Araya-Anchetta, A., Busch, J.D., Scoles, G.A., Wagner, D.M., 2015. Thirty years of tick population genetics: a comprehensive review. Infect. Genet. Evol. 29, 164–179. https://doi.org/10.1016/j.meegid.2014.11.008.

Archie, E.A., Ezenwa, V.O., 2011. Population genetic structure and history of a generalist parasite infecting multiple sympatric host species. Int. J. Parasitol. 41, 89–98. https://doi.org/10.1016/j.ijpara.2010.07.014.

Babos, S., 1964. Die Zeckenfauna Mitteleuropas. Akadémiai Kiadó, Budapest.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29 (4), 1165–1188. www.jstor.org/stable/2674075.

Blanchong, J.A., Robinson, S.J., Samuel, M.D., Foster, J.T., 2016. Application of genetics and genomics to wildlife epidemiology: genetics and wildlife epidemiology. J. Wildl. Manag. 80, 593–608. https://doi.org/10.1002/jwmg.1064.

Carpi, G., Kitchen, A., Kim, H.L., Ratan, A., Drautz-Moses, D.I., McGraw, J.J., Kazimirova, M., Rizzoli, A., Schuster, S.C., 2016. Mitogenomes reveal diversity of the European Lyme borreliosis vector *Ixodes ricinus* in Italy. Mol. Phylogenet. Evol. 101, 194–202. https://doi.org/10.1016/j.ympev.2016.05.009.

Casati, S., Bernasconi, M.V., Gern, L., Piffaretti, J.-C., 2008. Assessment of intraspecific mtDNA variability of European *Ixodes ricinus* sensu stricto (Acari: Ixodidae). Infect. Genet. Evol. 8, 152–158. https://doi.org/10.1016/j.meegid.2007.11.007.

Chitimia-Dobler, L., Rieß, R., Kahl, O., Wölfel, S., Dobler, G., Nava, S., Estrada-Peña, A., 2018. *Ixodes inopinatus* – Occurring also outside the Mediterranean region. Ticks Tick-Borne Dis. 9, 196–200. https://doi.org/10.1016/j.ttbdis.2017.09.004.

Dantas-Torres, F., 2015. Climate change, biodiversity, ticks and tick-borne diseases: the butterfly effect. Int. J. Parasitol. Parasites Wildl. 4, 452–461. https://doi.org/10.1016/j.ijppaw.2015.07.001.
De Meeûs, T., 2018. Revisiting FIS, FST, Wahlund effects, and null alleles. J. Hered. 109, 446–456. https://doi.org/10.1093/jhered/esx106.
Dharmarajan, G., Beasley, J.C., Rhodes, O.E., 2011. Heterozygote deficiencies in parasite populations: an evaluation of interrelated hypotheses in the raccoon tick, *Ixodes texanus*. Heredity 106, 253–260. https://doi.org/10.1038/hdy.2010.84.
Dinnis, R.E., Seelig, F., Bormane, A., Donaghy, M., Vollmer, S.A., Feil, E.J., Kurtenbach, K., Margos, G., 2014. Multilocus sequence typing using mitochondrial genes (mtMLST) reveals geographic population structure of *Ixodes ricinus* ticks. Ticks Tick-Borne Dis. 5, 152–160. https://doi.org/10.1016/j.ttbdis.2013.10.001.
Do, C., Waples, R.S., Peel, D., Macbeth, G.M., Tillett, B.J., Ovenden, J.R., 2014. NEESTIMATOR v2: re-implementation of software for the estimation of contemporary effective population size (N_e) from genetic data. Mol. Ecol. Resour. 14, 209–214. https://doi.org/10.1111/1755-0998.12157.
Dray, S., Dufour, A.-B., 2007. The ade4 package: implementing the duality diagram for ecologists. J. Stat. Softw. 22. https://doi.org/10.18637/jss.v022.i04.
Earl, D.A., vonHoldt, B.M., 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. 4, 359–361. https://doi.org/10.1007/s12686-011-9548-7.
Ehrmann, S., Liira, J., Gärtner, S., Hansen, K., Brunet, J., Cousins, S.A.O., Deconchat, M., Decocq, G., De Frenne, P., De Smedt, P., Diekmann, M., Gallet-Moron, E., Kolb, A., Lenoir, J., Lindgren, J., Naaf, T., Paal, T., Valdés, A., Verheyen, K., Wulf, M., Scherer-Lorenzen, M., 2017. Environmental drivers of *Ixodes ricinus* abundance in forest fragments of rural European landscapes. BMC Ecol. 17. https://doi.org/10.1186/s12898-017-0141-0.
Estrada-Peña, A., Gray, J.S., Kahl, O., 1996. Variability in cuticular hydrocarbons and phenotypic discrimination of *Ixodes ricinus* populations (Acarina: Ixodidae) from Europe. Exp. Appl. Acarol. 20, 457–466. https://doi.org/10.1007/BF00053309.
Estrada-Peña, A., Daniel, M., Frandsen, F., Gern, L., Gettinby, G., Gray, J.S., Jaenson, T.G.T., Jongejan, F., Kahl, O., Korenberg, E., Mehl, R., Nuttall, P.A., 1998. *Ixodes ricinus*Strains in Europe. Zentralblatt für Bakteriologie 287, 185–189. https://doi.org/10.1016/S0934-8840(98)80119-9.
Estrada-Peña, A., Nava, S., Petney, T., 2014. Description of all the stages of *Ixodes inopinatus*n. sp. (Acari: Ixodidae). Ticks Tick-Borne Dis. 5, 734–743. https://doi.org/10.1016/j.ttbdis.2014.05.003.
European Centre for Disease Prevention and Control and European Food Safety Authority. Tick maps 2019. https://ecdc.europa.eu/en/disease-vectors/surveillance-and-disease-data/tick-maps (Accessed 12 January 2020).
Evanno, G., Regnaut, S., Goudet, J., 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. Mol. Ecol. 14, 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x.
Gao, G.F., Hussain, M.H., Reid, H.W., Gould, E.A., 1993. Classification of a new member of the TBE *Flavivirus* subgroup by its immunological, pathogenetic and molecular characteristics: identification of subgroup-specific pentapeptides. Virus Res. 30, 129–144. https://doi.org/10.1016/0168-1702(93)90002-5.
Gilbert, L., Aungier, J., Tomkins, J.L., 2014. Climate of origin affects tick (*Ixodes ricinus*) host-seeking behavior in response to temperature: implications for resilience to climate change? Ecol. Evol. 4, 1186–1198. https://doi.org/10.1002/ece3.1014.
Gooding, R.H., 1996. Genetic variation in arthropod vectors of disease-causing organisms: obstacles and opportunities. Clin. Microbiol. Rrev. 9, 301–320.
Goudet, J., 2003. FSTAT (Version 2.9.4), a Program to Estimate and Test Population Genetics Parameters. Available from:. https://www2.unil.ch/popgen/softwares/fstat.htm.
Goudet, J., Jombart, T., 2018. hierfstat: Estimation and Tests of Hierarchical F-Statistics. http://www.r-project.org. http://github.com/jgx65/hierfstat.
Handeland, K., Qviller, L., Vikøren, T., Viljugrein, H., Lillehaug, A., Davidson, R.K., 2013. *Ixodes ricinus*infestation in free-ranging cervids in Norway—a study based upon ear examinations of hunted animals. Vet. Parasitol. 195, 142–149. https://doi.org/10.1016/j.vetpar.2013.02.012.
Hasle, G., Bjune, G., Edvardsen, E., Jakobsen, C., Linnehol, B., Røer, J.E., Mehl, R., Røed, K.H., Pedersen, J., Leinaas, H.P., 2009. Transport of ticks by migratory passerine birds to Norway. J. Parasitol. 95, 1342–1351. https://doi.org/10.1645/GE-2146.1.
Helyar, S.J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M.I., Ogden, R., Limborg, M.T., Cariani, A., Maes, G.E., Diopere, E., Carvalho, G.R., Nielsen, E.E., 2011. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Mol. Ecol. Resour. 11, 123–136. https://doi.org/10.1111/j.1755-0998.2010.02943.x.
Hijmans, R.J., 2017. Geosphere: Spherical Trigonometry. R Package Version 1.5-7. https://CRAN.R-project.org/package=geosphere.
Hillyard, P.D., 1996. Ticks of North-West Europe. In: Barnes, R.S.K., Crothers, J.H. (Eds.), Synopses of the British Fauna (New Series).
Hvidsten, D., Frafjord, K., Gray, J.S., Henningsson, A.J., Jenkins, A., Kristiansen, B.E., Lager, M., Rognerud, B., Slåtsve, A.M., Stordal, F., Stuen, S., Wilhelmsson, P., 2020. The distribution limit of the common tick, *Ixodes ricinus*, and some associated pathogens in north-western Europe. Ticks Tick-Borne Dis. 11, 101388. https://doi.org/10.1016/j.ttbdis.2020.101388.
Jakobsson, M., Rosenberg, N.A., 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23, 1801–1806. https://doi.org/10.1093/bioinformatics/btm233.
Jiang, Y., Bolnick, D.I., Kirkpatrick, M., 2013. Assortative mating in animals. Am. Nat. 181, E125–E138. https://doi.org/10.1086/670160.
Jombart, T., 2008. Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24, 1403–1405. https://doi.org/10.1093/bioinformatics/btn129.
Jombart, T., Devillard, S., Balloux, F., 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11, 94. https://doi.org/10.1186/1471-2156-11-94.
Jore, S., Viljugrein, H., Hofshagen, M., Brun-Hansen, H., Kristoffersen, A.B., Nygård, K., Brun, E., Ottesen, P., Sævik, B.K., Ytrehus, B., 2011. Multi-source analysis reveals latitudinal and altitudinal shifts in range of *Ixodes ricinus* at its northern distribution limit. Parasit. Vectors 4. https://doi.org/10.1186/1756-3305-4-84.
Kempf, F., de Meeûs, T., Arnathau, C., Degeilh, B., McCoy, K.D., 2009. Assortative pairing in *ixodes ricinus* (Acari: Ixodidae), the european vector of lyme borreliosis. J. Med. Entomol. 46, 471–474. https://doi.org/10.1603/033.046.0309.
Kempf, F., De Meeûs, T., Vaumourin, E., Noel, V., Taragel'ová, V., Plantard, O., Heylen, D.J.A., Eraud, C., Chevillon, C., McCoy, K.D., 2011. Host races in *Ixodes ricinus*, the European vector of Lyme borreliosis. Infect. Genet. Evol. 11, 2043–2048. https://doi.org/10.1016/j.meegid.2011.09.016.
Kozakiewicz, C.P., Burridge, C.P., Funk, W.C., VandeWoude, S., Craft, M.E., Crooks, K.R., Ernest, H.B., Fountain-Jones, N.M., Carver, S., 2018. Pathogens in space: advancing understanding of pathogen dynamics and disease ecology through landscape genetics. Evol. Appl. 11, 1763–1778. https://doi.org/10.1111/eva.12678.
Kriz, B., Daniel, M., Benes, C., Maly, M., 2014. The role of game (Wild Boar and Roe Deer) in the spread of tick-borne encephalitis in the Czech Republic. Vector-Borne Zoon. Dis. 14, 801–807. https://doi.org/10.1089/vbz.2013.1569.
Lang, K.R., Blanchong, J.A., 2012. Population genetic structure of white-tailed deer: understanding risk of chronic wasting disease spread. J. Wildl. Manag. 76, 832–840. https://doi.org/10.1002/jwmg.292.
Lindgren, E., Gustafson, R., 2001. Tick-borne encephalitis in Sweden and climate change. Lancet 358, 16–18. https://doi.org/10.1016/S0140-6736(00)05250-8.
Manangwa, O., De Meeûs, T., Grébaut, P., Ségard, A., Byamungu, M., Ravel, S., 2019. Detecting Wahlund effects together with amplification problems: cryptic species, null alleles and short allele dominance in *Glossina pallidipes* populations from Tanzania. Mol. Ecol. Resour. 19, 757–772. https://doi.org/10.1111/1755-0998.12989.
Marin, M.S., McKenzie, J., Gao, G.F., Reid, H.W., Antoniadis, A., Gould, E.A., 1995. The virus causing encephalomyelitis in sheep in Spain: a new member of the tick-borne encephalitis group. Res. Vet. Sci. 58, 11–13. https://doi.org/10.1016/0034-5288(95)90081-0.
Mccoy, K.D., Boulinier, T., Tirard, C., Michalakis, Y., 2001. Host specificity of a generalist parasite: genetic evidence of sympatric host races in the seabird tick *Ixodes uriae*. J. Evol. Biol. 14, 395–405. https://doi.org/10.1046/j.1420-9101.2001.00290.x.
Medlock, J.M., Hansford, K.M., Bormane, A., Derdakova, M., Estrada-Peña, A., George, J.-C., Golovljova, I., Jaenson, T.G., Jensen, J.-K., Jensen, P.M., 2013. Driving forces for changes in geographical distribution of *Ixodes ricinus* ticks in Europe. Parasit. Vectors 6, 1. https://doi.org/10.1186/1756-3305-6-1.
Norte, A.C., de Carvalho, I.L., Ramos, J.A., Gonçalves, M., Gern, L., Núncio, M.S., 2012. Diversity and seasonal patterns of ticks parasitizing wild birds in western Portugal. Exp. Appl. Acarol. 58, 327–339. https://doi.org/10.1007/s10493-012-9583-4.
Noureddine, R., Chauvin, A., Plantard, O., 2011. Lack of genetic structure among Eurasian populations of the tick *Ixodes ricinus* contrasts with marked divergence from north-African populations. Int. J. Parasitol. 41, 183–192. https://doi.org/10.1016/j.ijpara.2010.08.010.
Oksanen, F., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, F., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, H., Szoecs, E., Wagner, H., 2019. Vegan: Community Ecology Package. R Package Version 2.5-5. https://CRAN.R-project.org/package=vegan.
Pecl, G.T., Araújo, M.B., Bell, J.D., Blanchard, J., Bonebrake, T.C., Chen, I.-C., Clark, T.D., Colwell, R.K., Danielsen, F., Evengård, B., Falconi, L., Ferrier, S., Frusher, S., Garcia, R.A., Griffis, R.B., Hobday, A.J., Janion-Scheepers, C., Jarzyna, M.A., Jennings, S., Lenoir, J., Linnetved, H.I., Martin, V.Y., McCormack, P.C., McDonald, J., Mitchell, N.J., Mustonen, T., Pandolfi, J.M., Pettorelli, N., Popova, E., Robinson, S.A., Scheffers, B.R., Shaw, J.D., Sorte, C.J.B., Strugnell, J.M., Sunday, J.M., Tuanmu, M.-N., Vergés, A., Villanueva, C., Wernberg, T., Wapstra, E., Williams, S.E., 2017. Biodiversity redistribution under climate change: impacts on ecosystems and human well-being. Science 355https://doi.org/10.1126/science.aai9214. eaai9214.
Pérez-Eid, C., 2007. Les tiques - Identification, biologie, importance médicale et vétérinaire. Lavoisier, Provigny. .
Porretta, D., Mastrantonio, V., Mona, S., Epis, S., Montagna, M., Sassera, D., Bandi, C., Urbanelli, S., 2013. The integration of multiple independent data reveals an unusual response to Pleistocene climatic changes in the hard tick Ixodes ricinus. Mol Ecol . 22, 1666–1682. https://doi.org/10.1111/mec.12203.
Portillo, A., Santibáñez, P., Palomar, A.M., Santibáñez, S., Oteo, J.A., 2018. '*Candidatus* Neoehrlichia mikurensis' in Europe. New Microbes New Infect. 22, 30–36. https://doi.org/10.1016/j.nmni.2017.12.011.
Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959. https://www.genetics.org/content/155/2/945.
Pritchard, J.K., Wen, X., Falush, D., 2007. Documentation for Structure Software: Version 2.2. (Accessed 20 November 2019). https://web.stanford.edu/group/pritchardlab/software/structure22/readme.pdf.
Quillery, E., Quenez, O., Peterlongo, P., Plantard, O., 2014. Development of genomic resources for the tick *Ixodes ricinus*: isolation and characterization of single nucleotide polymorphisms. Mol. Ecol. Resour. 14, 393–400. https://doi.org/10.1111/1755-0998.12179.
R Core Team, 2019. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Rizzoli, A., Hauffe, H.C., Tagliapietra, V., Neteler, M., Rosà, R., 2009. Forest structure and roe deer abundance predict tick-borne encephalitis risk in Italy. PLoS One 4, e4336. https://doi.org/10.1371/journal.pone.0004336.
Røed, K.H., Hasle, G., Midthjell, V., Skretting, G., Leinaas, H.P., 2006. Identification and

characterization of 17 microsatellite primers for the tick, *Ixodes ricinus*, using enriched genomic libraries: PRIMER NOTE. Mol. Ecol. Notes 6, 1165–1167. https://doi.org/10.1111/j.1471-8286.2006.01475.x.

Røed, K.H., Kvie, K.S., Hasle, G., Gilbert, L., Leinaas, H.P., 2016. Phylogenetic lineages and postglacial dispersal dynamics characterize the genetic structure of the tick, *Ixodes ricinus*, in Northwest Europe. PLoS One 11, e0167450. https://doi.org/10.1371/journal.pone.0167450.

Rousset, F., 1997. Genetic differentiation and estimation of gene flow from FStatistics under isolation by distance. Genetics 145, 1219–1228. https://www.genetics.org/content/145/4/1219.

Schulz, M., Mahling, M., Pfister, K., 2014. Abundance and seasonal activity of questing *Ixodes ricinus* ticks in their natural habitats in southern Germany in 2011. J. Vector Ecol. 39, 56–65. https://doi.org/10.1111/j.1948-7134.2014.12070.x.

Séré, M., Thévenon, S., Belem, A.M.G., De Meeûs, T., 2017. Comparison of different genetic distances to test isolation by distance between populations. Heredity 119, 55–63. https://doi.org/10.1038/hdy.2017.26.

Smouse, P.E., 2010. How many SNPs are enough? Mol. Ecol. 19, 1265–1266. https://doi.org/10.1111/j.1365-294X.2010.04555.x.

Tabachnick, W.J., Black, W.C., 1995. Making a case for molecular population genetic studies of arthropod vectors. Parasitol. Today 11, 27–30. https://doi.org/10.1016/0169-4758(95)80105-7.

Tomkins, J.L., Aungier, J., Hazel, W., Gilbert, L., 2014. Towards an evolutionary understanding of questing behaviour in the tick *Ixodes ricinus*. PLoS One 9, e110028. https://doi.org/10.1371/journal.pone.0110028.

Vähä, J.-P., Erkinaro, J., Niemelä, E., Primmer, C.R., 2007. Life-history and habitat features influence the within-river genetic structure of Atlantic salmon. Mol. Ecol. 16, 2638–2654. https://doi.org/10.1111/j.1365-294X.2007.03329.x.

Van Zee, J., Piesman, J.F., Hojgaard, A., Black IV, W.C., 2015. Nuclear markers reveal predominantly north to south gene flow in *Ixodes scapularis*, the tick vector of the lyme disease spirochete. PLoS One 10, e0139630. https://doi.org/10.1371/journal.pone.0139630.

Weir, B.S., Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. Evolution 38, 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x.

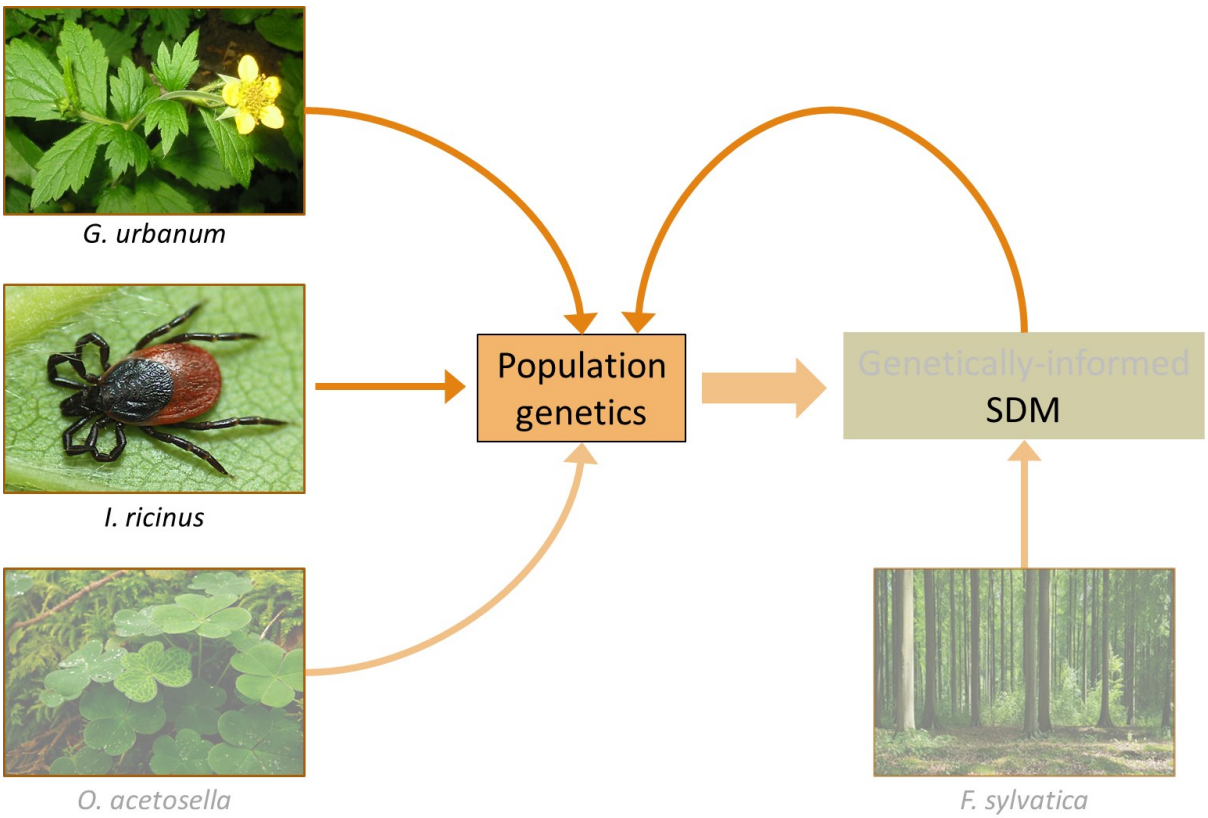
Welinder-Olsson, C., Kjellin, E., Vaht, K., Jacobsson, S., Wenneras, C., 2010. First case of human "*Candidatus* neoehrlichia mikurensis" infection in a febrile patient with chronic lymphocytic leukemia. J. Clin. Microbiol. 48, 1956–1959. https://doi.org/10.1128/JCM.02423-09.

Wessels, C., Matthee, S., Espinaze, M.P.A., Matthee, C.A., 2019. Comparative mtDNA phylogeographic patterns reveal marked differences in population genetic structure between generalist and specialist ectoparasites of the African penguin (*Spheniscus demersus*). Parasitol. Res. 118, 667–672. https://doi.org/10.1007/s00436-018-6150-x.

Wonham, M.J., Lewis, M.A., Renclawowicz, J., van den Driessche, P., 2006. Transmission assumptions generate conflicting predictions in host-vector disease models: a case study in West Nile virus. Ecol. Lett. 9, 706–725. https://doi.org/10.1111/j.1461-0248.2006.00912.x.

Younsi, H., Fares, W., Cherni, S., Dachraoui, K., Barhoumi, W., Najjar, C., Zhioua, E., 2020. *Ixodes inopinatus* and *ixodes ricinus* (Acari: Ixodidae) Are sympatric ticks in North Africa. J. Med. Entomol. 57, 952–956. https://doi.org/10.1093/jme/tjz216.

Chapitre 3 Towards better understanding population genetic structure: contribution of species distribution models



Introduction

Population genetic differentiation may originate from many different evolutionary processes, such as migration and gene flow, genetic drift, funding effects, as well as natural selection (Slatkin 1993, Pearse and Crandall 2004, Holsinger and Weir 2009). Understanding the genetic structure of populations and which of those individual processes (or combination of processes) influence the observed patterns has been one of the main goals of population genetics since its inception (Wright 1931, Pritchard *et al.* 2000, Bradburd *et al.* 2018). Observed genetic structures may be the result of the presence of a more or less impervious physical barrier to gene flow, and therefore could be accessed by correlating landscape features to metrics inferred from population genetics (Manel and Holderegger 2013). Genetic structure is also the result of the evolutionary history of populations. For instance, past climatic oscillations are considered to be an important cause of population differentiation (Hewitt 2000, 2004, Schmitt 2007). In the European context, the post-glacial expansion of many taxa has contributed to the genetic variation observed between populations of the same taxa (Magri *et al.* 2006, Niedziałkowska *et al.* 2011, Pedreschi *et al.* 2019).

Species Distribution Models (SDMs) are widely used in ecology to serve a multitude of applications. Among others, SDMs were applied to infer changes in habitat suitability from past to present conditions, allowing to better understand how species shifted their ranges following past climate changes (Svenning *et al.* 2008, 2011, Roces-Díaz *et al.* 2018). When coupled with genetic data, SDMs have also allowed to test several hypotheses that relate to the evolutionary history at the species and population levels, such as: (**i**) the degree of gene flow between populations located in glacial refugia during the Last Glacial Maximum (LGM) (Porretta *et al.* 2013, Wren and Burke 2019); (**ii**) population range expansion and lineage divergence (Diniz-Filho *et al.* 2016, Palma *et al.* 2017); (**iii**) the identification of past climatic refugia (Assis *et al.* 2016); and (**iv**) niche conservatism (Gutiérrez-Rodríguez *et al.* 2017, Meynard *et al.* 2017).

This third chapter aims at detecting how historical and contemporary bioclimatic conditions contribute to the genetic divergence and the genetic structure of two of the models species explored in chapter 2, i.e. the tick *Ixodes ricinus* and the herbaceous plant *Geum urbanum*. Considering the former results of the population genetics' analyses indicating that both species are structured in two genetic clusters mainly separated by latitude, i.e. a Northern cluster and a Southern cluster (see chapter 2), I hypothesised that the observed present genetic patterns of populations are the consequences of the species' range expansion dynamics since the LGM. In this case, contemporary climatic conditions (1970-2000) would have a minor influence on the overall spatial genetic structure of the two studied species, while changes in climatic conditions since the LGM should matter to explain the overall spatial genetic structure of the two studied species. To test those hypotheses for each of the two studied species I used SDMs to predict past and present habitat suitability as well as changes in habitat suitability, and then I applied a combination of correlative and regression methods to: (**i**) identify candidate loci under selection by bioclimatic variables and changes in habitat suitability; (**ii**) investigate the influence of habitat suitability on population structure in the near present and changes in habitat suitability compared to the LGM; and (**iii**) directly test the influence of the changes in habitat suitability on the spatial structuring of populations as inferred previously in chapter 2.

Material and Methods

Species Distribution Models

To test whether allele frequencies for each locus is related to contemporary or past climatic conditions and thus potentially under bioclimatic selection, we used SDMs to predict habitat suitability conditions during both the 1970-2000 period, as a surrogate for contemporary climate, and the LGM. Once habitat suitability values were projected across the study region of Europe at both periods and for both studied species, we extracted those values for each of the sampled populations to study the relationship between allele frequencies of each locus and habitat suitability conditions.

Present and past projections for Ixodes ricinus

The potential distribution of *I. ricinus* was calibrated under bioclimatic conditions during the period 1970-2000 as a function of five non co-linear bioclimatic variables, namely: (**i**) annual mean temperature (BIO 1); (**ii**) mean diurnal temperature range (BIO 2); (**iii**) isothermality (BIO 3: BIO 2/temperature annual range); (**iv**) temperature seasonality (BIO 4); and (**v**) mean annual precipitation (BIO 12). All bioclimatic variables were downloaded from the WorldClim 2 database (Fick and Hijmans 2017, - https://www.worldclim.org/data/index.html) at 5 arc-minute resolution (about 8.5 km²) and are representative of long-term bioclimatic conditions.

Occurrence data for *I. ricinus* were extracted from Poli et al. (2020), GBIF (GBIF.org, 2020), and VectorMap (http://vectormap.si.edu/). A total of 2,171 occurrences were kept for further analysis (**Figure 3.1**a). Four algorithms implemented in the biomod2 R package (Thuiler et al., 2020) were applied to model the potential distribution of *I. ricinus* during the period 1970-2000: (**i**) generalized linear models (GLMs); (**ii**) generalized boosted regression models (GBMs); (**iii**) generalized additive models (GAMs); and (**iv**) random forests (RFs). Each time, we used the default parameters. Ten separate datasets composed of all 2,171 occurrences and the same number of pseudo-absences or background data were built, each one with a different random selection of pseudo-absences drawn within a convex-hull around all occurrences. I decided to not search for pseudo-absences outside this

convex-hull as a way to limit the study area based on the actual spatial distribution of the species. The choice of the number of pseudo-absences is controversial (Barbet-Massin et al., 2012; Liu et al., 2019). According to Barbet-Massin et al. (2012) and Liu et al. (2019), model accuracy tends to be improved with a higher number of pseudo-absences for regression techniques (GLM, GAM), although this gain was modest in Barbet-Massin et al. (2012). The inverse pattern was observed for the Random Forest technique in both Barbet-Massin et al. (2012) and Liu et al. (2019), with and non-negligible lost in model accuracy with higher numbers of pseudo-absences. Considering those results and the fact that it was not the goal of this work to evaluate consequences of different techniques of pseudo-absence selection on the model performance, I decided for a same number of pseudo-absences as there were occurrences. Even if this choice could let to a loss in model accuracy for the regression techniques, it seems this loss would be less important than the gain in the Random Forest technique. For each of the original ten datasets, ten repetitions of the algorithms were run, each time setting aside 70% of the data for model calibration and the rest for an independent model validation. This results in a total of 40 models per dataset (4 algorithms × 10 repetitions). Those 40 models were assembled by weighting coefficient estimates based on true skill statistic (TSS) values, resulting in ten assembled models. Finally, a contemporary or near present distribution raster was built with the mean of the probabilities values from each of the 10 assembled projections based on bioclimatic conditions during 1970-2000. To evaluate the performance of the 10 assembled models, the area under the receiver operator curve (AUC), TSS and the continuous Boyce index (CBI – Hirzel *et al.* 2006) were calculated. Both AUC and TSS values were calculated using the biomod2 package, and CBI values were calculated with the ecospat R package (Broenniman et al., 2020).

Based on each of the ten assembled models for the period 1970-2000, the potential distribution of *I. ricinus* was hindcasted during the LGM, about 20,000 ybp. I used the same bioclimatic variables at the same spatial resolution as those used for model calibration were downloaded from the WorldClim 1.4 database https://www.worldclim.org/data/v1.4/paleo1.4.html - Hijmans *et al.*, 2005) for the three global circulation models (GCMs) that are available in the WorldClim database:

CCSM4; MIROC-ESM; and MPI-ESM-P. Three LGM potential distributions were projected based on each of those GCMs and a raster for the LGM was built from the mean values of the 30 LGM projections (10 repetitions x 3 GCMs).

Present and past projections for Geum urbanum

The general approach for modelling *G. urbanum* distribution during both the 1970-2000 period and the LGM period was the same as the one described for *I. ricinus*. The two differences concern the choice of bioclimatic variables and the selection of occurrences. Five bioclimatic variables were kept: (**i**) annual mean temperature (BIO 1); (**ii**) temperature seasonality (BIO 4); (**iii**) precipitation seasonality (BIO 15); (**iv**) precipitation of the wettest quarter (BIO16); and (**v**) precipitation of the driest quarter (BIO 12). It is important to notice that the potential distribution of *G. urbanum* is a rough approximation, since no soil projection (such as pH or texture) are available for the LGM period. Hence, the modelled distribution of *G. urbanum* reflects only the abiotic niche of the species based on the aforementioned bioclimatic variables without the aim to produce a precise and realistic map of *G. urbanum* distribution across Europe.

Occurrence data for *G. urbanum* consisted of the populations studied in chapter 1 (**Figure 2.5**) and occurrences extracted from GBIF (GBIF, 2020). Occurrences from GBIF were highly concentrated in West Europe, notably in Great Britain, France, and through West Germany. This initial set of occurrences was subject to a cleaning procedure to reduce biases due to oversampling effort in some regions over the studied area of Europe. More specifically, only one occurrence per spatial grid cell of about 45 km² was retained. After this spatial thinning procedure, 1,274 occurrences were kept for further analysis (**Figure 3.1**b). All the remaining steps followed the same approach as described for *I. ricinus*.

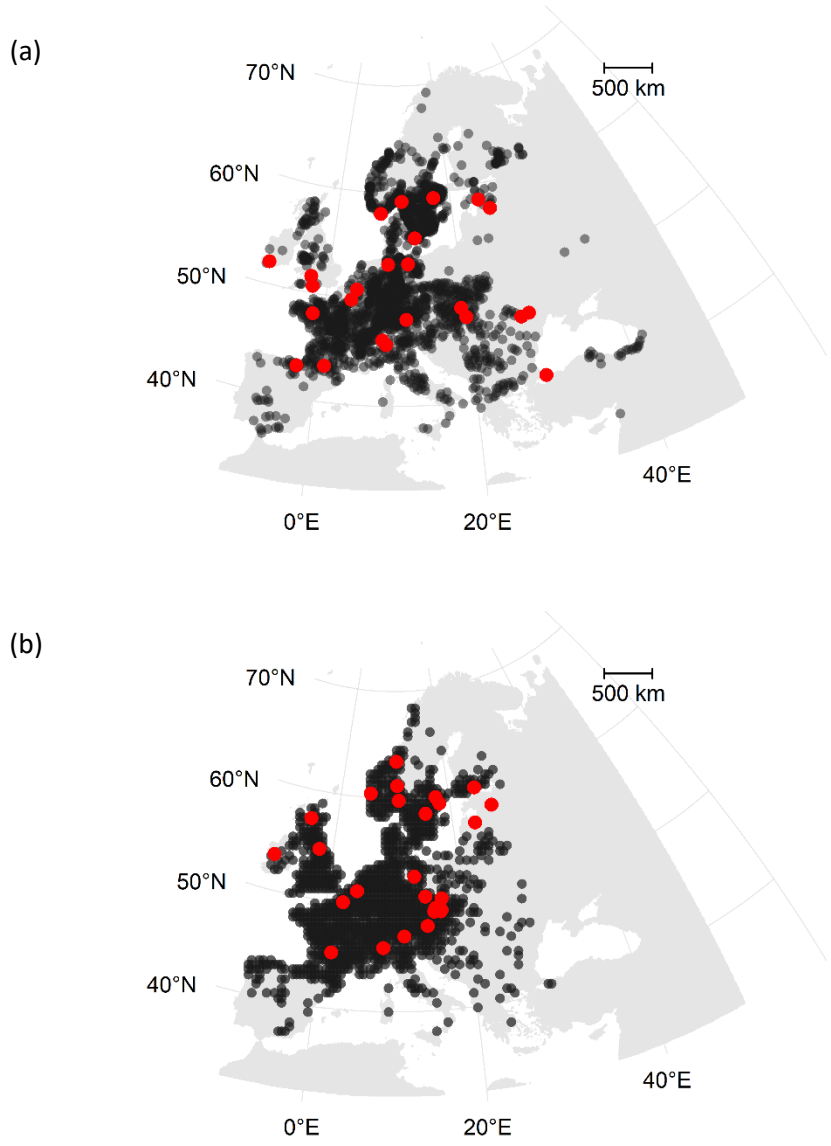


Figure 3.1. Occurrences used for model calibration and validation during the 1970-2000 period for both *Ixodes ricinus* (a) and *Geum urbanum* (b). A total of 2,171 occurrences were used for *I. ricinus*, and 1,274 for G. urbanum. The points in red represent the location of the populations of *I. ricinus* and *G. urbanum* (25 and 27 populations, respectively) analysed in chapter 2.

For each of the two studied species, the predicted habitat suitability values during both 1970-2000 (P_S) and the LGM (LGM_S) were extracted for each of the sampled populations for which genetic analyses have been carried out (see **Figure 3.1**). From these two habitat suitability values per sampled population, we computed the difference as follows: $\Delta_S = P_S - LGM_S$. Positive values of Δ_S indicate

that habitat suitability conditions were more favourable during 1970-2000 than it was during the LGM, while negative values indicate the opposite.

Influence of climate change on the spatial genetic structure of *Ixodes ricinus* and *Geum urbanum*

To better understand the role of bioclimatic factors on the population genetic structure presented in chapter 2 (sections '*Population Genetic Structure of Geum urbanum*' and '*Published article: Strong genetic structure among populations of the tick Ixodes ricinus across its range*'), I conducted a series of regressive and correlative analysis. Those analysis were based either on the allele frequencies across the 25 and 27 sampled populations for *I. ricinus* and *G. urbanum*, respectively, or on the probabilities of assignement of each of those populations to one of the two genetic cluster identified for each species (**Figure 3.2**).

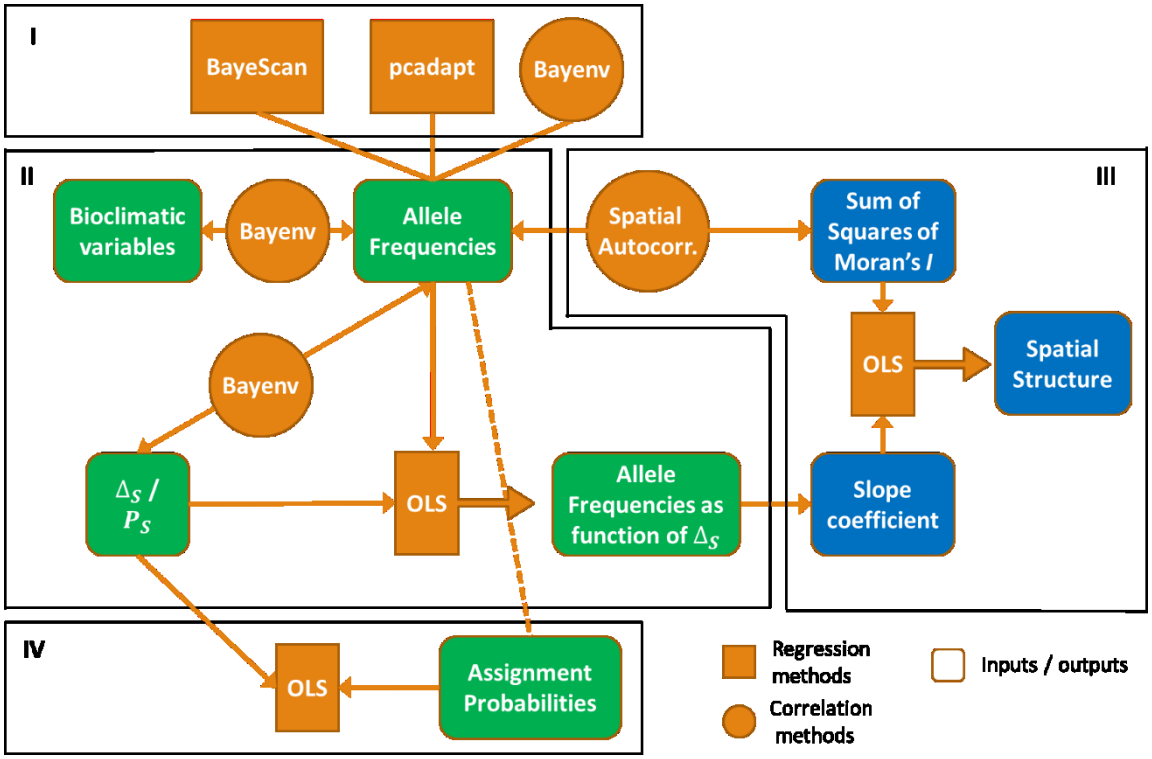


Figure 3.2. Schematic relationship of the different analyses carried out in this third chapter. Four main groups of analyses were conducted for both *Ixodes ricinus* and *Geum urbanum*: I. Candidate loci under selection were identified by correlational and regression methods that only take into account differences in allele frequencies across populations without testing potential links with bioclimatic variables or habitat suitability conditions; II. the relationship between allele frequencies and bioclimatic variables or habitat suitability values were investigated by correlation approaches (bioclimatic variables and habitat suitability) and ordinary least square regression (OLS) with habitat suitability only as the independent variable; III. the influence of habitat suitability values on the spatial structure of allele frequencies were investigated by autocorrelation analysis and regression analysis; and IV. the probabilities of assignment of populations to one of the two genetic clusters of each species (as inferred in chapter 2) will be regressed on the habitat suitability values during 1970-2000 as well as on the changes in habitat suitability values since the LGM. Green colours correspond to analysis with bioclimatic variables and habitat suitability values, and blue colours to spatial autocorrelation analysis.

Identifying candidate loci under selection

The first step of the analyses (box I in **Figure 3.2**) consisted in the identification of candidate loci under selection independently of environmental conditions. The methods described in this section take only genetic variables as input (multilocus genotypes of populations allele frequencies), and thus do not make direct inference of the influence of environmental variables and genetic differentiation between

populations. This first set of methods in the analytical workflow presented in the box I of **Figure 3.2** aims at measuring population differentiation (Günther & Coop, 2018) as consistently different allele frequencies of certain loci across populations. It thus serves here as an exploratory step to detect candidate loci under selection before testing potentially significant relationships between environmental conditions and allele frequencies, i.e. candidate loci under abiotic selection. Although there are several methods available in the scientific literature for identifying candidate loci under selection, some of them were developped for particular types of genetic markers, notably for bi-allelic SNPs as those of *I. ricinus*. Hence, I will describe next the chosen methods for each of the two model species, according to the genetic marker used in the population genetic structure analysis presented previously in chapter 2.

For *I. ricinus*, two methods developed specifically for bi-allelic marker like SNPs were applied. The first is the multivariate method implemented in the R package 'pcadapt' (Privé *et al.* 2020). This method applies a five steps analysis comprising: (**i**) the normalisation of the allele frequencies; (**ii**) the computation of a Principal Component Analysis (PCA) on the normalised allele frequencies; (**iii**) a multiple linear regression linking the allele frequencies of each locus to the coordinates of the PCA; (**iv**) the calculation of the Mahalanobis distance *D* on the z-scores of the previous regressions; and (**v**) the calculation a *p*-value for each locus by comparing the distances *D* for each locus to a chi-square distribution. The normalisation procedure follows (Patterson *et al.* 2006), where the allele frequency G_{ij} of the sample i at locus j is transformed by the function:

$$\widehat{G_{ij}} = \frac{G_{ij}}{\sqrt{\left(2 \times p_j(1-p_j)\right)}}$$

where p_j denotes the frequency of the allele less frequent. After running those analyses, it is recommended to apply a *p*-value correction to control for the false discovery rate (FDR) (Luu et al., 2016). I applied the procedure described by (Benjamini and Yekutieli 2001) in R. The second method applied to identify candidate loci under selection was the Bayenv method (Coop *et al.* 2010, Günther and Coop 2013). Bayenv is a Bayesian approach developed for identifying loci under selection and to

test the correlation between allele frequencies and environmental variables. The correlative approach of Bayenv is described in the next section. This method has the advantage of accounting for differences in sample size across populations and the intrinsic correlation in allele frequencies between populations (due to genetic drift and migration), and is reported to be more powerful and less error-prone than most of the available methods for identifying loci under selection (Lotterhos and Whitlock 2014). The first step in this method is to build a covariance matrix of allele frequencies across populations. This covariance matrix is then used as a null-model of the expected allele frequencies across populations. The posterior probabilities inferences under these models are performed using a Markov Chain Monte Carlo (MCMC) and are expressed as the Bayes factor (BF). Other than the correlation between allele frequencies and environmental variables, this method also calculates a population differentiation statistic for each locus independent of environmental variables, termed X^TX. Extreme values of X^TX indicate selection pressure over one loci.

Since the population genetics analyses for *G. urbanum* were based on multiallelic microsatellite loci, the two methods discussed above, i.e. the pcadapt and Bayenv method, could not be applied. Instead, I investigated loci under selection with BayeScan v.2.1 (Foll and Gaggiotti 2008). The program is also based on the differences in allele frequencies across populations to identify candidate loci under selection. BayeScan estimates the posterior probability by a MCMC approach of a neutral model where a particular locus is not under selection and an alternative model where it is. BayeScan was run with default parameters (100,000 MCMC iterations, with 50,000 burn-in), at the exception of the 'Prior odds of neutral model' parameter, which is the prior probability of a locus being under selection in the data set. The default value of 10 is the odds that in a ten loci dataset, one is under selection. BayeScan is considered to estimate a high number of false positives (Foll & Gaggiotti, 2008; Lotterhos & Whitlock, 2014), particularly for species that have experienced recent range expansion. Lotterhos & Whitlock (2014), suggest that values from 100 to 10,000 to this parameter tends to reduce the number of false positives. Three runs of BayeScan were conducted setting this parameter to 100, 1,000 and 10,000. The program automatically reports a *q*-value as described in

Benjamini and Hochberg 1995). Loci were considered under selection with this method with a FDR of 5% (q-value < 0.05).

Allele frequency, bioclimatic variables, habitat suitability and habitat suitability changes

The second group of analyses (Box II in **Figure 3.2**) consisted of investigating the individual influence of different environment variables on the allele frequencies across populations. The investigated environment variable were: (**i**) each of the bioclimatic variables used to calibrate the SDMs for each species (period 1970-2000); (**ii**) the predicted habitat suitability values during both 1970-2000 (P_S) from the individual SDMs; and (**iii**) the changes in habitat suitability from the LGM (Δ_S). Those analyses were conducted by a correlation analysis in Bayenv (for *I. ricinus*) and by a regression analysis for both species, as described below.

First, and only for *I. ricinus*, Bayenv was used to identify correlations between the environmental variables and allele frequencies. The method does not allow the test of multiple variables at the same time, so each of the environment variables (bioclimatic variables, P_S, and Δ_S) were tested separately. In addition to the Bayes factor (BF) estimate, the method also calculates the Spearman's ρ correlation coefficient for comparison. Since SNPs with high BF may be affected by outliers, it is advised to compare BF values to ρ values. If high BF values are also supported by high ρ values, the signal is considered to be robust (Günther & Coop, 2018). In any case, a BF value between 3 and 10 is generally considered as substantial evidence for the alternative hypothesis (in the present study, the correlation, either positive or negative, between allele frequencies and one environmental variable), while a BF higher than 10 is considered strong evidence for the alternative hypothesis (Kass and Raftery 1995, Wetzels *et al.* 2011). All loci that had a BF value higher than 3 were considered as significantly correlated to the focal environmental variable. For both the covariance matrix estimation and the environmental correlation, 10,000 iterations were run in Bayenv. Finally, for all the significant loci identified across the different methods, a BLAST query was conducted on the sequences

encompassing the variable base (i.e., the SNP context) to search for possible genes that could be linked to adaptation to climatic conditions.

Next, and for both species, I ran a separate linear regression model for each allele of each locus, with the focal allele frequency being the response variable and P_S and Δ_S being the two explanatory variables, such that $f_{ij} \sim P_S + \Delta_S$, where f_{ij} is the frequency of the allele i at locus j. For *I. ricinus*, the frequency of one allele being equal to 1 minus the frequency of the other allele (cf. bi-allelic SNPs), only one allele per locus was regressed. This regression approach helps to identify how changes in habitat suitability conditions between the LGM and 1970-2000 influence allele frequencies across populations, given a sufficient number of generations for this influence to express itself on allele frequencies. In other words, it helps identifying potential loci under bioclimatic selection. Next, I ran separate linear models of each allele of each locus, but this time with each bioclimatic variable used to calibrate the SDMs as explanatory variables one at a time, such as $f_{ij} \sim Bio_n$, where Bio_n is one focal bioclimatic variable. Noteworthy, allele frequencies may exhibit a spatial pattern (cf. the next section on Morans' *I* correlograms). For instance, given IBD, there is a general tendency for neighbouring populations to exhibit somewhat similar allele frequencies, i.e., a positive spatial autocorrelation signal. To account for this potential bias in the linear regression analyses, residuals of each regression were tested for spatial autocorrelation by the Moran's *I* test implemented in the R package *spdep* package (lm.morantest) (Bivand et al., 2013). Whenever this test was significant (i.e., spatial autocorrelation is present in the residuals), a simultaneous autoregressive (SAR) lagged model was applied. For both the Moran's *I* test and the SAR lagged model, a list of weights of connections between populations was built based on a Gabriel graph. In this sense, regression results were compared to results from existing methods (detailed hereafter) for the identification of loci under abiotic selection, according to the type of genetic marker under analyse.

For both species, the mean frequencies of the alleles significantly correlated with P_S and Δ_S were compared between genetic clusters. If the genetic structure described in chapter 2 are a

consequence of a post-Pleistocene range expansion for the two studied species, I expect a significant difference in the allele frequencies between the Northern and Southern clusters of the alleles correlated with the change in habitat suitability conditions between the LGM and the 1970-2000 period, i.e. Δ_S. In the other hand, if the observed genetic structure is a consequence of near present climatic conditions, a significant difference in allele frequencies between the two clusters should be observed for the alleles correlated with P_S. If the climatic conditions have no influence in the genetic structure of the studied populations (neither during the LGM nor during 1970-2000), then no significance should be observed. To test the differences in allele frequencies between the two clusters as described above, I applied a Mann-Whitney rank test.

Influence of habitat suitability on the spatial autocorrelation of alleles of *Ixodes ricinus* and *Geum urbanum*

In this third group of analysis (see Box III in **Figure 3.2**), I investigated the influence of habitat suitability on the spatial structure of allele frequencies for the two focal species. First, I used Moran's I correlograms to analyse the spatial structure of allele frequencies across the studied populations of *I. ricinus* and *G. urbanum*. As described above, the SNP markers of *I. ricinus* are bi-allelic (and therefore the frequency of the allele *p* is equal to 1-*q*), thus only one allele per locus was investigated. The frequency of all each individual microsatellite allele was investigated in the case of *G. urbanum*. The number of distance classes k to compute a Moran's *I* correlogram is a delicate choice and represents a compromise between the resolution and the power of the test statistics (Legendre & Legendre, 1998). Although a higher number of distance classes will allow to analyse spatial autocorrelation signals at a much finer spatial resolution, the number of population pairs or connections in a given distance class will be reduced, hence reducing the statistical power of the analysis. It is also important that distance classes have a similar number of connections so that the standard error of the Moran's *I* is comparable among classes (Diniz-Filho et al., 2016). In the present study, for both *I. ricinus* and *G.*

urbanum, eight distance classes were kept, eight being the maximum number of distance classes after which the number of connections became unbalanced between distance classes. Distances were chosen so that the number of connections was approximately constant among classes: aside from the first class, all classes had the same number of connections, for both species. In addition to analysing the spatial autocorrelation of the allele frequencies for each locus separately, using eight distance classes each time, the amount of spatial structure was also summarised across distance classes by the Sum of Squares of Moran's *I* $\sum I_k^2$ (Kissling and Carl 2007, Diniz-Filho *et al.* 2016), where k corresponds to each distance class. Moran's *I* were computed with the *spdep* R package (Bivand et al., 2013).

The influence of both the contemporary period (1970-2000) and historical changes in habitat suitability over the spatial structure of allele frequencies was then investigated on the basis of the Moran's *I* in each of the eight distance classes. Moran's *I* were regressed on the variation of allele frequencies as a function of P_S and Δ_S (the coefficient estimate of the slope of the regression described at the beginning of this section), and two strictly genetic variables: divergence estimated for each locus (F_{ST}) for the bi-allelic SNPs of *I. ricinus*, or the number of populations in which an allele occurs for the microsatellites of *G. urbanum.* A similar approach has been applied by Diniz-Filho et al. (2016) for disentangling the observed spatial genetic structure of a tropical tree based on microsatellite loci. The regressions were constructed in the form $I_d = \beta P_S + \beta \Delta_S + G$, where I_d is the Moran's *I* across loci in the distance class d, β is the coefficient estimate of the slope of the regression of allele frequencies and the habitat suitability measures, and G is the genetic variable according to the type of molecular marker (SNP or microsatellite). The same independent variable was also used in a regression of the sum of squares of Moran's *I*, a synthetic measure of the spatial autocorrelation in allele frequencies. AIC values of the complete model and of models with only a combination of the independent variables were compared. If the contemporary habitat suitability or the change in habitat suitability since the LGM are important variables defining the spatial autocorrelation structure, the β coefficients should be significant.

Habitat suitability and genetic structure

To directly test the influence of habitat suitability (both past and present) on the genetic structure identified previously (chapter 2) by STRUCTURE (*I. ricinus* and *G. urbanum*), DAPC (*I. ricinus*) and PCoA (*G. ubanum*), I regressed the populations' probabilities of assignment to one of the two clusters against P_S and Δ_S (see Box IV in **Figure 3.2**). The STRUCTURE analysis pipeline with Structure Harvester already give an average probability of assignment for each population, while multivariate methods calculate values for at least two principal components (axes). Also, the DAPC method calculates probabilities of assignment for each multilocus genotypes (individuals) instead of populations, and thus I calculated population probabilities as the average of the individual probabilities within each population. For the multivariate methods, I conducted the regression analysis with values from the first principal components as dependent variables. Residuals from the linear regressions were tested for spatial autocorrelation using Moran's *I* tests implemented in the *spdep* R package (Bivand et al., 2013) and using the same eight distance classes as described above.

Finally, I applied a cophenetic approach to compare visually how environmental distances (habitat suitability) among populations are related to genetic distances. I used the Ward's hierarchical clustering method to construct dendrograms based either: (**i**) on the distance of present habitat suitability (P_S); (**ii**) on the distance of changes in habitat suitability (Δ_S); (**iii**) on the genetic distances among sampled populations using the unbiased parameter θ (Weir and Cockerham, 1984). A cophenetic correlation (Sokal and Rohlf 1962) between all possible pairs of dendrograms was then calculated to assess how similar the genetic and the habitat suitability dendrograms are. A cophenetic correlation coefficient is a measure of how well a dendrogram represents original distance matrix, or how similar two dendrograms are with each other. This last approach helps understanding qualitatively how changes in habitat suitability mimic the overall genetic structure of populations observed for both species. All the hierarchical clustering analyses were implemented in R with the stats package.

Results

Species Distribution Models

For both studied species (*G. urbanum* & *I. ricinus*), model performances ranked from good to excellent. AUC values ranged from 0.956 to 0.960 for *I. ricinus* and from 0.876 to 0.888 for *G. urbanum*. Similarly, TSS values varied from 0.771 to 0.777 for the former and from 0.557 to 0.597 for the later. Finally, CBI values ranged from 0.925 to 0.991 for *I. ricinus* and were more variable across runs for *G. urbanum*, ranging from 0.721 to 0.907. For both studied species, SDMs showed a shift in habitat suitability conditions from a south-western distribution during the LGM to a central-northern distribution during 1970-2000 (**Figure 3.3** and **Figure 3.4**). The projected distribution of *I. ricinus* during 1970-2000 (**Figure 3.3**a) is more or less similar to the present species distribution (see **Figure 1.8** of the chapter 1). The projected distribution of *G. urbanum* (**Figure 3.4**a) seems, however, more narrowly distributed across Europe than the present species distribution, probably as a result of the biased spatial occurrence distribution and the lack of important abiotic variables, such as edaphic parameters.

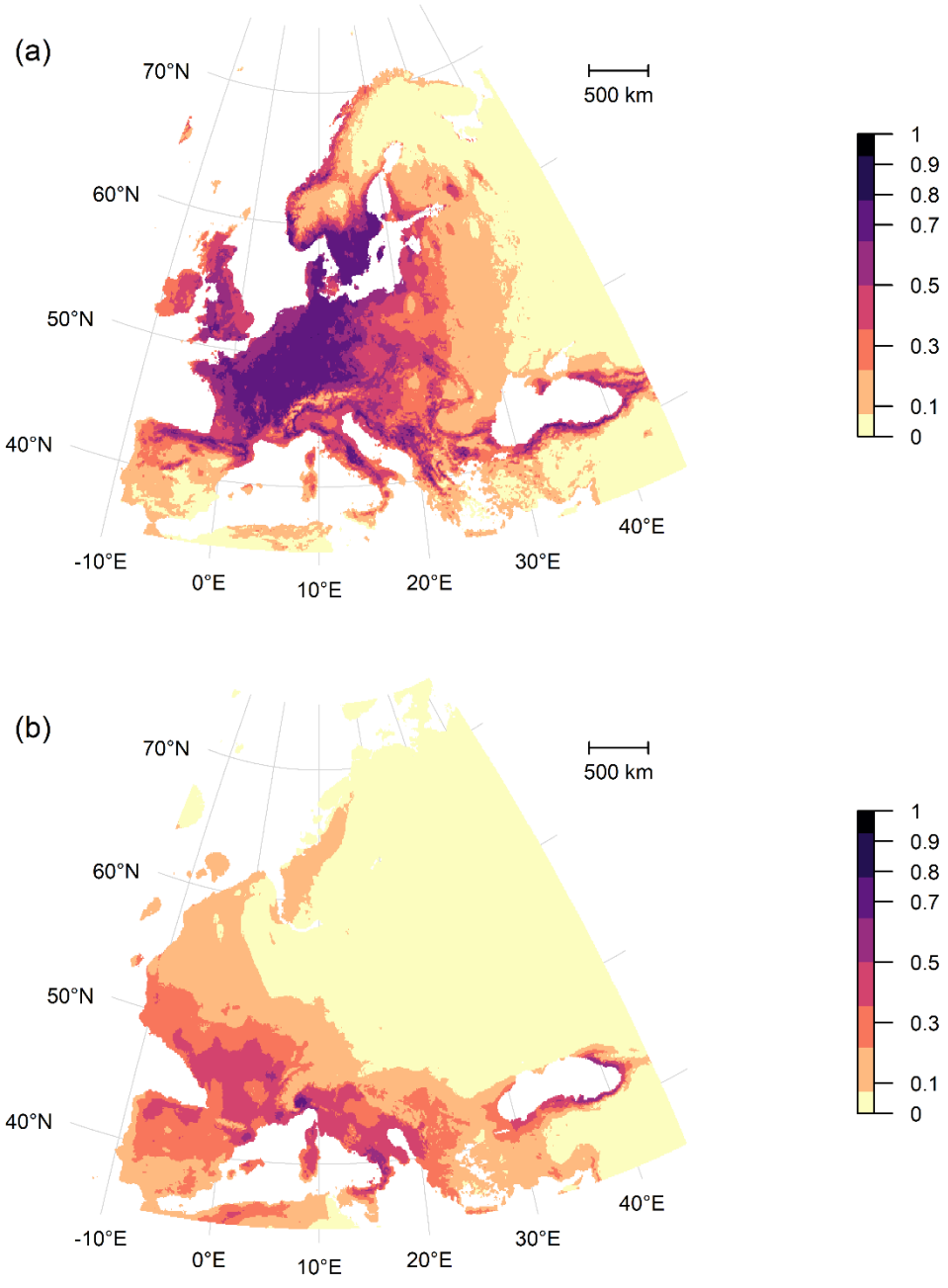


Figure 3.3. Predicted habitat suitability values (0-1) for *Ixodes ricinus* during 1970-2000 (a) and the LGM (b).

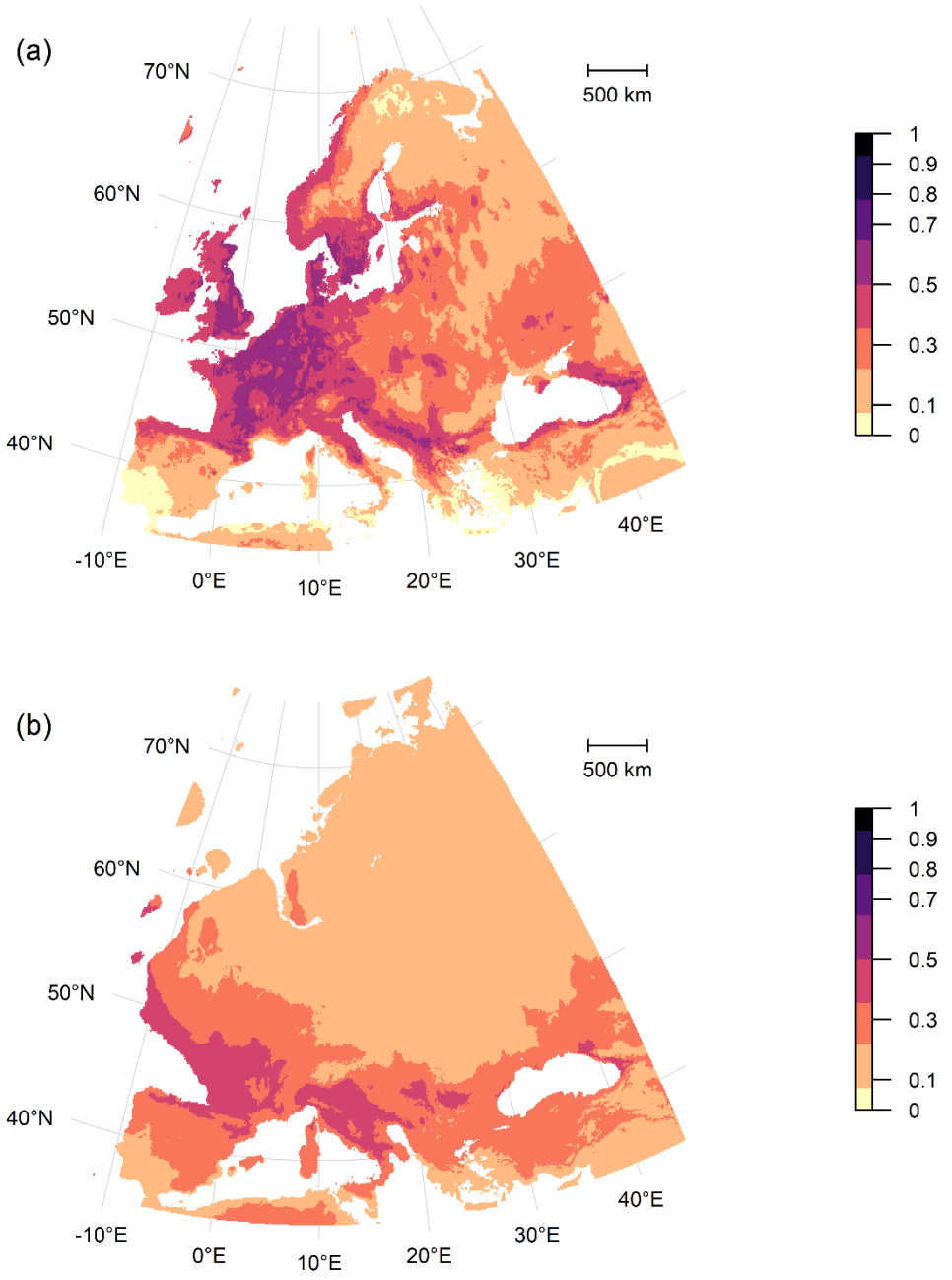


Figure 3.4. Predicted habitat suitability values (0-1) for *Geum urbanum* during 1970-2000 (a) and the LGM (b).

Candidate loci under selection

For *I. ricinus*, the Bayenv analysis identified 6 loci as outliers, with high values of the X^TX indices (greater than 25.5, **Table 3-1**). This is a much smaller number of outliers than the 18 loci identified by pcadapt, even after *p*-values correction. For *G. urbanum*, only one locus was identified as an outlier by BayeScan (WGU228, *q*-value = 0.0020), when the 'Prior odds of neutral model' parameter was set to 1,000 or 10,000. Setting this parameter to 100 suggests 3 loci under selection: WGU228 (*q*-value < 0.0001); WGU210 (*q*-value = 0.0078); and WGU248 (*q*-value = 0.0023).

Allele frequencies as a function of habitat suitability and habitat suitability changes

When considering the correlations with environmental variables, there was a remarkable coincidence on the loci varying significantly with environmental variables for *I. ricinus* across the different approaches I used: the pcadapt; Bayenv and the regressions of allele frequencies on habitat suitability on 1970-2000 (P_S) and the difference in habitat suitability to the LGM (Δ_S). Eleven loci were significantly correlated to environmental variables according to the Bayenv approach (**Table 3-1** and **Figure 3.6**). The BIO 12 bioclimatic variable (annual precipitation) was the only variable not significantly correlated with allele frequencies according to Bayenv. More important, Δ_S was correlated to 5 of the previous loci according to Bayenv, and had a significant influence on nineteen out of the 125 SNP loci according to the regression approach (**Figure 3.5**). All the loci potentially under selection according to the two previous methods also responded significantly to Δ_S. Habitat suitability conditions during 1970-2000 or P_S, on the other hand, showed very different influence over loci. Bayenv did not identify any loci correlated to P_S, and the regression analysis identified a significant influence on 23 loci, from which only three were identified by the other methods. After a BLAST query on the SNP context of the loci potentially under selection, six of them were found (**Table 3-1**), with the query cover varying from 93% to 100%. Allele frequencies of five of those loci varied significantly with Δ_S, and only one was also correlated to the mean annual temperature (BIO 1) according to both Bayenv. The sixth loci only showed a significant relationship to P_S. Finally, the frequencies of the loci positively correlated with

the change in habitat suitability or Δ_S were significantly higher in the Northern populations than in the Southern populations (Mann-Whitney rank test, p = 0.0002), while no such difference in allele frequencies between both clusters was observed when looking at allele frequencies of the loci positively correlated with habitat suitability conditions during 1970-2000 or P_S (Mann-Whitney rank test, p = 0.6165) (**Figure 3.7**). For *G. urbanum*, the frequencies of seventeen alleles varied significantly with P_S, Δ_S, or one of the five bioclimatic variables used to build the SDMs (**Table 3-2**). As for *I. ricinus*, the mean frequencies of the alleles positively correlated with Δ_S were higher in the Northern populations, but this represented only 6 alleles.

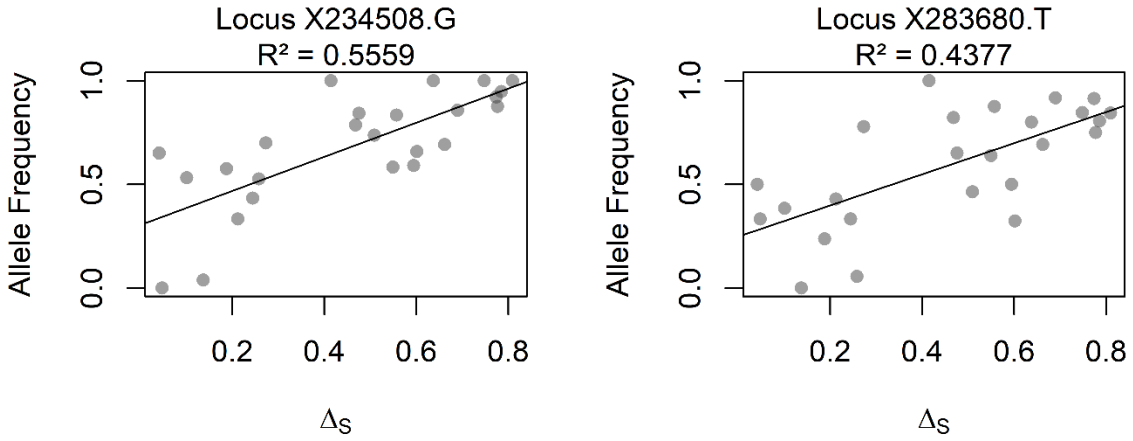


Figure 3.5. Variation in the allele frequencies of two of the SNP loci significantly correlated to Δ_S for *I. ricinus*. The letter after the locus name correspond to the nucleic acid base of the plotted allele, 'G' for guanine and "T" for thymine. The relation is inversed for the other allele.

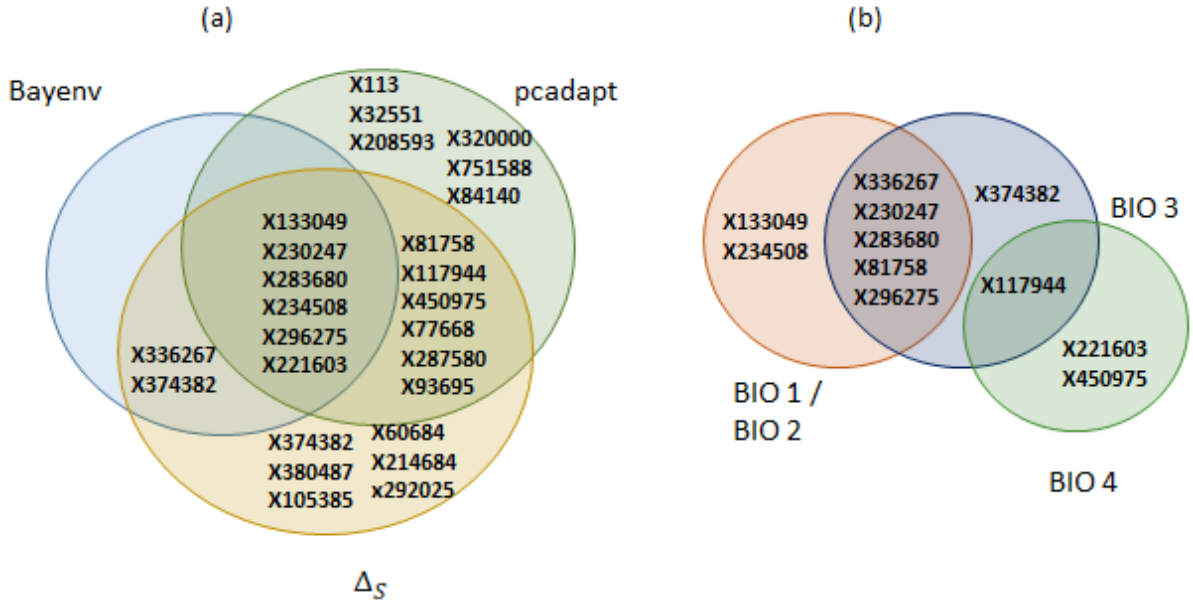


Figure 3.6. SNP loci of *I. ricinus* potentially under selection according to Bayenv, pcadapt and the regression of allele frequencies on changes in habitat suitability from the LGM to the period 1970-2000 (a), and loci correlated to four bioclimatic variables (BIO 1, BIO 2, BIO 3, and BIO 4) according to the Bayenv approach (b).

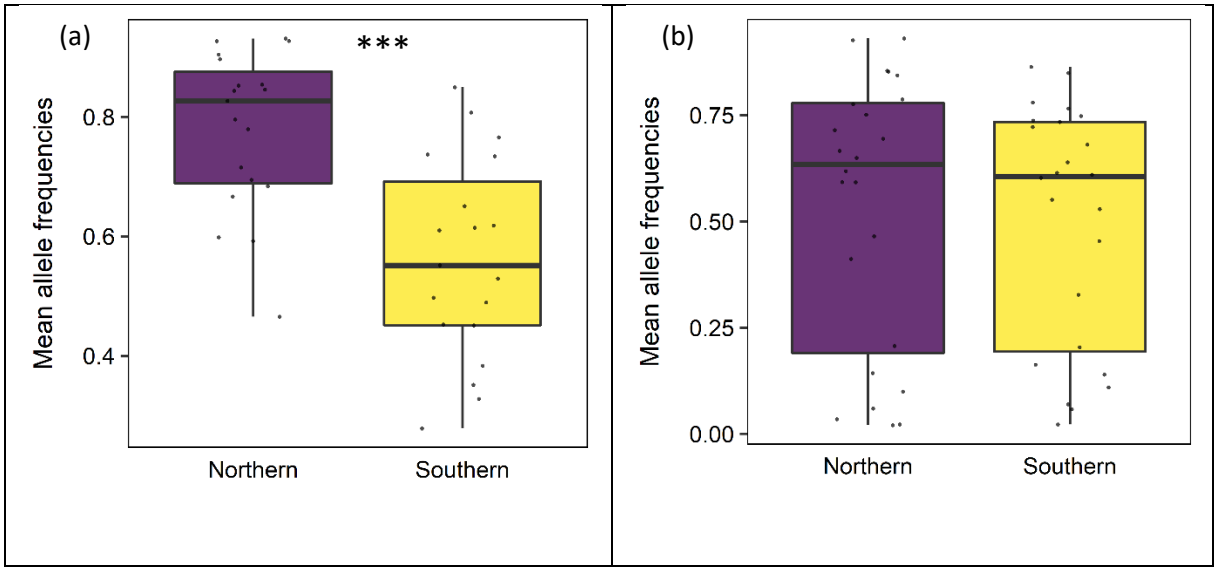


Figure 3.7. Frequencies of the 19 alleles from the 125 bi-allelic SNPs of *I. ricinus* positive and significantly correlated with changes in habitat suitability Δ_S from the LGM to the contemporary conditions (period 1970-2000) (a) and with contemporary habitat suitability (P_S) in the two Northern (violet) and Southern (yellow) genetic clusters (Mann-Whitney rank test, p = 0.0002 (a) and p = 0.6165 (b)).

Spatial autocorrelation and influence of habitat suitability

For *I. ricinus*, Moran's *I* correlogram exhibited a significant autocorrelation signal in the first two distance classes (0 to 800 km) for about 21% of the investigated loci. More precisely, 27 loci showed a significant positive signal in at least one of the two first distance classes, with 17 and 18 loci in the first and second distance class, respectively, of which 8 loci showed significant positive autocorrelation in both classes. The mean values of Moran's *I* across loci tended to gradually decrease after the second distance class (**Figure 3.8**), although the number of loci that showed a significant signal per distance class tended to increase continuously after the sixth distance class. This was almost exclusively due to the increase in the number of loci having significantly negative Moran's *I* values in the last distance classes (**Figure 3.9**), indicating a negative autocorrelation signal of allele frequencies between distant populations.

For *G. urbanum*, most of the significant autocorrelation values were observed in the second, third (12 out of 89 alleles in both classes) and eighth distance classes (9 out of 89 alleles) (**Figure 3.10**). This increase in the number of alleles showing significant Morans' *I* values in the last distance classes was, again, almost exclusively due to the increase in the number of significantly negative values (**Figure 3.11**), indicating a negative autocorrelation signal of allele frequencies between distant populations.

For both species, habitat suitability conditions during 1970-2000 did not affect Moran's *I* values across distance classes nor the sum of squares of Moran's *I* values. Actually, AIC values of the models including P_S were always higher and regression coefficient values were always smaller than the models without this variable. This way, the regression results presented here does not take into account the present habitat suitability (P_S). For *I. ricinus*, regression coefficients varied greatly between distance classes, from 0.07 in the last distance class to 0.23 in the sixth distance class. F_{ST} was highly significant in the first, second, sixth, and seventh distance classes (values of *p* varying from <0.0001 to 0.0077), while Δ_S was only significant in the sixth distance class. The sum of squares of Moran's *I* varied significantly with F_{ST} ($p < 0.0001$) and Δ_S ($p = 0.0204$), and the adjusted R^2 of the regression was 0.3422. For *G. urbanum*,

regression coefficients varied from 0.01 in the third distance class to 0.13 in the first distance class. The number of populations in which the allele was present was the most important variable, being significant in the first, second, seventh, and eighth distance classes (values of *p* from 0.0002 to 0.0331), while Δ_S was only significant in the first distance. The sum of squares of Moran's *I* values varied significantly with the number of populations in which the allele was present ($p < 0.0001$) and with Δ_S ($p = 0.0102$), and the adjusted R^2 was 0.3859.

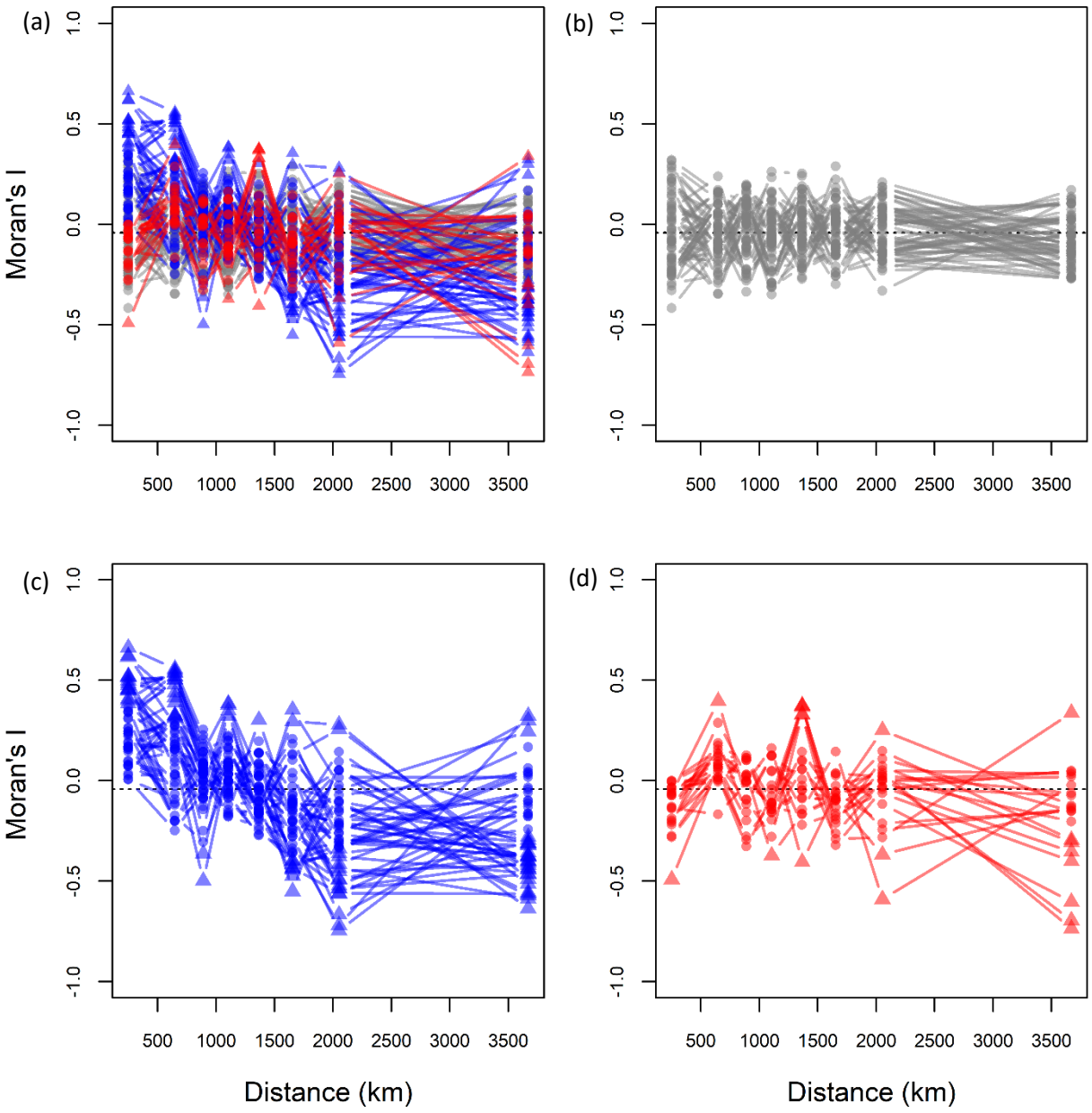


Figure 3.8. Moran's *I* values for 125 SNP loci (a) across 25 populations of *I. ricinus* in Eurasia. Grey lines in (b) show only the loci that exhibited no autocorrelation signal in the eight distance classes; blue lines in (c) show only the loci that exhibited a positive and significant autocorrelation signal in the first distance class (0-500 km); red lines in (d) show the loci that exhibited a negative and significant autocorrelation signal in the first distance class. Triangles indicate a significance at 5% ($p < 0.05$) for one particular locus in one distance class.

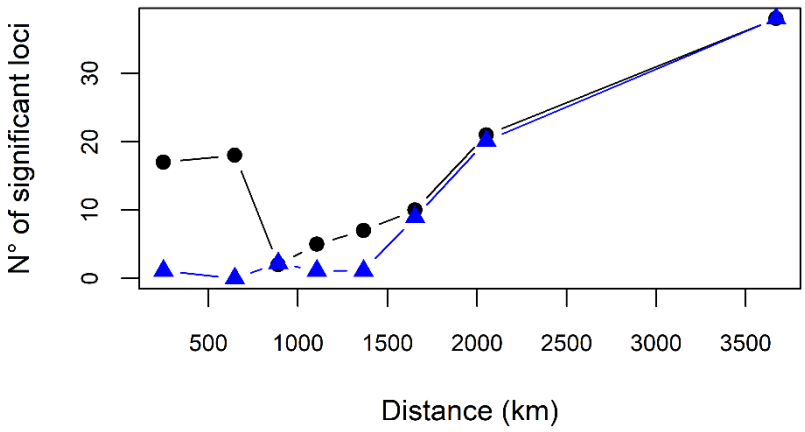


Figure 3.9. Number of significant Moran's *I* values in each of the eight distance classes based on the correlograms of allele frequencies for 125 bi-allelic SNPs across the 25 sampled populations of *I. ricinus.* Blue triangles and the blue line show the number of significantly negative Moran's *I* values while black dots and the black line show the number of significant Moran's *I* values being either positive or negative. The number of significant values of Moran's *I* increases rapidly after the sixth distance class, mostly due to significant negative values, while the first distance classes are dominated by significantly positive values.

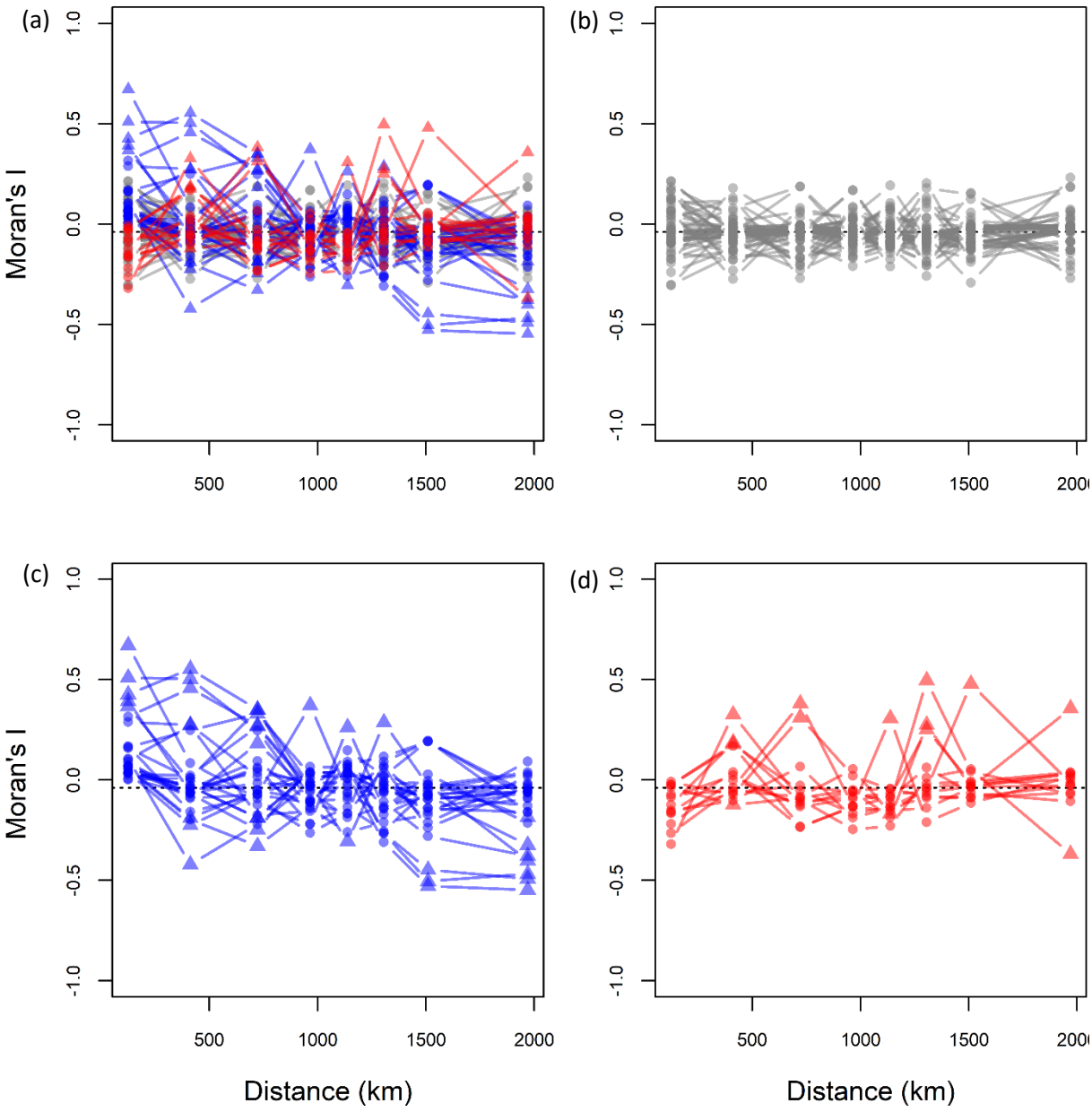


Figure 3.10. Moran's *I* values for 89 microsatellites alleles (a) across 27 populations of *G. urbanum* in Europe. Grey lines in (b) show only the loci that exhibited no autocorrelation signal in the eight distance classes; blue lines in (c) show only the loci that exhibited a positive and significant autocorrelation signal in the first distance class (0-412 km); red lines in (c) show the loci that exhibited a negative and significant autocorrelation signal in the first distance class. Triangles indicate a significance at 5% ($p < 0.05$) for one particular locus in one distance class.

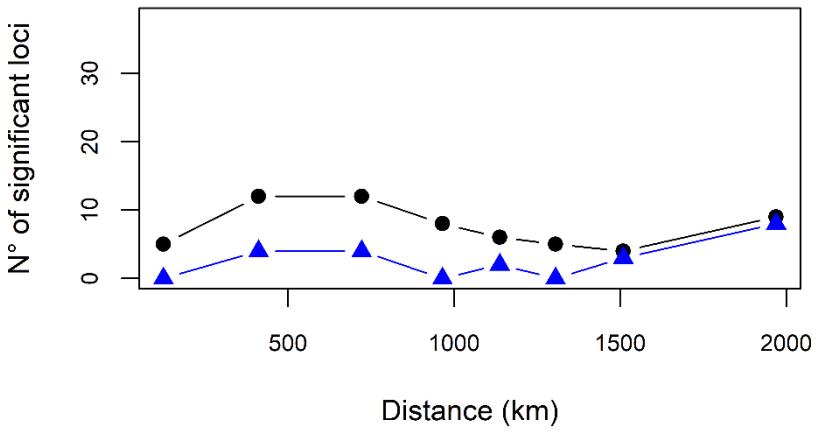


Figure 3.11. Number of significant Moran's *I* values in each of the eight distance classes based on the correlograms of allele frequencies for the 89 microsatellites investigated across the 27 sampled populations of *G. urbanum.* Blue triangles and the blue line show the number of significantly negative Moran'*I* values while black dots and the black line show the total number of significant Moran's *I* values being either positive or negative. The number of significant values of Moran's *I* reach the maximum at the second and third distance classes.

Genetic structure as a function of habitat suitability

For both species, no autocorrelation was observed in the residuals of the multiple regressions of population probabilities of assignment (or principal coordinates) and habitat suitability, and thus I kept the ordinary least square regressions. For *I. ricinus*, probabilities of assignment from Structure had a significant relationship with Δ_S (p < 0.0001), but none with P_S (p = 0.1460), with an adjusted R^2 of 0.5153. Similar results were observed for the within populations mean individual probabilities coordinates from the DAPS (Δ_S p-value < 0.0001; P_S p-value = 0.1780; adjusted R^2 = 0.5224).

For *G. urbanum*, habitat suitability during 1970-2000 and changes in habitat suitability between the LGM and 1970-2000 had no significant influence over the populations probabilities of assignment from Structure (Δ_S p-value = 0.418; P_S p-value = 0.3880), and the adjusted regression coefficient was surprisingly small (R^2 = 0.0881). The regression of the first principal coordinates of the PCoA, on the other hand, showed a significant influence of Δ_S (p = 0.0227) and no significance of P_S (p = 0.6741), with an adjusted R^2 of 0.3147.

Hierarchical clustering analysis

The topologies of hierarchical clustering dendrograms constructed from genetic or habitat suitability distances (**Figure 3.12** and **Figure 3.13**) closely mirrors the overall genetic structure identified for both species (see sections '*Published article: Strong genetic structure among populations of the tick Ixodes ricinus across its range*' and '*Population Genetic Structure of Geum urbanum*' for *I. ricinus* and *G. urbanum*, respectively). For *I. ricinus,* the genetic dendrogram clearly separates the Eurasian populations into two genetic branches, corresponding to the Southern and Northern clusters identified by both STRUCTURE and DAPC methods. For *G. urbanum*, the genetic dendrogram reproduces nearly exactly the genetic structure inferred from both STRUCTURE and PCoA analyses. Two particular differences are worth noticing: the Belgium population BEL_1 that was previously placed in the Southern cluster belong now to the Northern cluster, and the Latvian population LVA_1 that was situated in the Northern cluster belongs now to the Southern one.

The dendrogram based on habitat suitability during 1970-2000 (P_S) does not reflect the genetic structure of *I. ricinus* (**Figure 3.12**). On the other hand, the dendrogram based on the changes in habitat suitability since the LGM better mirrors the genetic structure of I. *ricinus*, grouping together geographically distant but genetically close populations, e.g. the group formed by the Iranian (IRN), the Southern France (FRA-S), the Turkish (TUR) and the Italian (ITA-V) populations, and the marked proximity between the Western France (FRA-W) and Spanish (ESP) populations. The cophenetic correlation (**Table 3-3**) between the Δ_S and the genetic dendrogram was higher (0.4492) than the one estimated between P_S and the genetic dendrograms (0.0960).

For *G. urbanum*, similar to the results for *I. ricinus*, the dendrogram based on habitat suitability during 1970-2000 (P_S) did not reproduce the global structure found with genetic distances, while the dendrogram based on Δ_S distances group populations from the Northern and Southern clusters more closely (**Figure 3.13**). Two notable exceptions are the Norwegian and Estonian populations (NOR_4 and EST_1). It seems that the spatial location where those populations are found nowadays have

experienced similar changes in habitat suitability since the LGM, which are not mimicked by the genetic distance between those populations. Again, the cophenetic correlation (**Table 3-3**) between the Δ_S and the genetic dendrograms was higher than the one estimated between P_S and the genetic dendrograms (0.6763 and 0.3476, respectively).

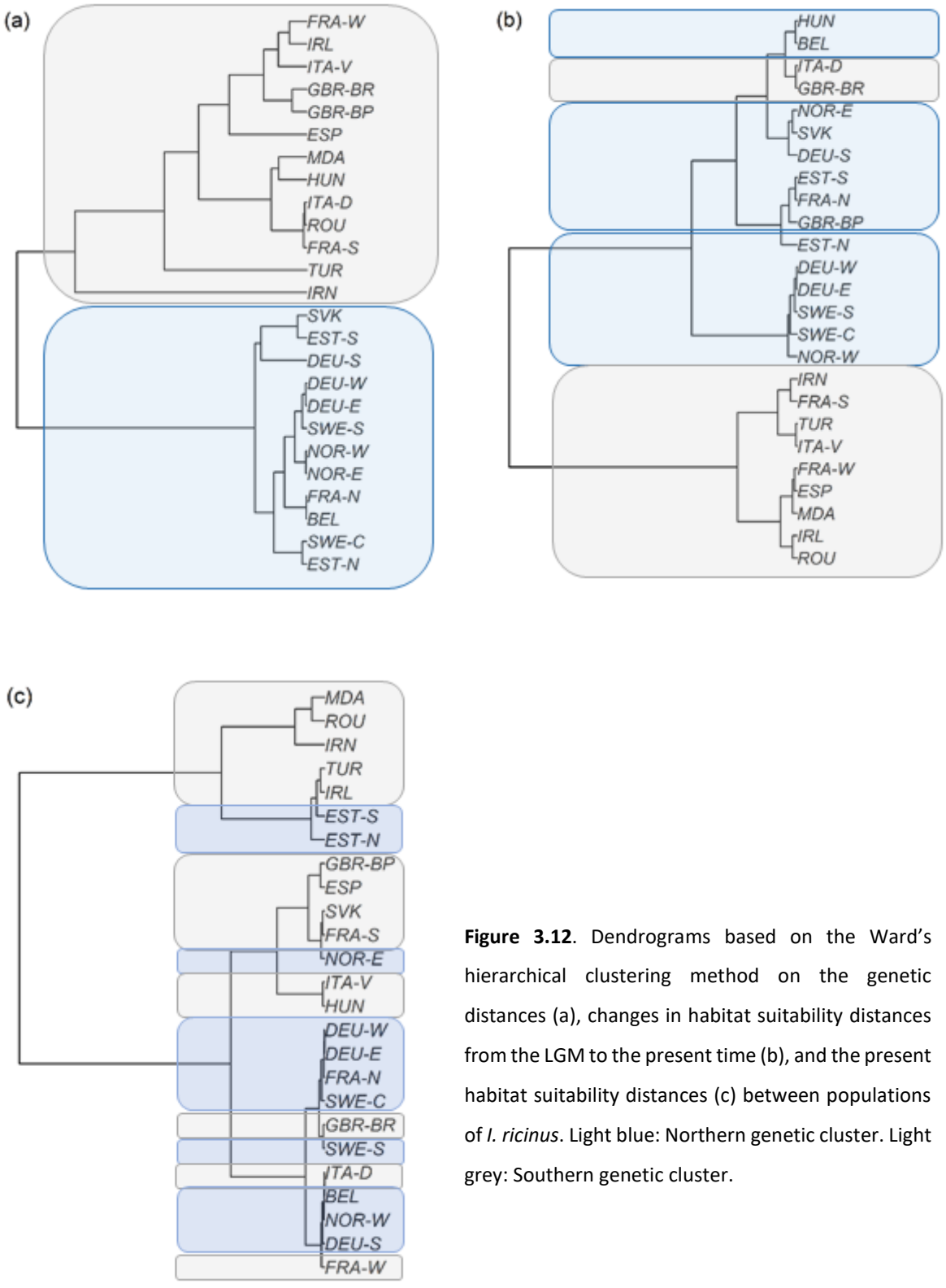


Figure 3.12. Dendrograms based on the Ward's hierarchical clustering method on the genetic distances (a), changes in habitat suitability distances from the LGM to the present time (b), and the present habitat suitability distances (c) between populations of *I. ricinus*. Light blue: Northern genetic cluster. Light grey: Southern genetic cluster.

Figure 3.13. Dendrograms based on the Ward's hierarchical clustering method on the genetic distances (a), changes in habitat suitability distances from the LGM to the present time (b), and the present habitat suitability distances (c) between populations of *G. urbanum.* Light green: Northern genetic cluster. Light orange: Southern genetic cluster.

Table 3-1 *(cont.)*. SNP loci of *I. ricinus* exhibiting a significant relationship between allele frequencies across 25 populations showing: the Bayes Factor (BF) from the correlation analysis with Bayenv in columns Δ_S, BIO 1, BIO 2, BIO 3, and BIO 4; the X^TX statistic from Bayenv for the loci showing the 5% highest values; the *p*-values from the linear regression with present habitat suitability P_S and the differences in habitat suitability between present and LGM climatic conditions Δ_S; and, the candidate loci under selection according to the pcadapt method. Results of the BLAST query for those loci are shown in the last four columns. The two *p*-values in bold are the results of the lagged models.

Locus	Significance tests									GenBank records			
		Bayenv						Regressions					
	pcadapt	X^TX	Δ_S	Bio1	Bio2	Bio3	Bio4	P_S	Δ_S	Species	Gene	Locus	Query cover
X133049	<0.0001	25.50	4.43	5.47	5.47				0.001				
X336267				3.97	3.97	4.61			**0.043**				
X230247	<0.0001	26.52	4.22	4.53	4.53	9.39			0.002				
X283680	<0.0001	27.49	5.18	6.01	6.01	9.54			0.001				
X234508	<0.0001	33.02	17.35	7.91	7.91				**0.022**	*I. scapularis*	G protein-coupled receptor kinase 1	XM_029970359.1	96%
X81758	<0.0001			3.65	3.65	12.98			0.002				
X296275	<0.0001	29.61	10.98	8.59	8.59	4.95			0.044				
X221603	0.0009	25.78					8.71		0.037				
X117944	0.0009					4.44	20.38	0.003	0.001				
X450975	0.0003						3.27	0.002	0.004				
X374382						8.48		0.049	0.034				
X198227								0.017					
X251320								0.003					

Table 3-1 *(cont.)*. SNP loci of *I. ricinus* exhibiting a significant relationship between allele frequencies across 25 populations showing: the Bayes Factor (BF) from the correlation analysis with Bayenv in columns Δ_S, BIO 1, BIO 2, BIO 3, and BIO 4; the X^TX statistic from Bayenv for the loci showing the 5% highest values; the *p*-values from the linear regression with present habitat suitability P_S and the differences in habitat suitability between present and LGM climatic conditions Δ_S; and, the candidate loci under selection according to the pcadapt method. Results of the BLAST query for those loci are shown in the last four columns. The two *p*-values in bold are the results of the lagged models.

Locus	Significance tests									GenBank records			
	pcadapt	Bayenv						Regressions					
		X^TX	Δ_S	Bio1	Bio2	Bio3	Bio4	P_S	Δ_S	Species	Gene	Locus	Query cover
X225377								0.012					
X380487								0.006	0.008	*I. scapularis*	Uncharacterized	XM_029990341.1	98%
X105385								0.007	**0.016**				
X77668	<0.0001							0.007	0.001	*I. scapularis*	Uncharacterized	XM_029988614.1	100%
X399212								0.040					
X233961								0.015					
X145634								0.019					
X307361								0.018		*I. scapularis*	FRAS-1	XM_029984399.1	98%
X287805	0.0019							0.035	0.034				
X116335								0.010					

Table 3-1 *(cont.)*. SNP loci of *I. ricinus* exhibiting a significant relationship between allele frequencies across 25 populations showing: the Bayes Factor (BF) from the correlation analysis with Bayenv in columns Δ_S, BIO 1, BIO 2, BIO 3, and BIO 4; the X^TX statistic from Bayenv for the loci showing the 5% highest values; the *p*-values from the linear regression with present habitat suitability P_S and the differences in habitat suitability between present and LGM climatic conditions Δ_S; and, the candidate loci under selection according to the pcadapt method. Results of the BLAST query for those loci are shown in the last four columns. The two *p*-values in bold are the results of the lagged models.

Locus	Significance tests									GenBank records			
	pcadapt	Bayenv						Regressions					
		X^TX	Δ_S	Bio1	Bio2	Bio3	Bio4	P_S	Δ_S	Species	Gene	Locus	Query cover
X60684								0.025	0.014				
X313057								0.012		*I. scapularis*	Putative nuclease HARBI1	XM_029967969.1	93%
X214684								0.006	0.002				
X487540								0.024					
X84140	0.0077							0.028					
X446758								0.006					
X200386								0.003					
X783090								0.011					
X225801								0.013					
X93695	0.0086								0.004	*I. scapularis*	Neprilysin-2	XM_029976277.1	96%
X292025									0.015				

Table 3-1 *(cont.)*. SNP loci of *I. ricinus* exhibiting a significant relationship between allele frequencies across 25 populations showing: the Bayes Factor (BF) from the correlation analysis with Bayenv in columns Δ_S, BIO 1, BIO 2, BIO 3, and BIO 4; the X^TX statistic from Bayenv for the loci showing the 5% highest values; the *p*-values from the linear regression with present habitat suitability P_S and the differences in habitat suitability between present and LGM climatic conditions Δ_S; and, the candidate loci under selection according to the pcadapt method. Results of the BLAST query for those loci are shown in the last four columns. The two *p*-values in bold are the results of the lagged models.

Locus	Significance tests									GenBank records		
		Bayenv						Regressions				
	pcadapt	X^TX	Δ_S	Bio1	Bio2	Bio3	Bio4	P_S	Δ_S	Species	Gene	Locus	Query cover
X1133	<0.0001												
X32551	0.0042												
X208593	<0.0001												
X320000	<0.0001												
X751588	0.0003												

Table 3-2. Significant values of the simple regressions of allele frequencies of six microsatellite loci of *G. urbanum* as a function of the five bioclimatic variables used to calibrate the SDMs, present habitat suitability (P_S), and changes in habitat suitability from the LGM to the present (Δ_S). Values in bold are from the lagged models.

Allele	Bio1	Bio8	Bio12	Bio15	Bio18	P_S	Δ_S
WGU210_201						0.0384	
WGU210_225	0.0065			0.0032		0.0002	0.0012
WGU210_237			0.0000		0.0004		
WGU228_174		**0.0167**					**0.0031**
WGU228_180		**0.0319**					**0.0018**
WGU248_204			0.0011		0.0039		
WGU61_195							**0.0074**
WGU65_216			0.0010		0.0110		
WGU65_234	0.0228					0.0001	0.0001
WGU65_246	0.0048						
WGU67_162	0.0162			0.0082		0.0051	
WGU67_174		**0.0160**					**0.0025**
WGU67_177	0.0169					0.0061	0.0028
WGU67_180		0.0211					0.0378
WGU67_183						0.0486	0.0356
WGU67_204		0.0206	0.0030		0.0278		
WGU67_219	0.0017						

Table 3-3. Cophenetic correlations between dendrograms of genetic (Θ), changes in habitat suitability from the LGM to the present (Δ_S), and present habitat suitability (P_S) for *Ixodes ricinus* (upper triangle) and *Geum urbanum* (lower triangle).

G. urbanum \ *I. ricinus*	Θ	Δ_S	P_S
Θ		0.4492	0.0960
Δ_S	0.6763		0.2373
P_S	0.3476	0.5758	

Discussion

Coupling habitat suitability predictions from SDMs with genetic data on both allele frequencies and probabilities of assignment to a given cluster allowed to identify alleles under selection and improve our understanding of the potential drivers underlying the current spatial genetic structure of two forest-dwelling species widely distributed in Europe: the castor tick *Ixodes ricinus* and the wood avens *Geum urbanum.* Besides the differences in the biology of the two studied species and the type of genetic markers employed to disentangle their genetic structure, the results suggest that the genetic structure of both species across Europe is likely very much influenced by the post-LGM expansion dynamics, as I initially assumed. Habitat suitability conditions during the contemporary period of 1970-2000 seem to be of much less importance to explain the current spatial genetic structure of *I. ricinus* and *G. urbanum*.

For both species, SDMs' predictions indicate a northward range shift from the main southern refugia, chiefly located across the Italian peninsula, around the Black Sea as well as in south-western France and the north of the Iberian Peninsula. For *I. ricinus*, the distribution projected during the LGM match and support a former study on the past distribution of *I. ricinus* (Porreta et al., 2013). During the LGM, the species range was probably restricted to the southern limits of its current range, from the north of the Iberian Peninsula to the south-eastern coasts of the Black Sea, with some other favourable

and more cryptic glacial refugia located slightly to the north at the foothill of the French Alps. For *G. urbanum*, we are not aware of any previous work projecting the species distribution during the LGM. As for *I. ricinus*, predictions also suggest that *G. urbanum* was restricted to the southern limits of its current range during the LGM, with four main glacial refugia located in south-western France, the Italian peninsula, the north of the Balkan peninsula and on the eastern side of the Black Sea.

The genetic structure observed for both species, indicating a Northern and a Southern cluster, seem to be largely influenced by long-term climatic changes since the LGM. Probabilities of assignment for *I. ricinus* populations are strongly correlated with changes in habitat suitability between the LGM and 1970-2000, irrespective of the clustering method used to generate the probabilities of assignment (STRUCTURE and DAPC). This interpretation of a potentially strong influence of past climate changes since the LGM is less obvious for *G. urbanum*, since the assignment probabilities from STRUCTURE do not seem to correlate with the change in habitat suitability conditions since the LGM. There is however, for *G. urbanum*, a significant influence of the deep-time change in habitat suitability conditions over the coordinates from the first principal components of the PCoA. The matching topologies of the Ward's dendrograms are another strong indication that the genetic differentiation between the two clusters is (at least partially) a result of the post-LGM expansion dynamic. For both species, the dendrogram based on habitat suitability differences between 1970-2000 and the LGM reproduced with a high level of fidelity the topology of the dendrogram based on the pairwise genetic distances. Those dendrograms differentiated from one another mostly inside the two main clusters, as there are of course other variables than the change in habitat suitability conditions between 1970-200 and the LGM that may influence the local and regional gene flow. Particularly for *I. ricinus*, this analysis has brought to light one of the probable main reasons for the close genetic proximity between geographically very distant populations, notably the genetic proximity between the group of the Iranian and Turkish (IRN and TUR) populations and the group of the Southern French and Italian (FRA-S and ITA-V) populations. Porreta et al. (2013) have already claimed that populations across Southern

Europe during the LGM may have formed a connected metapopulation, and the results presented here support that claim.

Population differentiation can generally be well explained through gradients of (or barriers for) gene flow, drift, directional selection, or a combination of each of these processes (Slatkin 1996, Rieseberg *et al.* 2002, Friesen *et al.* 2007, Hofer *et al.* 2009). Range expansion is considered an important cause of population differentiation. Slatkin (1996) argued that, after the founding event, i.e. after a sample of a small number of individuals from a parental population has already established, the rapid growth of that founding population will relax genetic drift and selection will cause the fixation of low-frequency alleles or the combination of alleles. Although the methods applied here do not aim at precisely disentangling selection from gene flow (and even less the role of founding effects in population differentiation), the spatial genetic patterns obtained clearly suggest that the bioclimatic changes between the LGM and 1970-2000, and as a consequence, the expansion of both species after the Pleistocene, may be in part responsible for the partitioning of populations into a Northern *vs*. a Southern genetic cluster.

For *I. ricinus* in particular, it is noteworthy that alleles significantly correlated to changes in habitat suitability since the LGM are more frequent within the Northern genetic cluster than the Southern one. This suggests a selection pressure over those loci as a consequence of the evolutionary history at least partially independent from the Southern cluster. From the results of this study, this hypothesis seems to be true. The identification of loci under bioclimatic selection is not a simple task, as both the sampling design and the shared history between populations may contribute to the observed patterns in allele frequencies (Excoffier *et al.* 2009, Hancock *et al.* 2011). Here, I applied different but complementary methods to identify those loci and confirms those identifications throughout the prisms of the different methods I used. The results across those different methods are in agreement for some loci, which reinforces the conclusion that those loci are indeed under bioclimatic selection pressure. More precisely, since most of those loci were significantly related to

both bioclimatic variables and changes in the overall bioclimatic niche, it is plausible to suppose that the population across the species' range are under different selection pressures. While a founder-flush effect (Slatkin, 1996) is almost certain, it is likely that population differentiation is also a result of a directional selection. Acknowledging those phenomena is of importance to forecast the future dynamics of species, especially so for vector species like *I. ricinus*, which can carry a lot of diseases of potential threat for human health. In fact, the range of *I. ricinus* is already shifting rapidly northward in latitude and upward in elevations (Lindgren and Gustafson 2001b, Jore *et al.* 2011, Hvidsten *et al.* 2020). These recent evidence of contemporary range shift, coupled with what appears to be a directional selective pressure during past climate changes, might lead to the fixation of new combination of alleles in populations at the leading edge of the shifting range, which could represent new challenges for the human society.

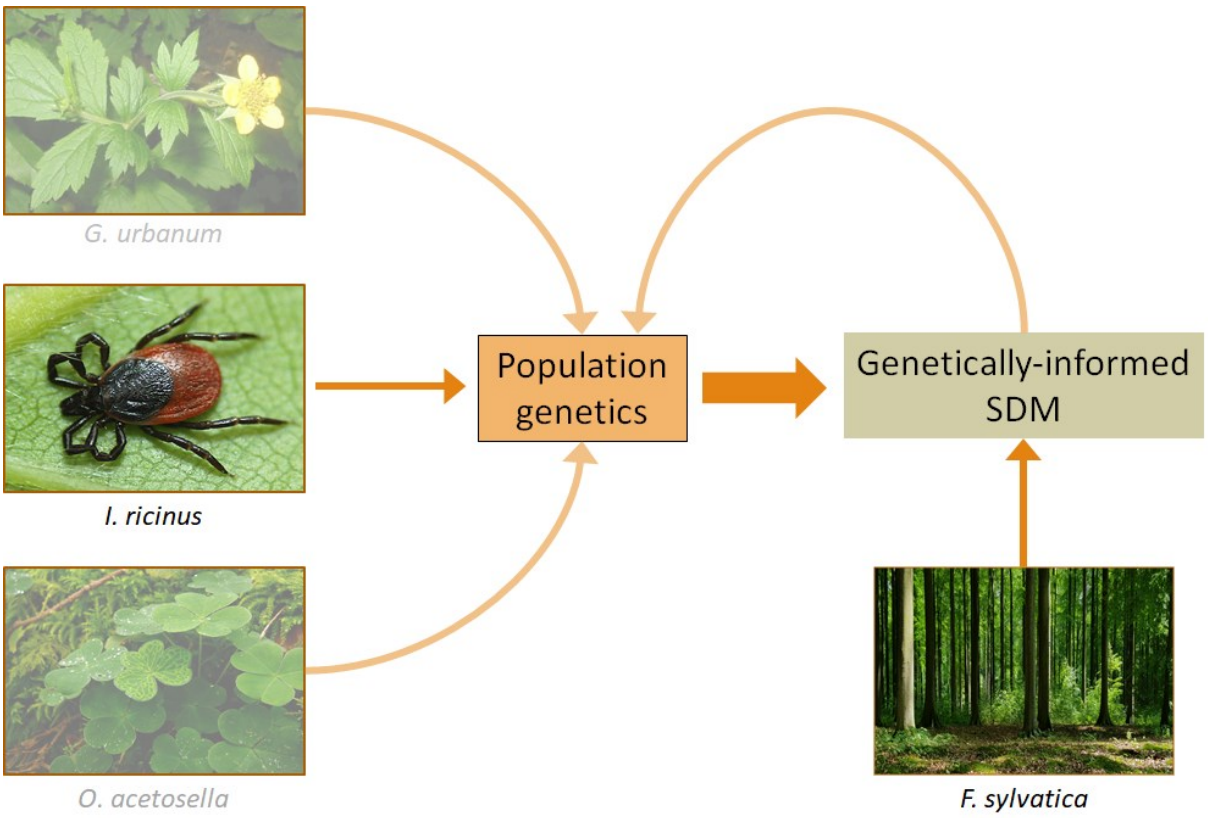
For *G. urbanum*, some of the alleles exhibited a significant relation with the investigated bioclimatic variables and changes in habitat suitability since the LGM, but they represent a very small proportion of the allelic richness of each locus, varying from 10% to 40% of the total alleles in each locus. The higher proportion was found in the WGU228 locus (two significant alleles out of 5). Interestingly this locus was identified as under selection by BayecScan in the three levels of prior odds. Although those results suggest that these loci are under some kind of selection pressure, it is hard to be confident in this at this point.

Another particularity of *G. urbanum* is the high genetic homogeneity of the Northern cluster, as demonstrated by the analysis of multiple values of *K* in STRUCTURE, the lack of correlation between genetic and geographic distance, and the seemingly lack of influence of habitat suitability within the genetic cluster. Regarding those results, it is possible that those populations have differentiated from the Southern populations early on during the northward post-LGM expansion dynamic and since then they did not experience important pressures for differentiation other than genetic drift, as discussed in Chapter 2. In any case, it seems that habitat suitability changes since the LGM or habitat suitability

conditions during 1970-2000 are not enough to explain the within clusters pattern observed for *G. urbanum*.

To conclude, coupling habitat suitability predictions from SDMs with genetic data on both allele frequencies and probabilities of assignment to a given cluster has helped to enlighten some of the mechanisms behind the observed patterns of genetic structure at the European extent. The results from both genetic and spatial analysis strongly suggests that the range expansion dynamic from southern main glacial refugia in Europe is the main cause of the observed differentiation between a Northern and a Southern cluster. At regional and local scales, other factors not regarded here may be influencing the population differentiation. Finally, the quality of the information provided from the present approach seem to depend on the species biology, but also on the type of genetic markers used in the population genetics analyses.

Chapitre 4 Chapter 4: Genetically-informed Species Distribution Models



Presentation

In the two previous Chapters, I explored the population genetic structure of three of the fours model species (Chapter 2) and one approach of applying SDMs to validate hypothesis of the process behind the observed genetic patterns (Chapter 3). In this fourth Chapter, I explore how species-only SDMs may be improved by informing the models of the species' population genetic structure or phylogeography. Two of the four model species were used in this analysis, the tick *I. ricinus* and the tree *F. sylvatica*, for which the phylogeography has already been extensively explored by Magri et al. (2006, 2008). This chapter corresponds to an article entitled 'How does incorporating genetic information improve species distribution models?', submitted to the journal Global Ecology and Biogeography in October 28th 2020. At the time of this manuscript writing, the article was yet under review. For this reason, the article is appended here almost as submitted, including the reference list, and only changing the placement and cross-references of figures and tables for better reading. Supplementary Information cited this Chapter are presented in the **Appendix 2**.

Submitted article: How does incorporating genetic information improve species distribution models?

Poli, P., Lenoir, J., Guiller, A. Submited to Global Ecology and Biogeography.

Introduction

Species distribution models (SDMs) are widely implemented in ecology, biogeography, and conservation biology (Guisan & Thuiller, 2005; Grimmett, Whitsed and Horta, 2020). Most SDMs' applications include one or a combination of the following: (i) supporting conservation planning (Hannah et al., 2007; Guisan et al., 2013); (ii) forecasting or hindcasting species distribution changes in response to climate change (Thuiller et al., 2008; Alkishe et al., 2017); (iii) simulating the spread of invasive alien species (Guisan et al., 2014; Hattab et al., 2017); and (iv) investigating biogeographic and evolutionary hypotheses (Svenning et al., 2008; Ives & Helmus, 2011). While a vast set of modelling techniques are available, most rely on correlative approaches between species presence (and absence or pseudo-absence) data and contemporary environmental variables (Guisan et al., 2017). During the last two decades, many studies investigated how analytical decisions may affect the performance of correlative SDMs, such as the relative proportion of presence vs. absence data (i.e. prevalence) (Pearson et al., 2007; Liu et al., 2018), the resampling strategy (Braunisch et al., 2008; Mainali et al., 2015), the set of predictor variables used for model building (Araújo et al., 2019), the model selection strategy, and the performance of different algorithms (Hallgren et al., 2019).

More recently, some attention has been given to the incorporation of intra-specific variation into correlative SDMs (Pearman et al., 2010; Smith et al., 2018). Incorporating genetic information into correlative SDMs may especially help address concerns in case of low degree of niche conservatism between distant populations (Cooper et al., 2010; Schulte et al., 2012; Wasof et al., 2013), or large genetic divergence among lineages due to local adaptation (Cooper *et al.*, 2011; Meynard *et al.*, 2017). When distinct genetic groups respond differently to environmental variables, modelling those genetic

units independently could theoretically provide more reliable predictions of the species response as a whole (Pearman et al., 2010; Peterson et al., 2018; Smith et al., 2018). Based on those premises, a few studies have arrived at better predictions when modelling lineages independently (Palma *et al.*, 2017; Lecocq *et al.*, 2019; Chardon *et al.*, 2020). For two rodent species in the Andes, Palma et al. (2017) used SDMs to model each species individually as well as individual lineages separately for each species. They showed that stacking all individual lineage models performed either similar or better, depending on the study species, than using the traditional SDM approach of modelling the species as a whole. Lecocq et al. (2019) also investigated the performance of individual lineage models compared to a whole species model for three bumblebee species in the West-Palaearctic. The authors concluded that assembling individual lineage-based models does not increase model performance, whatever the metric used. Overall and as far as we know, all studies comparing a combined set of lineage-specific SDMs, hereafter genetically-informed SDMs, against traditional SDMs do not provide a crystal clear answer regarding the performance of genetically-informed models over traditional whole species models.

Here, we aim at investigating further this question of whether or not informing SDMs with genetic information (phylogeography or population genetic structure) could improve model performance over traditional approaches (i.e., modelling all occurrences as one homogeneous single unit). We constructed both genetically-informed and traditional SDMs for two widespread European species for which there is a clear genetic structure published in the scientific literature: common beech *Fagus sylvatica* (Magri et al., 2006) and the castor bean tick *Ixodes ricinus* (Poli et al., 2020). For *F. sylvatica*, we could also rely on accurate pollen and macrofossil distribution data from the Mid-Holocene period to test the predictive performances of both modelling approaches during a time period that was warmer than the present-day climate for Europe and thus a good candidate period for anticipating the potential future distribution of *F. sylvatica*. We hypothesized that the overall species distribution predicted from an assemblage of genetically-informed models would better predict the entire

potential distribution of the focal species than a traditional SDM approach and perform better at identifying past occurrences during the Mid-Holocene warm period for *F. sylvatica*.

Material and Methods

All the data preparation steps and statistical analyses described in the following subsections were conducted in R (R Core Team, 2019), using the following suite of packages: raster (Hijmans, 2020); rgdal (Bivand et al., 2020); sp (Pabesma & Bivand, 2005); biomod2 (Thuiller et al., 2020); usdm (Naimi et al., 2014); ade4 (Dray & Dufour, 2007); and viridis (Garnier, 2018). All maps and spatial information displayed in the figures were projected using the Lambert azimuthal equal-area (LAEA) Europe (EPSG:3035) projection system and all raster layers (e.g. bioclimatic variables and model predictions) were set to a spatial resolution of 5 arc-minutes (about 9.3 by 9.3 km at the equator) as a compromise between spatial accuracy and computational power.

Study species and genetic data

Here, we focused on *Fagus sylvatica* L. (1753) and *Ixodes ricinus* L. (1758) which are two widely distributed species in Europe. Both studied species have detailed published records on phylogeography or genetic structure, thus allowing us to build SDMs with genetic information at the intra-species level. The phylogeography of *F. sylvatica* was investigated by Magri et al. (2006), where nine lineages were identified based on isozyme data. The population genetic structure of *I. ricinus* was recently investigated by Poli et al. (2020), where two well distinct Eurasian genetic clusters were identified (one Northern and the other Southern of Eurasia). Besides, thanks to pollen and macrofossil data, *F. sylvatica* is a good candidate species to validate our models against empirical data of its past distribution during the Mid-Holocene period (~6000 ybp).

Distribution data for *Fagus sylvatica*

Current occurrence data for *F. sylvatica* were extracted from both Magri et al. (2006) and the EU-Forest dataset (Mauri *et al.* 2017). Spatial locations of occurrence data from Magri et al. (2006) are already assigned to particular lineages (n = 9) based on isozyme data. Some of the lineages identified in Magri et al. (2006) rely on small sample sizes. More specifically, lineages #3 and #6 had six and five occurrences, respectively, and thus were excluded from further analysis. Lineages #2, #4, and #5 were grouped as they form a monophyletic group, as were lineages #7, #8, and #9 belonging to another monophyletic group. These two monophyletic groups (lineage #245 and lineage #789), as well as lineage #1, are hereafter referred to as 'lineages' for simplicity. To increase the number of occurrence data for each of these three lineages, we first extracted all spatial locations from the EU-Forest dataset where *F. sylvatica* is occurring in Europe (n = 35,862). To assign each of these occurrences from the EU-Forest dataset to a particular lineage, we relied on the spatial proximity between occurrence data in the EU-Forest dataset and occurrence data in Magri et al. (2006) with an identified lineage membership (#1, #245, #789) (Fig. S1). First, we build a circular buffer of 100 km around the spatial location of each lineage-labelled occurrence in Magri et al. (2006). The buffer size was optimized to maximize the number of occurrences for each lineage while minimizing overlap among neighbouring lineages. Then, all overlapping circular buffers of a particular lineage were dissolved. Whenever the spatial location of an EU-Forest occurrence for *F. sylvatica* fell exclusively inside the buffer area of a given lineage, it was assigned to that particular lineage (Fig. S1). Spatial locations of occurrence data from the EU-Forest dataset falling within several overlapping lineages' buffer zone or outside of all three lineage buffers were excluded (Fig. S1). After resampling, the final set of present-day occurrences for *F. sylvatica* reached 1,140, including 609, 281, and 239 occurrences for lineages #1, #245, and #789, respectively (Fig. S2a).

Because the EU-Forest dataset is a quasi-exhaustive survey of all European tree species co-occurring within a given forest plot, the fact that the name of a given tree species is not recorded within the focal

plot is interpreted as an absence data. Moreover, *F. sylvatica* is a non-cryptic and easily identifiable tree species. Hence, the inferred absences from locations where the species was not recorded in the EU-Forest dataset can be considered as true absences, and not pseudo-absences or background data. From this absence dataset (n = 214,716), we randomly selected as many absences as the total number of occurrences we used for *F. sylvatica* (n = 1,140), such that the prevalence of the total set of presence-absence data is 0.5 as suggested by Barbet-Massin et al. (2012).

Historical data on *F. sylvatica* distribution during the Mid-Holocene period was extracted from the European pollen database (EPD; http://www.europeanpollendatabase.net/) for the period between 5000 and 7000 ybp. Pollen data from the EPD was combined with that available in Magri et al. (2006). A great variation in the abundance of pollen records dated between 5000 and 7000 ybp was observed between sites recorded in the EPD (from 1 to 709 records per site per time). Since the species is wind-pollinated, the presence of pollen in a site at a certain time may not necessarily reflect past occupancy of that particular site. Yet, a greater amount or abundance of pollen data within a site increases the chances that this site, or a neighbouring location, was effectively occupied by *F. sylvatica*. To account for this possible artefact, we only retained sites where a given threshold of pollen records of *F. sylvatica* dated at the same period (between 5000 and 7000 ybp) was observed. Two thresholds of pollen record distribution were considered: the first quartile (minimum of 4 pollen records) and the medium (minimum of 37 pollen records). After this filtering, a total of 246 grid cells (n = 145 when using the median instead of the first quartile) contained a sufficient amount of pollen records among sites dated between 5000 and 7000 ybp to consider that *F. sylvatica* was present during the Mid-Holocene period in that particular cell. Macrofossil data (e.g. charcoal) from the same period were also extracted from Magri et al. (2006) as well as from Lafontaine et al. (2014). A total of 19 grid cells contained macrofossils of *F. sylvatica* from the Mid-Holocene period across the study area.

Distribution data for *Ixodes ricinus*

Current occurrence data for *I. ricinus* were extracted from Poli et al. (2020), GBIF (GBIF.org, 2020), and VectorMap (http://vectormap.si.edu/). Similar to Magri et al. (2006) for *F. sylvatica*, occurrence data used in Poli et al. (2020) (n = 497) were already assigned to particular genetic clusters: the northern vs. southern Eurasian clusters. Occurrence data extracted from GBIF (n = 2,929) and VectorMap (n = 880) were assigned to one of the two Eurasian clusters according to the geographic locations of occurrences. First, we build an interpolation map of probabilities of assignment to a given genetic cluster based on Poli et al. (2020) (Fig. S3). Occurrence data falling within a high probability zone for one cluster (>50%) were assigned to that cluster. A total of 2,171 occurrences, including 884 and 1287 occurrences for the southern and northern cluster, respectively, were kept for further analysis (Fig. S2b). An equal number of pseudo-absences or background data were randomly selected within a convex-hull around all occurrences.

Bioclimatic variables

To build both traditional and genetically-informed SDMs, we relied on several bioclimatic variables as predictor variables. We first downloaded contemporary bioclimatic variables at 5 arc-minute resolution representative of long-term conditions during 1970-2000 from the WorldClim 2 database (Fick & Hijmans, 2017 - https://www.worldclim.org/data/index.html). For each of the two studied species, bioclimatic variables used as predictor variables in our models were selected based on knowledge from the scientific literature (Durrant et al., 2016; Svenning et al., 2011; Alkishe et al., 2017). We excluded collinear variables based on the results of the variance inflation factor (VIF), using the *vifcor* function from the R package usdm (Naimi et al., 2014). Bioclimatic variables kept as predictors in our SDMs are shown in **Table 4-1**. The same set of bioclimatic variables used to model *F. sylvatica* was downloaded for the Mid-Holocene period (~6000 ybp) from the WorldClim 1.4 database (https://www.worldclim.org/data/v1.4/paleo1.4.html - Hijmans *et al.*, 2005). The bioclimatic layers

from the nine global circulation models (GCMs) of the Mid-Holocene period available in WorldClim 1.4 were used to predict the probability of occurrence of *F. sylvatica* during the Mid-Holocene period.

Model calibration during present-day climate

For each species and each lineage (*F. sylvatica*) or genetic cluster (*I. ricinus*), we build a presence-absence dataset composed of all current occurrences assigned to that particular lineage (or genetic cluster) plus the same number of absences (*F. sylvatica*) or pseudo-absences (*I. ricinus*) randomly assigned to each lineage (or genetic cluster) from the total pool of absences (or pseudo-absences) described above. Two 'whole-species' or 'total' datasets were also created, one for each species, corresponding to the combination of occurrences from all intra-species level entities belonging to a given species and all the available absences (*F. sylvatica*) or pseudo-absences (*I. ricinus*). Those last two datasets were used to build traditional SDMs, without incorporating any intra-species level information, being either lineages or genetic clusters. In total, four and three presence-absence datasets, with a prevalence of 0.5 each time, were constructed for *F. sylvatica* (Lineage #1, Lineage #245, Lineage #789, and whole-species dataset) and *I. ricinus* (Southern cluster, Northern cluster, and whole-species dataset), respectively. For each of those seven datasets, 30% of presence-absence data were set aside and kept for the final external validation step once ensemble models were generated (see next section on model validation and comparison during present-day climate), while the remaining 70% of presence-absence data were used for model calibration. Since absences (*F. sylvatica*) and pseudo-absences (*I. ricinus*) of each lineage and genetic cluster, respectively, were randomly sampled/assigned from the total pool of absences (*F. sylvatica*) or pseudo-absences (*I. ricinus*), we repeated this process 20 times, building a total of 140 datasets, all with a prevalence of 0.5, but different combinations of absences/pseudo-absences, each time keeping 70% of data for model calibration and 30% for the final external validation step.

For each of the above-mentioned 140 calibration datasets, we related the presence-absence data to the set of selected bioclimatic variables (predictors). Note that the set of predictors used at the

species level was identical at the intra-species level. We used four algorithms implemented in the biomod2 R package (Thuiler et al., 2020) to model the 'current' (1970-2000) potential distribution of each lineage, genetic cluster, or whole-species level: generalized linear models (GLMs); generalized boosted regression models (GBMs); generalized additive models (GAMs); and random forests (RFs). Each time, we used the default parameters. For each of the 140 calibration dataset and each of the four algorithms, we ran 10 repetitions by further setting aside, randomly, 70% of the presence-absence data for algorithm-specific calibration and the remaining 30% for algorithm-specific validation, which makes 1,400 different calibration-validation datasets per algorithm and thus 5,600 models. To assess the models' performance, we computed the true skill statistic (TSS), and all predicted probabilities of occurrence greater than 0.5 were transformed into presences while probability values lower or equal to 0.5 were transformed into absences. A fixed value for the threshold was necessary for comparing different runs. A threshold of 0.5 was chosen after previous modelling with subsets of the presence-absence dataset that suggested this to be the value that maximized The TSS values. For each of the 140 original calibration datasets, the forty models we calibrated (4 algorithms by 10 repetitions) were subsequently assembled in an ensemble model by weighting coefficient estimates based on TSS values. Mind that the 140 original calibration datasets we used here included 80 (4 × 20) and 60 (3 × 20) original calibration datasets for *F. sylvatica* and *I. ricinus*, respectively.

Table 4-1. Bioclimatic Variables from WorldClim (Hijmans et al., 2005) used for building SDMs for each species.

Bioclimatic Variable	Description	*F. sylvatica*	*I. ricinus*
Bio1	Annual Mean Temperature	X	X
Bio2	Mean Diurnal Range (Mean of monthly (max temp - min temp))		X
Bio3	Isothermality (BIO2/BIO7) (* 100)		X
Bio4	Temperature Seasonality (standard deviation *100)	X	X
Bio12	Annual Precipitation	X	X
Bio14	Precipitation of Driest Month	X	
Bio15	Precipitation Seasonality (Coefficient of Variation)	X	
Bio18	Precipitation of Warmest Quarter	X	
Bio19	Precipitation of Coldest Quarter	X	

Model validation and comparison during present-day climate

To construct a genetically-informed SDM for a given study species, we combined predictions obtained from all ensemble models across all lineages or genetic groups of the focal species into an ensemble probability map. More specifically, we overlaid all three (two) predicted probability of occurrence of the three (two) lineages (genetic clusters) of *F. sylvatica* (*I. ricinus*), and kept only the maximum value, as proposed by Lecocq et al. (2019). This layer of lineage-specific maximum probability of occurrences representing the main outcome of a genetically-informed SDM was then compared with the

corresponding layer of predicted probability of occurrences of the traditional or control SDM based on the whole-species dataset without distinguishing between lineages or genetic groups.

Table 4-2. Discrimination metrics used to access model performance. TP: True positives. FN: False negatives. TN: True negatives. FP: False positives. p_o: proportion of agreement. p_e: expected proportion of agreement.

Metric	Definition	References
Sensitivity	TP/(TP+FN)	Fielding and Bell (1997)
Specificity	TN/(TN+FP)	Fielding and Bell (1997)
True Skill Statistics (TSS)	Sensitivity+Specificity-1	Allouche et al. (2006)
AUC	Area Under the Receiver Operating Characteristic curve	Lobo et al. (2007)
Sørensen's similarity index	2TP/(FN+2TP+FP)	Leroy et al. (2018)
Overprediction Rate (OPR)	FP/(TP+FP)	Marcia Barbosa et al. (2013)

For a fair comparison of model performances between genetically-informed and traditional SDMs, we generated six different but complementary discrimination metrics (**Table 4-2**). There is a vast literature available about the limits, applications, and information provided by these different discrimination metrics used to assess SDMs' performances (Fielding & Bell, 1997; Liu *et al.*, 2009; Liu *et al.*, 2011; Leroy *et al.*, 2018; Shabani *et al.*, 2018). Transformations of the predicted probabilities of occurrence into presence-absence for both the genetically-informed and traditional SDMs were based on a threshold of 0.5. We used the set of observed presence-absence data from the 140 validation datasets that we set aside earlier to compute confusion matrices. For each of the 20 random iterations used to generate the 140 validation datasets, we combined the three lineages of *F. sylvatica* and the two genetic clusters of *I. ricinus* into a single dataset to compute all the metrics of model performance for the genetically-informed and traditional SDM approach. This led to a total of 20 values for each metric and each type of SDM (traditional vs. genetically-informed SDM). We ran a Mann-Whitney rank

test to assess the significance of the difference between the two SDM approaches across the 20 repetitions.

Model evaluation and comparison during the Mid-Holocene period

Based on the ensemble models that we calibrated for *F. sylvatica* during the present-day climate, we hindcasted the potential distribution of *F. sylvatica* during the Mid-Holocene period (about 6,000 ybp) using all nine GCMs for each of the three lineages separately as well as at the species level. This resulted in a total of 36 raster layers, including 27 (9 × 3) layers at the lineage level and nine layers at the species level. For each grid cell and each lineage separately, as well as at the species level, we then averaged probabilities of occurrences (i.e. habitat suitability values) across the nine GCMs. Similar to the maximum lineage projection described above, we kept only the maximum probability of occurrence among the three studied lineages of *F. sylvatica* as the outcome of the genetically-informed SDM. To evaluate and compare the model performances to hindcast the species' past distribution between traditional and genetically-informed SDMs, we extracted all the predicted probability of occurrences across all grid cells that contain pollen records or macrofossils. We then applied two complementary approaches to test for potential differences in the potential distribution of *F. sylvatica* during the Mid-Holocene period between the two studied SDM approaches (traditional vs. genetically-informed). First, a Mann-Whitney rank test was used to compare the distribution of probability (habitat suitability) values of each cell where pollen or macrofossil records occurred (presence-only test). Then, as a second and more quantitative approach to account for the abundance of pollen records in each grid cell, we extracted the total (sum), mean, median, and maximum abundance of pollen records per grid cell. For each of these four summary statistics, we regressed pollen abundance (log-transformed) against the probability of occurrence, separately for each of the two studied SDM approaches, using a linear modelling approach: $Y = \beta_0 + \beta_1 X$, where Y is the log-transformed pollen abundance, β_0 is the intercept, β_1 in the slope of the regression, and X is the predicted species probability of occurrence. We tested for statistical differences in the slope estimate between the genetically-informed and

traditional SDM approach by using a randomization procedure. Probabilities of occurrence in each of the grid cells where pollen records occurred were randomly assigned to either the traditional or genetically-informed SDM approach. Next, for each of the two randomly assigned SDM approaches, we applied the same linear model as the one mentioned above for empirical data. Last, the randomized slope coefficient estimate β_1 from the traditional SDM approach was subtracted from the genetically-informed approach and this randomized difference stored. This process was repeated a thousand times, creating a simulated distribution of the randomized differences in the slope coefficient estimate between both SDM approaches. The observed difference in the slope coefficient estimate between the two SDM approaches was then compared to the random distribution of 1,000 randomized difference in the slope coefficient estimate to compute a non-parametric *p*-value. We decide for the regression approach descried above to avoid adding more uncertainty to the analysis as would be the case of a transformation of pollen occurrences to a presence-absence data since even though sites with a great abundance in pollen of *F. sylvatic* represent most probably real presences, the absence of pollen records does suggest a real absence of the species.

Results

Model comparison and validation during the present-day climate (*Fagus sylvatica* & *Ixodes ricinus*)

For both species (*F. sylvatica* & *I. ricinus*) and both modelling approaches (traditional vs. genetically-informed SDMs), predictive performances under the present-day climate ranked from good to excellent (**Figure 4.1** and Table S1). We found high AUC values, ranging from 0.90 to 0.93 for *F. sylvatica* and from 0.80 to 0.84 for *I. ricinus*. Similarly, TSS values were relatively high and ranged between 0.60 and 0.71 for *F. sylvatica* and between 0.60 and 0.68. for *I. ricinus*. In all individual models, the RFs algorithm outperformed the other three algorithms, followed by GAMs and GBMs, while GLMs were always the less performant modelling algorithm (Fig. S4). In general, we found similar distribution patterns between both modelling approaches and for both studied species, matching with the known

present-day distribution of both species (**Figure 4.2**). Noteworthy, probability values tended to be higher for genetically-informed than for traditional SDMs.

For *I. ricinus*, neither AUC nor TSS values differed between the two modelling approaches, whereas for *F. sylvatica*, TSS values were higher when using the traditional over the genetically-informed SDM approach (**Figure 4.1**). We found similar Sørensen values between the two modelling approaches for both species. Noteworthy, for both species, the genetically-informed SDM approach reached systematically higher sensitivity and OPR values, but systematically lower specificity values than the traditional SDM approach.

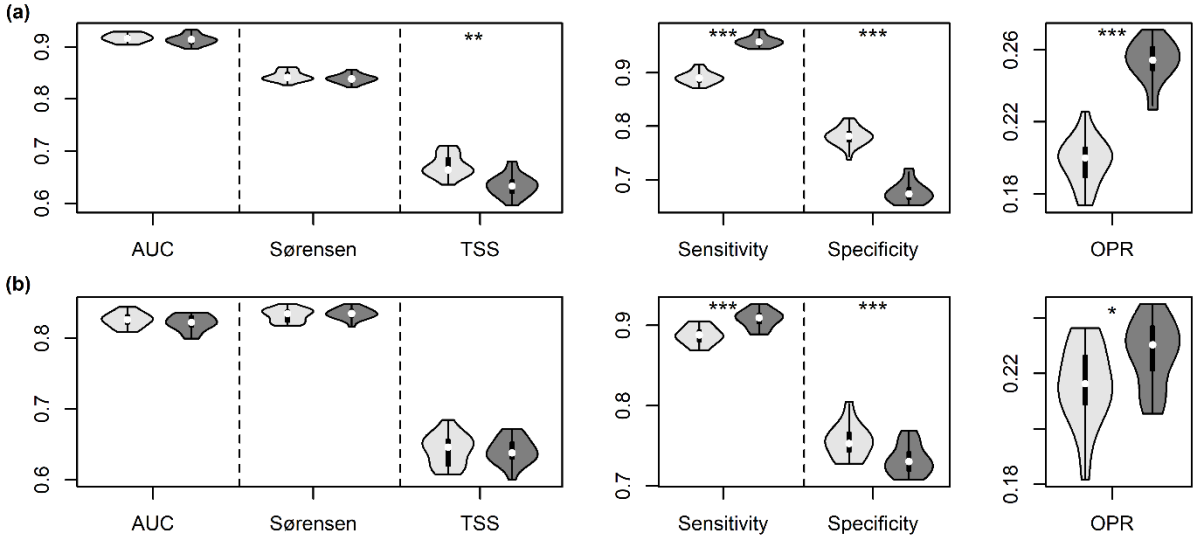


Figure 4.1. Comparison of the resulting values of the six discrimination metrics used across 20 repetitions for (a) *Fagus sylvatica* and (b) *Ixodes ricinus*. Light grey: traditional species distribution model (SDM) approach. Dark grey: genetically-informed SDM approach. Significances are indicated by asterisks (***: $p < 0.001$; **: $0.001 < p < 0.01$; *: $0.01 < p < 0.05$).

Model evaluation and comparison during the Mid-Holocene period for *Fagus sylvatica*

Predicted probabilities of occurrence during the Mid-Holocene period at locations where pollen data occurred were higher for the genetically-informed SDM approach than the traditional one for both the first quartile and the median threshold (Mann-Whitney, $p = 0.010$ and $p = 0.004$ respectively, **Figure 4.3a** and S5). Although predicted probabilities of occurrence from the genetically-informed SDM approach tended to be higher than the traditional one at locations where macrofossils (i.e. charcoal)

were found, the average difference was not significant (Mann-Whitney, p = 0.359, **Figure 4.3b**). Irrespective of the summary statistic used (mean, median, sum, or maximum abundance of pollen per grid cell), we found a stronger relationship between pollen abundance (log-transformed) and the probability of occurrence of *F. sylvatica* during the Mid-Holocene period when using the genetically-informed SDM approach (R^2 ranging from 0.36 to 0.42) than when using the traditional one (R^2 ranging from 0.21 to 0.26) (**Figure 4.4** and Fig. S6). The slope coefficient estimates for the genetically-informed SDM approach were always higher ($p < 0.05$) than for the traditional SDM approach (Fig. S7 and Table S2). Note that increasing the threshold to the median value (37 or more pollen records) instead of the first quartile did not change the main findings (Figs. S8-S9 and table S2). Noteworthy, the genetically-informed SDM approach showed a much better match with pollen data in the Balkans, Italy, around the Dead Sea, and the northern Iberian regions than the traditional SDM approach (**Figure 4.5** and **Figure 4.6**).

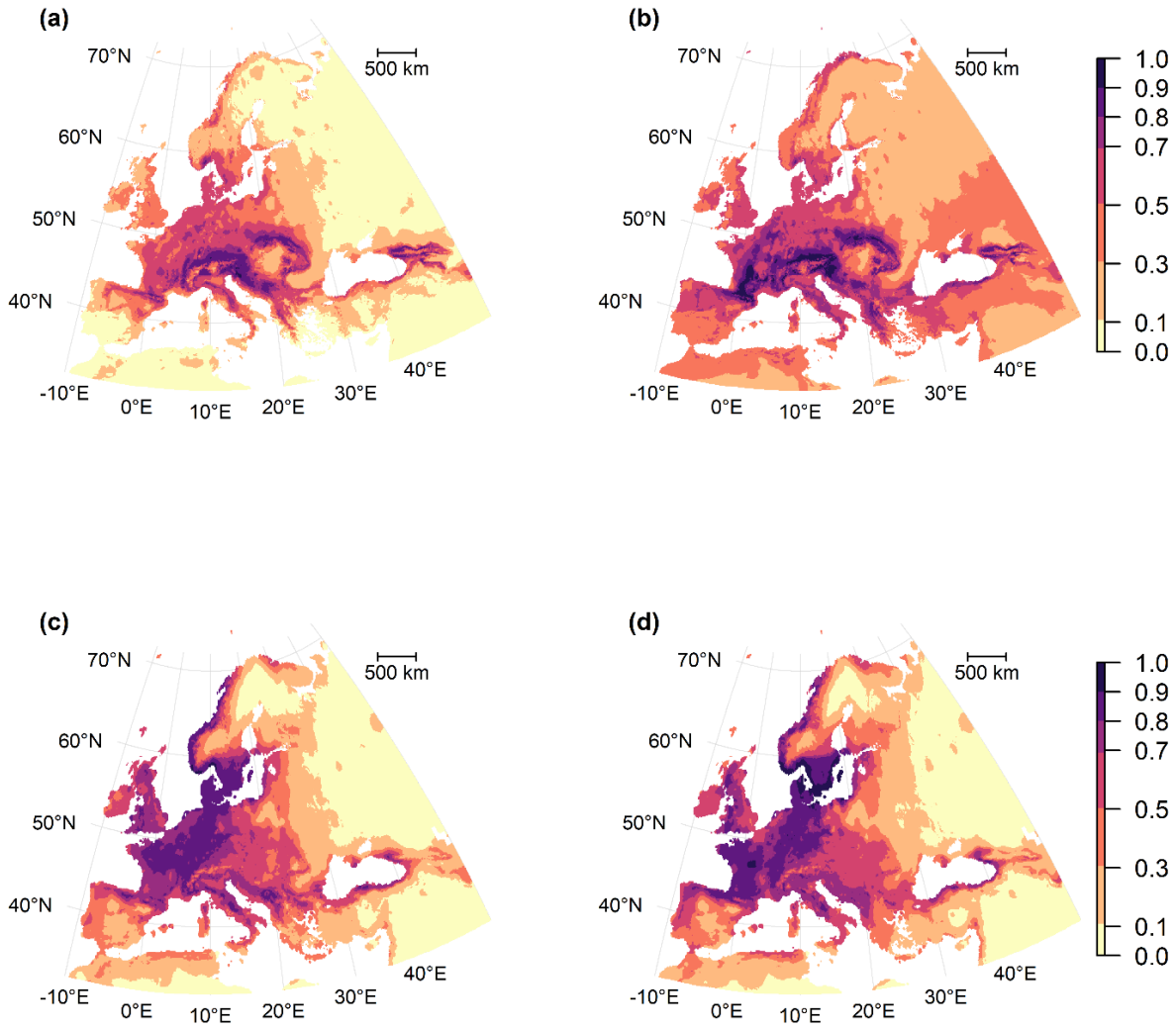


Figure 4.2. Predictions of the probability of presence or habitat suitability for *Fagus sylatica* (a, b) and *Ixodes ricinus* (c, d) during present day climate (1970-2000) using both the traditional species distribution model (SDM) approach (left) and the genetically-informed SDM approach (right).

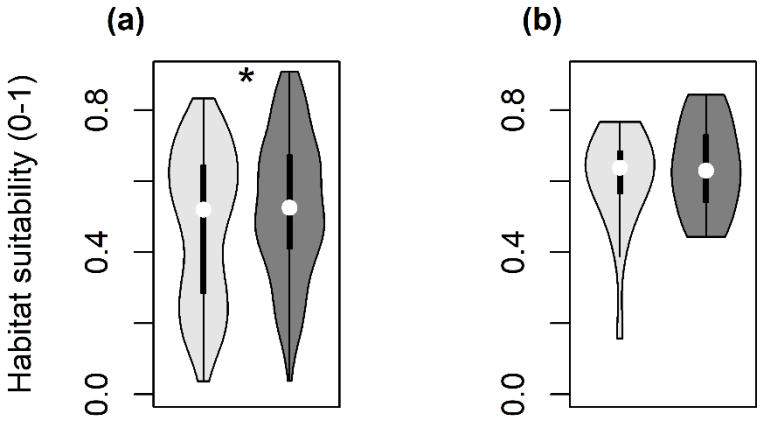


Figure 4.3. Distribution of the probabilities of presence of *Fagus sylvatica* during the Mid-Holocene period at spatial locations where fossil records of *F. sylvatica* from the Mid Holocene period have been found for pollen (first quartile threshold, n>=4 pollen records per site and time) (a) and macrofossil (charcoal) (b) records. Light grey: traditional species distribution model (SDM) approach. Dark grey: genetically-informed SDM approach. Stars display the significance level based on a Mann-Whitney test of difference between the two SDM approaches (*, p = 0.010).

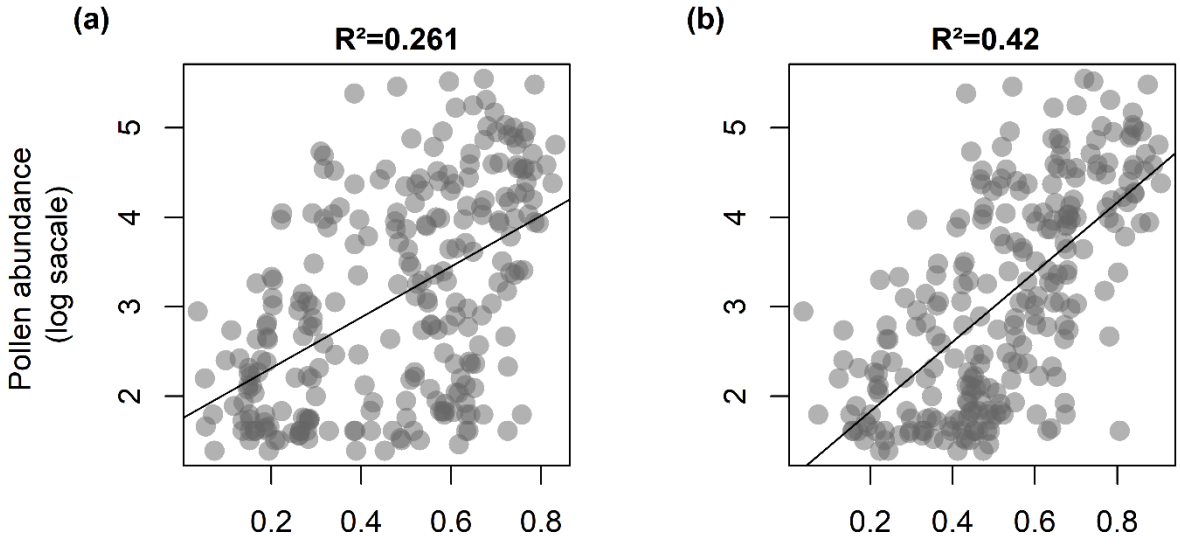


Figure 4.4. Linear regression relating the mean pollen abundance (log-transformed) with a threshold of 4 pollen records (first quartile) across sites co-occurring in the same grid cell as a function of the probability of occurrence or habitat suitability according to the traditional species distribution model (SDM) approach (a) and the genetically-informed SDM approach (b) for *F. sylvatica*. For results based on summary statistics other than the mean per grid cell (e.g. median or maximum abundance per grid cell), please refer to S4. Note also that these results are based on the first quartile across all pollen records as a threshold to exclude locations with very limited pollen abundance. For results using a more restrictive threshold (median value), please refer to S6.

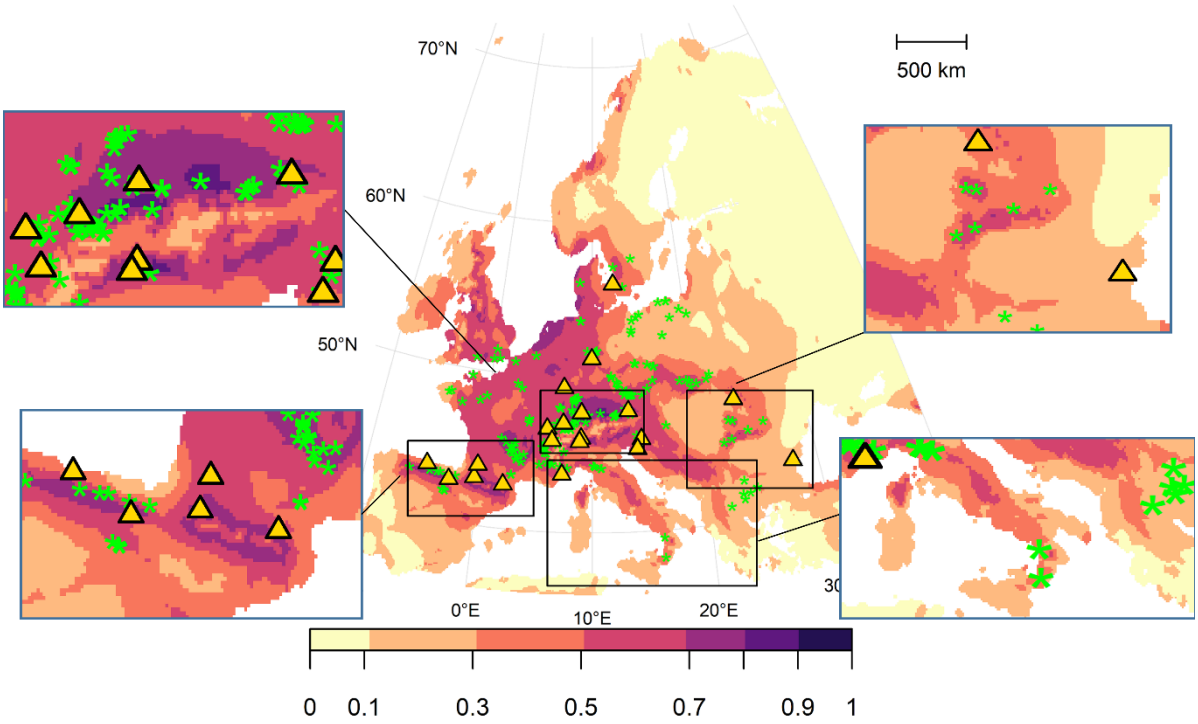


Figure 4.5. Fossil records and probability distributions during the Mid Holocene period for *Fagus sylvatica*. Probability of occurrence or habitat suitability values are based on the traditional species distribution model (SDM). Yellow triangles: macrofossil records. Green asterisks: pollen records.

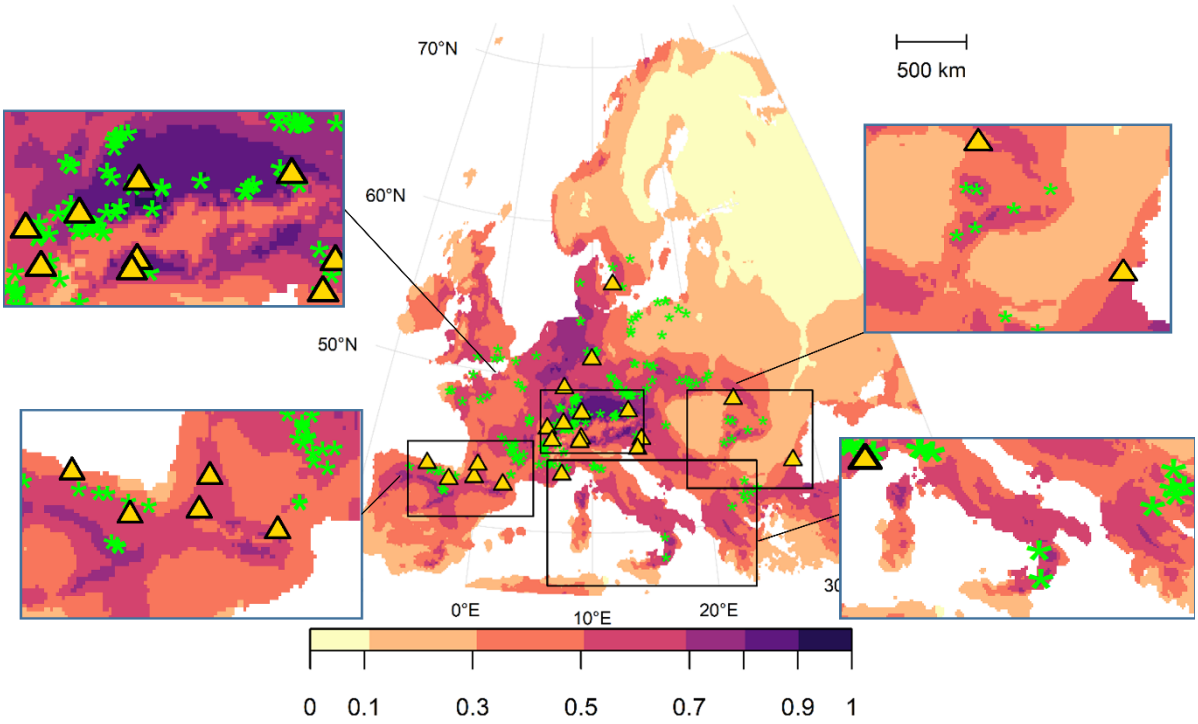


Figure 4.6. Fossil records and probability distributions during the Mid Holocene period for *Fagus sylvatica*. Probability values are based on the genetically-informed species distribution model (SDM). Yellow triangles: macrofossil records. Green asterisks: pollen records. The genetically-informed SDM approach was capable to assign higher (> 0.5) probability values than the traditional SDM approach (**Figure 4.5**), most notably at the southern edge of the species distribution (see zooming windows), coinciding with locations where fossil records occur.

Discussion

Our findings suggest that genetically-informed SDMs tend to increase the probability to detect potentially suitable sites and thus potential cryptic refugia for the focal species at locations where traditional SDMs fail to do so. Albeit the well-known and widely used metrics of SDMs' performances (i.e. AUC) did not differ between the traditional and genetically-informed SDM approaches, ranging from good to excellent in both cases, our genetically-informed SDMs were able to systematically identify suitable habitats otherwise neglected by the traditional SDM approach, as shown by the higher sensitivity values we found for both *Fagus sylvatica* and *Ixodes ricinus*. However, the genetically-

informed SDM approach also tended to overestimate the rate of predicted occurrences at locations where in fact absences (for *F. sylvatica*) or pseudo-absences (for *I. ricinus*) have been recorded, as translated by the higher OPR values. Similarly, traditional SDMs performed better at identifying absences or pseudo-absences, as shown by the higher specificity values. These discrepancies observed between the two approaches when using metrics focusing more specifically on the models' ability to correctly predict either the occurrences (e.g. sensitivity and OPR) or the absences (e.g. specificity) almost disappeared when using metrics combining the models' ability to correctly predict occurrences and absences simultaneously (e.g. AUC, TSS, and Sørensen). Although the tendency for a model to underestimate absences (low specificity values) or overestimate occurrences (high OPR values) may seem like bad news at first sight, one needs to interpret these performance metrics in light of the quality, reliability, or trustworthiness of the data used as "field observations" to validate model predictions (see Box 1 in De Kort *et al.*, 2020). Indeed, although it may seem obvious and straightforward to trust a record of species occurrence from field observations, it is not as obvious with absence records from field observations, and even less so for pseudo-absences or background data (Lobo *et al.*, 2010; De Kort *et al.*, 2020). There are several important reasons for not necessarily trusting an absence record, even from field observations (Lobo et al., 2010). For instance, even an expert taxonomist may simply overlook the focal species that is occurring at a given location, either because it is an inconspicuous species or because it is not the right phenological window to observe the species. Those are typically referred to as methodological absences. Finally, even if an absence record observed in the field is real and not a methodological artefact, it is difficult to attribute this absence to environmental limitations only (environmental absences), as other contingencies may explain it in the field (contingent absences *sensu* Lobo et al., 2010), such as biotic interactions, historical factors, habitat fragmentation or simply dispersal limitations (Hattab et al. 2017). Hence, one needs to keep in mind the high uncertainty around absence records (Lobo et al., 2010) when interpreting model performance metrics partly relying on "observed" absences. Confirmed absences from repeated field observations are required to generate synthetic metrics (e.g. TSS) of model performance that are

trustworthy and unlikely to underestimate the model's ability to detect a potentially suitable site. For that reason, one cannot rely on one single metric of model performance to compare SDM outputs. Instead, it is of utmost importance to multiply the complementary metrics (e.g. sensitivity and specificity) to get a better understanding of the weaknesses and strengths of a given model.

Independent model evaluation for *Fagus sylvatica*

Focusing on *F. sylvatica* for which we have extensive knowledge of its past distribution thanks to pollen and macrofossil data, we were able to evaluate the ability of our genetically-informed SDM approach to accurately predict the potential occurrence of *F. sylvatica* during the Mid-Holocene period and compare it to the control situation of using traditional SDMs. The genetically-informed SDM approach showed better performances than traditional SDMs one to identify suitable habitat during the Mid Holocene period. This was demonstrated by the higher habitat suitability values extracted from the genetically-informed SDM at locations where pollen records were found as well as by the higher proportion of variance in pollen abundance that is explained by the habitat suitability predicted from the genetically-informed SDM. In fact, the magnitude of the positive relationship between pollen abundance and habitat suitability was much stronger when using the genetically-informed SDM approach, suggesting that this approach performs better at predicting the potential distribution of *F. sylvatica* during the Mid-Holocene period.

Both the genetically-informed and traditional SDM approach assigned higher probabilities of occurrence in the core of the predicted species past distribution. Nonetheless, the genetically-informed SDM approach was able to correctly identify suitable habitats at the southern limits of *F. Sylvatica* past distribution, around the Dead Sea, the Balkans, the Italian Peninsula, the northern part of the Iberian Peninsula, and south-western France. Most of those regions were not identified as highly suitable by the traditional SDM approach. Unfortunately, not much macrofossil data was available to confirm the overall tendency that genetically-informed SDMs are better at predicting the potential distribution of the focal species than traditional SDMs. Indeed, the non-significant signal in the

difference between the two modelling approaches as measured by macrofossil records is most likely due to the small sample size (n = 19). In any case, our findings suggest that combining lineage-specific models could help identify cryptic refugia (i.e., zones with high probabilities of occurrence) otherwise missed by traditional SDMs.

Implications for forecasting species redistribution under future climate change

According to our findings, the genetically-informed SDM approach has the potential to increase the reliability of future scenarios of biodiversity redistribution by increasing the model's abilities to detect potentially suitable areas, which also means a potential for reducing omission errors (i.e. false negative: predicting a future absence where it will be a presence). However, similar to the principle of communicating vessels, while reducing omission errors, the genetically-informed SDM approach also has the potential to increase commission errors (i.e. false positive: predicting a future occurrence where it will be an absence) in comparison with a more traditional SDM approach. Reducing omission errors at the expense of commission errors has the advantage to limit the risk of missing future suitable areas. For instance, for predicting the distribution of rare and endemic species with limited dispersal abilities, it has been demonstrated that reducing omission errors is of greater importance than reducing commission errors (Pearson et al., 2007; Liu et al., 2016). In those cases, SDMs' future forecasts could benefit from incorporating the genetic structure of those species.

Problems in the SDMs' performances to predict species' range margins have already been pointed out in the scientific literature (Braunisch *et al.*, 2008; Vale *et al.*, 2014). For both *F. sylvatica* and *I. ricinus*, traditional SDMs tended to underestimate suitable regions at the range margins in comparison with genetically-informed SDMs. This has important implications for the forecasts of future biodiversity redistribution given population dynamics at range margins.

While some pioneer studies have already suggested that incorporating genetic information into SDMs improve model performance (Palma et al., 2017; Chardon et al., 2020), others have suggested that it does not improve model accuracy (Lecocq et al., 2019), thus reopening the debate on whether

incorporating genetic information into SDMs can improve our understanding of species distribution. Noteworthy, all these studies mainly focused on visual inspections from predicted maps to make comparisons of model performances or used general metrics of model performances (e.g. Palma et al., 2017) systematically relying on presence-only data (Chardon et al., 2020) to validate their models and draw their conclusions, which may limit our abilities to evaluate the performance of genetically-informed SDMs. As far as we are aware, no study before ours has specifically tested the ability of genetically-informed SDMs to better explain the abundance of pollen found in the fossil records, providing truly independent evidence that genetically-informed SDMs outperform traditional SDMs in their abilities to detect potential refugia. Here, showed that genetically-informed SDMs can outperform traditional SDMs in identifying potentially suitable habitat of a given species.

Limits of the genetically-informed SDM approach

Aside from the obvious limitation related to the availability of genetic data at the infra-species level, it is important to note that the genetically-informed SDM approach we propose here is probably not applicable across all species. Both our studied species have a geographically explicit genetic structure or phylogeography. In such cases, modelling individual genetic units serves as a method for dismembering the dataset in geographical subunits coherent with the species life-history, ecology, and possibly local adaptation. In cases where the target species does not show a spatially explicit genetic structure, modelling and assembling individual genetic groups may not necessarily have an advantage over traditional SDM methods.

Data availability statement

Pollen records extracted from the European Pollen Database can be accessed are from Figshare with the following link: https://figshare.com/s/a4560da2e568e87cb6ac. Upon acceptance, a public DOI will be reserved and replace the above link.

References

Alkishe, A.A., Peterson, A.T. & Samy, A.M. (2017) Climate change influences on the potential geographic distribution of the disease vector tick *Ixodes ricinus*. *PLOS ONE*, **12**, e0189092.

Allouche, O., Tsoar, A. & Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS): Assessing the accuracy of distribution models. *Journal of Applied Ecology*, **43**, 1223–1232.

Araújo, M.B., Anderson, R.P., Márcia Barbosa, A., Beale, C.M., Dormann, C.F., Early, R., Garcia, …, Rahbek, C. (2019) Standards for distribution models in biodiversity assessments. *Science Advances*, **5**, eaat4858.

Bivand, R., Keitt, T. & Rowlingson, B. (2020). rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.5-10. https://CRAN.R-project.org/package=rgdal

Braunisch, V., Bollmann, K., Graf, R.F. & Hirzel, A.H. (2008) Living on the edge—Modelling habitat suitability for species at the edge of their fundamental niche. *Ecological Modelling*, **214**, 153–167.

Chardon, N.I., Pironon, S., Peterson, M.L. & Doak, D.F. (2020) Incorporating intraspecific variation into species distribution models improves distribution predictions, but cannot predict species traits for a wide-spread plant species. *Ecography*, **43**, 60–74.

Cooper, N., Freckleton, R.P. & Jetz, W. (2011) Phylogenetic conservatism of environmental niches in mammals. *Proceedings of the Royal Society B: Biological Sciences*, **278**, 2384–2391.

Cooper, N., Jetz, W. & Freckleton, R.P. (2010) Phylogenetic comparative approaches for studying niche conservatism: Comparative approaches for niche conservatism. *Journal of Evolutionary Biology*, **23**, 2529–2539.

De Kort, H., Baguette, M., Lenoir, J. & Stevens, V.M. (2020) Toward reliable habitat suitability and accessibility models in an era of multiple environmental stressors. *Ecology and Evolution*.

Dray, S. & Dufour, A. (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, **22**(4), 1-20. doi: 10.18637/jss.v022.i04

Houston Durrant, T., de Rigo, D., Caudullo, G. (2016) *Fagus sylvatica* and other beeches in Europe: distribution, habitat, usage and threats. In: San-Miguel-Ayanz, J., de Rigo, D., Caudullo, G., Houston Durrant, T., Mauri, A. (Eds.), European Atlas of Forest Tree Species. Publication Office of the European Union, Luxembourg

Fick, S.E. & Hijmans, R.J. (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, **37**, 4302–4315.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Grimmett, L., Whitsed, R. & Horta, A. (2020) Presence-only species distribution models are sensitive to sample prevalence: Evaluating models using spatial prediction stability and accuracy metrics. *Ecological Modelling*, **431**, 109194.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., …, Buckley, Y.M. (2013) Predicting species distributions for conservation decisions. *Ecology Letters*, **16**, 1424–1435.

Guisan, A., Petitpierre, B., Broennimann, O., Daehler, C. & Kueffer, C. (2014) Unifying niche shift studies: insights from biological invasions. *Trends in Ecology & Evolution*, **29**, 260–269.

Guisan, A., Thuiller, W., Zimmermann, N. (2017). Habitat suitability and distribution models with applications in R. Cambridge, Cambridge University Press.

Hallgren, W., Santana, F., Low-Choy, S., Zhao, Y. & Mackey, B. (2019) Species distribution models can be highly sensitive to algorithm configuration. *Ecological Modelling*, **408**, 108719.

Hannah, L., Midgley, G., Andelman, S., Araújo, M., Hughes, G., Martinez-Meyer, E., Pearson, R. & Williams, P. (2007) Protected area needs in a changing climate. *Frontiers in Ecology and the Environment*, **5**, 131–138.

Hattab, T., Garzón-López, C.X., Ewald, M., Skowronek, S., Aerts, R., Horen, H., …, Lenoir, J. (2017) A unified framework to model the potential and realized distributions of invasive species within the invaded range. *Diversity and Distributions*, **23**, 806–819.

Hijmans, R. (2020). raster: Geographic Data Analysis and Modeling. R package version 3.1-5. https://CRAN.R-project.org/package=raster

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Ives, A.R. & Helmus, M.R. (2011) Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs*, **81**, 511–525.

de Lafontaine, G., Amasifuen Guerra, C.A., Ducousso, A. & Petit, R.J. (2014) Cryptic no more: soil macrofossils uncover Pleistocene forest microrefugia within a periglacial desert. *New Phytologist*, **204**, 715–729.

Lecocq, T., Harpke, A., Rasmont, P. & Schweiger, O. (2019) Integrating intraspecific differentiation in species distribution models: Consequences on projections of current and future climatically suitable areas of species. *Diversity and Distributions*, **25**, 1088–1100.

Leroy, B., Delsol, R., Hugueny, B., Meynard, C.N., Barhoumi, C., Barbet-Massin, M. & Bellard, C. (2018) Without quality presence-absence data, discrimination metrics such as TSS can be misleading measures of model performance. *Journal of Biogeography*, **45**, 1994–2002.

Liu, C., Newell, G. & White, M. (2018) The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. *Ecography*, **42**, 535–548.

Liu, C., Newell, G. & White, M. (2016) On the selection of thresholds for predicting species occurrence with presence-only data. *Ecology and Evolution*, **6**, 337–348.

Liu, C., White, M. & Newell, G. (2011) Measuring and comparing the accuracy of species distribution models with presence-absence data. *Ecography*, **34**, 232–243.

Liu, Canran, White, M. & Newell, G. (2009) Measuring the accuracy of species distribution models: a review. 8.

Lobo, J.M., Jiménez-Valverde, A. & Hortal, J. (2010) The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, **33**, 103–114.

Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2007) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Magri, D., Vendramin, G.G., Comps, B., Dupanloup, I., Geburek, T., Gomory, D., Latalowa, M., Litt, T., Paule, L., Roure, J.M., Tantau, I., van der Knaap, W.O., Petit, R.J. & de Beaulieu, J.-L. (2006) A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New Phytologist*, **171**, 199–221.

Mainali, K.P., Warren, D.L., Dhileepan, K., McConnachie, A., Strathie, L., Hassan, G., Karki, D., Shrestha, B.B. & Parmesan, C. (2015) Projecting future expansion of invasive species: comparing and improving methodologies for species distribution modeling. *Global Change Biology*, **21**, 4464–4480.

Marcia Barbosa, M. A., Real, R., Muñoz, A.-R. & Brown, J.A. (2013) New measures for assessing model equilibrium and prediction mismatch in species distribution models. *Diversity and Distributions*, **19**, 1333–1338.

Mauri, A., Strona, G. & San-Miguel-Ayanz, J. (2017) EU-Forest, a high-resolution tree occurrence dataset for Europe. *Scientific Data*, **4**.

Meynard, C.N., Gay, P.-E., Lecoq, M., Foucart, A., Piou, C. & Chapuis, M.-P. (2017) Climate-driven geographic distribution of the desert locust during recession periods: Subspecies' niche

differentiation and relative risks under scenarios of climate change. *Global Change Biology*, **23**, 4739–4749.

Naimi, B., Na, H. Groen, T.A., Skidmore, A.K. & Toxopeus, A.G. (2014). Where is positional uncertainty a problem for species distribution modelling. *Ecography*, **37**, 191-203. doi:10.1111/j.1600-0587.2013.00205.x

Pebesma, E.J. & Bivand, R. (2005). Classes and methods for spatial data in R. R News 5 (2), https://cran.r-project.org/doc/Rnews/.

Palma, R.E., Gutiérrez-Tapia, P., González, J.F., Boric-Bargetto, D. & Torres-Pérez, F. (2017) Mountaintops phylogeography: A case study using small mammals from the Andes and the coast of central Chile. *PLOS ONE*, **12**, e0180231.

Pearman, P.B., D'Amen, M., Graham, C.H., Thuiller, W. & Zimmermann, N.E. (2010) Within-taxon niche structure: niche conservatism, divergence and predicted effects of climate change. *Ecography*, **33**, 990–1003.

Pearson, R.G., Raxworthy, C.J., Nakamura, M. & Townsend Peterson, A. (2007) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar: Predicting species distributions with low sample sizes. *Journal of Biogeography*, **34**, 102–117.

Peterson, M.L., Doak, D.F. & Morris, W.F. (2018) Incorporating local adaptation into forecasts of species' distribution and abundance under climate change. *Global Change Biology*, **25**, 775–793.

Poli, P., Lenoir, J., Plantard, O., Ehrmann, S., Røed, K.H., Leinaas, H.P., Panning, M. & Guiller, A. (2020) Strong genetic structure among populations of the tick Ixodes ricinus across its range. *Ticks and Tick-borne Diseases*, **11**, 101509.

Schulte, U., Hochkirch, A., Lötters, S., Rödder, D., Schweiger, S., Weimann, T. & Veith, M. (2012) Cryptic niche conservatism among evolutionary lineages of an invasive lizard: Intraspecific niche conservatism. *Global Ecology and Biogeography*, **21**, 198–211.

Shabani, F., Kumar, L. & Ahmadi, M. (2018) Assessing Accuracy Methods of Species Distribution Models: AUC, Specificity, Sensitivity and the True Skill Statistic. 13.

Simon Garnier (2018). viridis: Default Color Maps from 'matplotlib'. R package version 0.5.1. https://CRAN.R-project.org/package=viridis

Smith, A.B., Godsoe, W., Rodríguez-Sánchez, F., Wang, H.-H. & Warren, D. (2018) Niche Estimation Above and Below the Species Level. *Trends in Ecology & Evolution*, **34**, 260–273.

Svenning, J.-C., Normand, S. & Kageyama, M. (2008) Glacial refugia of temperate trees in Europe: insights from species distribution modelling. *Journal of Ecology*, **96**, 1117–1127.

Thuiller, W., Albert, C., Araújo, M.B., Berry, P.M., Cabeza, M., Guisan, A., …, Zimmermann, N.E. (2008) Predicting global change impacts on plant species' distributions: Future challenges. *Perspectives in Plant Ecology, Evolution and Systematics*, **9**, 137–152.

Thuiller, W., Georges, D., Engler,R. & Breiner, F. (2020). biomod2: Ensemble Platform for Species Distribution Modeling. R package version 3.4.12.

Václavík, T. & Meentemeyer, R.K. (2012) Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion: Equilibrium and invasive species distribution models. *Diversity and Distributions*, **18**, 73–83.

Vale, C.G., Tarroso, P. & Brito, J.C. (2014) Predicting species distribution at range margins: testing the effects of study area extent, resolution and threshold selection in the Sahara-Sahel transition zone. *Diversity and Distributions*, **20**, 20–33.

Wasof, S., Lenoir, J., Gallet-Moron, E., Jamoneau, A., Brunet, J., Cousins, S.A.O., De Frenne, P., Diekmann, M., Hermy, M., Kolb, A., Liira, J., Verheyen, K., Wulf, M. & Decocq, G. (2013) Ecological niche shifts of understorey plants along a latitudinal gradient of temperate forests in north-western Europe: Species' realized-niche shifts across latitude. *Global Ecology and Biogeography*, **22**, 1130–1140.

Zimmermann, N.E., Edwards, T.C., Graham, C.H., Pearman, P.B. & Svenning, J.-C. (2010) New trends in species distribution modelling. *Ecography*, **33**, 985–989.

Chapitre 5 General Discussion

In this thesis I explored some of the advantages of coupling SDMs and genetic information at intraspecific level. Among the four model species originally meant to develop the research, it was possible to accomplish all of the thesis objectives with three of them: *Ixodes ricinus*, *Geum urbanum* and *Fagus sylvatica. Oxalis acetosella* was the only one for which phylogeographic or population genetics analyses could not be carried. The main reason for incorporating both *O. acetosella* and *Geum urbanum* on the thesis was their contrasting level of specialization (specialist *vs.* generalist), mode of reproduction (mix of cross-fertilisation and self-fertilisation *vs.* preferably self-fertiliser), and mode of dispersion (autochory *vs.* adhesive dispersion). Those different life-strategies have the potential to generate different patterns of genetic structure (Vandepitte *et al.* 2007, Schmidt *et al.* 2009, Aoki *et al.* 2019), and could have allowed for a deeper investigation of the benefits and limitations of the intersections between SDMs and population genetics. I was able nonetheless to compare the other answer the main questions of this thesis.

By analysing dynamics of gene flow, genetic variability, and spatial structuring of populations at different temporal and spatial scales, phylogeography and population genetics - in a broad definition - helps to examine how historical process and microevolutionary processes as natural selection, genetic drift and migration, shape genetic variation through time and space in order to postulate hypothesis about the evolution of a given species. By correlating one species distribution and environmental variables, SDMs allows to project the species' niche into the geographic space. In this thesis, I showed that SDMs can be applied to test hypothesis from population genetics, and that genetic units clustering and phylogeographic analysis may help improve SDM projections (**Figure 5.1**).

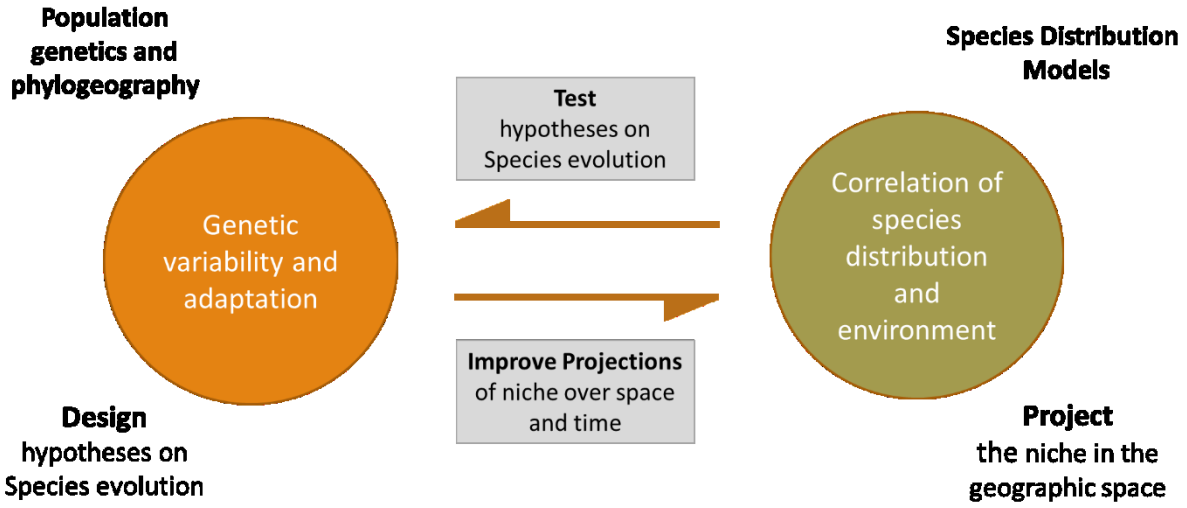


Figure 5.1. Schematic organisation of the contributions of coupling the information about the spatial genetic variability of a species and SDMs to understand its evolutionary history, as developed in this thesis. Population genetics and phylogeography helps to design hypothesis concerning the evolution of species, while SDMs may be applied to test those hypotheses. Conversely, the distribution of the genetic variation within a species may improve the niche projections over space and time.

Contributions of species distribution models to understand the evolutionary history of species

Species distribution models (SDMs) have great potential to improve our understanding of some of the mechanisms behind the observed patterns of genetic structure among populations. Considering *Ixodes ricinus* and *Geum urbanum*, it provided support to the hypothesis that post-glacial expansion dynamics are a main driving force that shaped current population genetic structure. As summarised in chapter 3 for two of our studied species, SDMs allowed: (**i**) the identification of particular loci (or alleles) under bioclimatic selection and potentially responding to changes in habitat quality since the last glacial maximum (LGM); (**ii**) the confirmation of a within group differentiation; and (**iii**) the discussion of hypotheses underlying the evolution of those species.

Applying SDMs to test hypotheses of gene flow

Regarding *G. urbanum*, albeit an isolation by distance (IBD) pattern was detected throughout the whole study area of Europe, the IBD pattern was neither observed in the Northern cluster nor in the Southern genetic cluster. My first hypothesis was that the separation of a Northern and a Southern cluster as a result of range expansion after the LGM was responsible for the contrast of a continental IBD pattern and absence of the same pattern within clusters. This hypothesis would roughly correspond to one of the assumptions that Slatkin (1993) advanced to explain a similar case of spatial genetic differentiation at global *versus* regional scales where the global IBD pattern was simply the result of recent colonisation by one of two genetic groups he investigated. The genetic data analysed in chapter 1 alone did not allow to test this hypothesis, but applying the regression analysed of genetic differentiation as a function of changes in habitat suitability helped a lot. When the species' genetic structure was tested against changes in habitat suitability according to contemporary and LGM projections, it became clear that range expansion alone cannot explain the observed (lack of) IBD, and that other factors are in play. Previous landscape genetics studies focusing on *G. urbanum* did not find any IBD pattern between populations across relatively restricted (regional) geographical extents (Vandepitte *et al.* 2007, Schmidt *et al.* 2009), suggesting an important degree of isolation between populations, especially when those results are compared with the high levels of inbreeding measured by F_{IS} across all those studies. The species mode of fertilisation could explain those results since gene exchange is mainly occurring from dispersed seeds. In this respect, the IBD pattern observed at the continental scale of Europe could be a remnant imprint of an older range expansion dynamic. As highlighted in the introduction of this thesis, our research team (EDYSAN, UMR CNRS 7058) is conducting, within the framework of the Woodnet BiodivERsA and the FORHAIE projects, a landscape genetic study comparing the influence of different land uses between contrasting landscape windows in France (**Figure 5.2**) to test if it affects the genetic structure of different taxa, including *G*. *urbanum*. Although covering a regional extent, the preliminary results from those projects seem to agree with

the idea that populations are structured by past gene flow. Bayesian coalescent-based methods (Beerli and Felsenstein 2001, Wilson & Rannala 2003) performed to unravel historical from contemporary effects of gene flows on genetic structure of populations suggests that *G. urbanum* responds to habitat fragmentation with a delay of several generations (**Figure 5.3**a, adapted from Guiller et al., unpublished). The clustering analysis using the Bayesian method implemented in STRUCTURE also shows that even populations from two landscape windows close to each other (around 27 km) still exhibit strong isolation (**Figure 5.3**b). Hence, considering those previous studies and the results of the coupled analyse of SDM and the genetic structure of the species presented in chapter 3, the continental IBD pattern found in chapter 2 could correspond to a genetic imprint of a 'historical' meta-population dynamics. In this sense, IBD is not only a matter of the geographical extent being covered (Slatkin 1993) but also a matter of time.

Figure 5.2. Arial photograph of the intensively managed openfield landscape window (OT) (a) and the low managed bocage landscape window (BT) (b) investigated in the Woodnet and FORHAIE projects, showing the studied forest patches (dark green) located in North of France (orange dot).

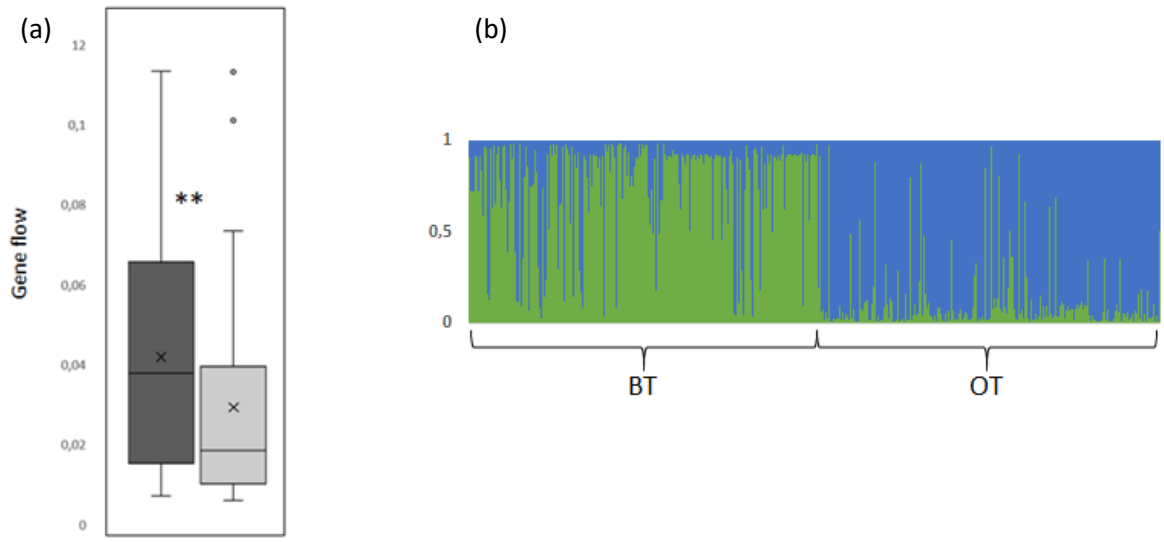


Figure 5.3. Genetic analyses of *Geum urbanum* occurring across two contrasting landscape windows (OT: Openfield Thiérache; BT: Bocage Thiérache) of 25 km^2 each in the Thièrache region (Picardy, Hauts-de-France) (source: Woodnet and FORHAIE projects) (modified from Guiller, et al., unpublished). In (a) is shown the comparison of historical (dark grey) *vs* contemporary (light grey) gene flow as estimated in Migrate v4 (Beerli & Falsenstein, 2001) and BAYESASS 1·3 (Wilson & Rannala, 2003) (** p = 0.0037). In (b) the individual assignment probabilities to one of the two clusters identified in STRUCTURE (Pritchard et al., 2000). The two clusters in (b) coincides with populations (fragments) from each of the two landscape windows.

Perspectives for integrating population genetics into species distribution models

The scientific literature focusing on the interface between population genetics and niche modelling suggests that modelling independently genetically differentiated populations (or lineages) could provide more reliable predictions of the species response as a whole (Pearman *et al.* 2010, Peterson *et al.* 2018, Smith *et al.* 2018, Boyer *et al.* 2020). As discussed in chapter 4, this general hypothesis may not be true in all cases. Nonetheless, incorporating a species genetic structure into SDMs gives important information about niche differentiation among different lineages belonging to the same species (Palma *et al.* 2017) and thus about the species niche conservatism hypothesis (Gutiérrez-Rodríguez *et al.* 2017, Meynard *et al.* 2017).

Collart *et al.* (2020) pointed out that splitting the occurrence data at the infra-species level and calibrating separated SDMs is not advisable unless there is solid evidence of niche divergence between intraspecific groups. Niche divergence is frequently only evaluated by comparing the response curves to climatic variables within each group. However, if presence-absence data or presence-only data for two lineages are geographically separated, there is no doubt that the response curves (the modelled niche) of those individualized lineages will be different. To the best of my knowledge, all the published studies that applied the 'lineage-modelling' approach dealt with geographically separated units. Yet, the geographic isolation of populations and lineages may be the result of barriers to gene flow and genetic drift, and not necessarily the result of changes in the potential niche due to local adaptation. In other words, differences in the response curves of intraspecific genetic units (lineages, genetic clusters or populations) may not represent the fundamental niche of those groups. One way of assessing niche divergence as a consequence of the species evolution is to investigate the differences in allele frequencies of genes correlated to climatic variables between the infra-species genetic groups (lineages or genetic clusters). Significant differences in allele frequencies of those genes would be a strong argument of niche differentiation as a consequence of divergent evolution between those groups and not a geographic coincidence or a demographic effect.

True niche differentiation seems to be the case for *I. ricinus*. All the loci identified as being under potential bioclimatic selection by three complementary methods also responded significantly to habitat suitability changes since the LGM. In this sense, it was possible to identify that at least some of those loci participate on important physiological and cytological functions. At least one SNP loci seems to be under strong selection (X234508), as it was identified by all applied methods and was correlated to two important bioclimatic variables (mean annual temperature and mean diurnal temperature range). The Bayes factor for the correlation between allele frequencies across and changes in habitat suitability between from the LGM to the period 1970-2000 was also the strongest among those correlations. After the Blast query, this SNP was target as part of the coding region of a G-protein-

coupled kinase 1. The family of transmembrane G-protein-coupled receptor mediate cellular response to various stimuli. Among others, they respond to changes in light and presence of odorants and hormones (Bockaert 1999, Gurevich & Gurevich 2019). Although I could not find any article on the role of this receptor in *Ixodes* species (nor on Acarians in general), it is possible that the protein is engaged in similar roles. In any case, the strong selection signal could represent a gain in fitness linked to one allele in populations from one of the two clusters, or at least a strong imprint of the species evolution in Europe. Another interesting case was the loci X313057, whose allele frequencies were varied significantly with the near present habitat suitability. The locus is associated with the production of mRNA for the putative nuclease HARBI1. In ants, this protein under expressed in conditions of cold stress (Tonione *et al.* 2020), and it is thus possible that it have the same role in *Ixodes* species. It is important to note though that this locus was not identified as under selection by any of the applied methods, and this significant response could be a false positive. Since the genome of *I. ricinus* is not entirely known, it was not possible to identify the physiological role of all the loci supposedly under selection. According to those results, it is likely that niche differences between populations of *I. ricinus*, as measured by SDM, are the result of directional selection. Coupling SDMs and analysis of the distribution of allele frequencies in loci under selection could be an important tool to identify truly niche divergence.

Modelling invasive species and range expansion

Modelling the distribution of invasive alien species is a typical case for which the equilibrium assumption of SDMs is violated, notably in the early stages of the invasion (Zimmermann *et al.* 2010). The equilibrium assumption violation could lead to an underestimation of the potential range of the invading species (Václavík & Meentemeyer 2012) or, in other words, a reduction in the number of commission errors at the expense or omission errors. It is intuitive to suppose that the increase in the number of omission errors could be more pronounced at the edge of the environmental space potentially available for the species. In this case, informing SDMs of the genetic structure of the target

invasive alien species could reduce omission errors and lead to more accurate and reliable predictions of the invasion risk, which is especially relevant for early detection. An interesting case study would be *G. urbanum*. It would be clarifying to study niche divergence between the native and the invaded range of the species. Population genetics and phylogeographic analysis can help identify the origin of invasive populations in the native range, while SDMs coupled with the genetic analyses can help identify niche divergence between native and invasive "entities". Bringing together those two methods could allow to better understand how the species range expansion (in the invaded range) are influencing the genetic differentiation of the two groups (native *vs.* invasive) and eventually justify modelling independently the future distribution of those entities.

Another case where the equilibrium assumption is violated is when species are shifting their range to track the isotherms that are also shifting due to climate change. As demonstrated by Garnier and Lewis (2016), this disequilibrium dynamic will be amplified the faster the climate is changing, because species might be unable to keep track of the high-speed pace at which isotherms are moving. This mismatch is likely to be, even more problematic when the expansion dynamic at the leading edge is accompanied by a concomitant contraction dynamic at the trailing edge. According to (Alkishe *et al.* 2017), this is exactly the case for *I. ricinus*, for which the spatial extent of suitable habitat is expected to increase at the northern range limit of the species distribution while it is expected to reduce at the southern range limit. In this scenario and considering the literature on species range expansion (Slatkin 1996, Excoffier & Ray 2008, Excoffier *et al.* 2009, Neve *et al.* 2009, Garnier & Lewis 2016), it is expected that rare alleles, a new combination of alleles, and new mutations become fixed in the populations located at the leading edge, possibly leading to a more pronounced niche divergence between populations simple due to demographic dynamics. In this case, genetically-based SDMs could better predict the species' future potential distribution. This could be one of the reasons why the genetically informed SDMs of *I. ricinus* showed a better performance than those of *F. sylvatica* in the calibration period. Since the mean response to the bioclimatic variables is truly different between the two *I. ricinus*

genetic clusters, modelling each cluster individually can bring into light the finer response of each cluster. The assembled model than reflects the individual bioclimatic niche of each of those genetic units.

General limits

One important limitation of many of the analyses in chapters 3 and 4 are related to the model calibration itself. The pipeline analysis described in chapter 3 was mainly based on the differences in habitat suitability between two geological times. Projecting the focal species distributions in time is, therefore, a precondition to this approach. In this case, the main limitation is that some important variables defining a species niche may not be available for both periods. Since the performance of SDM is dependent on the choice of the proper predictor variables (Araújo *et al.* 2019), the lack of adequate data could negatively impact the model projections, leading to unrealistic probabilities of occurrence. The magnitude of this negative impact is proportional to the importance of the missing variable to the species niche, and so will be the reliability of the comparisons between projected habitat suitability and population genetics metrics.

Another relevant limitation of the results presented in this thesis concerns the distribution of occurrences and their assignment to one particular genetic unit. As previously mentioned, splitting occurrences into infra-specific genetic groups leads to a smaller dataset compared to the species level (Stockwell and Peterson 2002, Collart *et al.* 2020), potentially leading to model overfitting (Breiner *et al.* 2015) and loss in model accuracy (Wisz *et al.* 2008). In chapter 4, I have circumvented this issue by assigning presences from public databases to one of the analysed genetic groups by geographic proximity. This method is only possible in cases where the focal species has a geographically explicit genetic structure. This issue becomes more relevant when applying the pipeline presented in chapter 3 since all the analyses were based on allele frequencies across populations, which are not available in public databases. Recently, Breiner *et al.* (2015, 2018) developed a set of techniques called Ensembles of Small Models (ESMs) to address the issue of small datasets, but the performance of those

techniques, when employed in infra-specific genetic groups, was put into question by the results of Collart et al. (2020). Of course, the best way to truly address the problem of small datasets in intraspecific genetically informed SDMs is to augment the number of sampled populations in the study range, but then the cost of genetic analysis becomes another relevant issue. In this sense, sampling should be carefully planned, specifically to test the hypothesis that can be addressed by coupling SDMs, population genetics, and phylogeography such as (but not limited to) niche differentiation between lineages, evolution consequences of range expansion, and intraspecific responses to climate change. Sampling design should search for the best compromise between the number of samples per site and the total number of sampled locations. It is also important to sample the species range in as much regular grid as possible. The uniform distribution of samples would provide higher level of certitude of the geographical discontinuities between genetic entities (groups or lineages), possibly allowing higher confidence when assigning occurrences from public databases to one of those genetic units.

Conclusion

Habitat suitability models aim at projecting in the geographical space some of the dimensions of the ecological space (i.e. the niche) of a certain level of biological organisation (population, species, clades, communities, etc.). The results of my thesis do suggest that the research fields of SDMs and population genetics (and phylogeography) can benefit each other almost reciprocally, by helping to identify allele under bioclimatic selection and validate hypotheses of genetic differentiation as a result of climate changes (chapter 3), or by informing SDMs (chapter 4) with the species genetic structure so that it can, in some cases, help improve model performance and the identification of cryptic suitable habitat. Further investigation is nonetheless needed to determine exactly how and when to incorporate this genetic information into SDMs. There are two questions of utmost importance concerning the geographical distribution and the niche divergence of intraspecific groups. Most of the studies that

investigated the incorporation of the species genetic structure into SDMs dealt with somewhat geographically well separated genetic groups, including the species we analysed here. In this sense, species that show genetic population divergence in the same geographic extent could be good models to understand the limits of incorporating genetic information into SDMs. The question of niche divergence between genetic groups belonging to the same species should also be investigated more specifically, since an absence of niche divergence should preclude the use of lineage-based SDMs, simply because modelling individual intraspecific groups would represent a more complex model without any biological relevance. In this sense, I would advocate for deeper research about the concomitant application of the whole species SDM and the investigation of genetic variation between populations (or any other level of organisation) in candidate loci under pressure selection from the environmental variables used in model calibration.

References

Abellán, P. and Svenning, J.-C., 2014. Refugia within refugia - patterns in endemism and genetic divergence are linked to Late Quaternary climate stability in the Iberian Peninsula: Climatic Stability and Biodiversity. *Biological Journal of the Linnean Society*, 113 (1), 13–28.

Alkishe, A.A., Peterson, A.T., and Samy, A.M., 2017. Climate change influences on the potential geographic distribution of the disease vector tick *Ixodes ricinus*. *PLOS ONE*, 12 (12), e0189092.

Aoki, S., Ohi-Toma, T., and Murata, J., 2019. Taxonomic Revision of Oxalis subsect. Oxalis (Oxalidaceae), 70, 14.

Araújo, M.B., Anderson, R.P., Márcia Barbosa, A., Beale, C.M., Dormann, C.F., Early, R., Garcia, R.A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R.B., Zimmermann, N.E., and Rahbek, C., 2019. Standards for distribution models in biodiversity assessments. *Science Advances*, 5 (1), eaat4858.

Arens, P., Durka, W., Wernke-Lenting, J.H., and Smulders, M.J.M., 2004. Isolation and characterization of microsatellite loci in Geum urbanum (Rosaceae) and their transferability within the genus Geum. *Molecular Ecology Notes*, 4 (2), 209–212.

Assis, J., Coelho, N.C., Lamy, T., Valero, M., Alberto, F., and Serrão, E.Á., 2016. Deep reefs are climatic refugia for genetic diversity of marine forests. *Journal of Biogeography*, 43 (4), 833–844.

Baguette, M., Blanchet, S., Legrand, D., Stevens, V.M., and Turlure, C., 2013. Individual dispersal, landscape connectivity and ecological networks: Dispersal, connectivity and networks. *Biological Reviews*, 88 (2), 310–326.

Barbet-Massin, M., Jiguet, F., Albert, C.H., and Thuiller, W., 2012. Selecting pseudo-absences for species distribution models: how, where and how many?: *How to use pseudo-absences in niche modelling? Methods in Ecology and Evolution*, 3 (2), 327–338.

Bateman, B.L., Pidgeon, A.M., Radeloff, V.C., VanDerWal, J., Thogmartin, W.E., Vavrus, S.J., and Heglund, P.J., 2016. The pace of past climate change vs. potential bird distributions and land use in the United States. *Global Change Biology*, 22 (3), 1130–1144.

Beerli, P. and Felsenstein, J., 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences*, 98 (8), 4563–4568.

Bellwood, D.R., Goatley, C.H.R., and Bellwood, O., 2017. The evolution of fishes and corals on reefs: form, function and interdependence: The evolution of fishes and corals on reefs. *Biological Reviews*, 92 (2), 878–901.

Benjamini, Y. and Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57 (1), 289–300.

Benjamini, Y. and Yekutieli, D., 2001. THE CONTROL OF THE FALSE DISCOVERY RATE IN MULTIPLE TESTING UNDER DEPENDENCY. *The Annals of Statistics*, 29 (4), 1165–1188.

Berg, H. and Redbo-torstensson, P., 1998. Cleistogamy as a bet-hedging strategy in *Oxalis acetosella*, a perennial herb. *Journal of Ecology*, 86 (3), 491–500.

Berg, H., 2000. Differential seed dispersal in Oxalis acetosella, a cleistogamous perennial herb. *Acta Oecologica*, 21 (2), 109–118.

Berg, M.P., Kiers, E.T., Driessen, G., van der HEIJDEN, M., Kooi, B.W., Kuenen, F., Liefting, M., Verhoef, H.A., and Ellers, J., 2010. Adapt or disperse: understanding species persistence in a changing world. *Global Change Biology*, 16 (2), 587–598.

Biswas, S.B. and Sarkar, A.K., 1970. Deoxyribonucleic acid base composition of some angiosperms and its taxonomic significance. *Phytochemistry*, 9 (12), 2425–2430.

Bivand, R. S., Pebesma, E., Gomez-Rubio, V. (2013). Applied spatial data analysis with R, Second edition. Springer, NY. http://www.asdar-book.org/

Bockaert, J., 1999. Molecular tinkering of G protein-coupled receptors: an evolutionary success. *The EMBO Journal*, 18 (7), 1723–1729.

Boyer, I., Cayuela, H., Bertrand, R., and Isselin-Nondedeu, F., 2020. Improving biological relevance of model projections in response to climate change by considering dispersal amongst lineages in an amphibian. *Journal of Biogeography*.

Bradburd, G.S., Coop, G.M., and Ralph, P.L., 2018. Inferring Continuous and Discrete Population Genetic Structure Across Space. *Genetics*, (210), 33–52.

Breiner, F.T., Guisan, A., Bergamini, A., and Nobis, M.P., 2015. Overcoming limitations of modelling rare species by using ensembles of small models. *Methods in Ecology and Evolution*, 6 (10), 1210–1218.

Breiner, F.T., Nobis, M.P., Bergamini, A., and Guisan, A., 2018. Optimizing ensembles of small models for predicting the distribution of species with few occurrences. *Methods in Ecology and Evolution*, 9 (4), 802–808.

Brito, P.H., 2005. The influence of Pleistocene glacial refugia on tawny owl genetic diversity and phylogeography in western Europe. *Molecular Ecology*, 14 (10), 3077–3094.

Carson, H.L. and Templeton, A.R., 1984. Genetic revolutions in relation to speciation phenomena: the founding of new populations. *Annual Review of Ecology and Systematics*, 15, 97–131.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J.H., 2011. *Stacks* : Building and Genotyping Loci *De Novo* From Short-Read Sequences. *G3: Genes|Genomes|Genetics*, 1 (3), 171–182.

Caudullo, G., Welk, E., and San-Miguel-Ayanz, J., 2017. Chorological maps for the main European woody species. *Data in Brief*, 12, 662–666.

Chardon, N.I., Pironon, S., Peterson, M.L., and Doak, D.F., 2020. Incorporating intraspecific variation into species distribution models improves distribution predictions, but cannot predict species traits for a wide-spread plant species. *Ecography*, 43 (1), 60–74.

Charlesworth, D. and Willis, J.H., 2009. The genetics of inbreeding depression. *Nature Reviews Genetics*, 10 (11), 783–796.

Clark, P.U., Dyke, A.S., Shakun, J.D., Carlson, A.E., Clark, J., Wohlfarth, B., Mitrovica, J.X., Hostetler, S.W., and McCabe, A.M., 2009. The Last Glacial Maximum. *Science*, 325 (5941), 710–714.

Collart, F., Hedenäs, L., Broennimann, O., Guisan, A., and Vanderpoorten, A., 2020. Intraspecific differentiation: Implications for niche and distribution modelling. *Journal of Biogeography*.

Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J.K., 2010. Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics*, 185 (4), 1411–1423.

Cooper, N., Jetz, W., and Freckleton, R.P., 2010. Phylogenetic comparative approaches for studying niche conservatism: Comparative approaches for niche conservatism. *Journal of Evolutionary Biology*, 23 (12), 2529–2539.

Davis, M.B. and Shaw, R.G., 2001. Range Shifts and Adaptive Responses to Quaternary Climate Change. *Science*, 292 (5517), 673–679.

Dawe, K.L. and Boutin, S., 2016. Climate change is the primary driver of white-tailed deer (*Odocoileus virginianus*) range expansion at the northern extent of its range; land use is secondary. *Ecology and Evolution*, 6 (18), 6435–6451.

Delord, C., Lassalle, G., Oger, A., Barloy, D., Coutellec, M., Delcamp, A., Evanno, G., Genthon, C., Guichoux, E., Le Bail, P., Le Quilliec, P., Longin, G., Lorvelec, O., Massot, M., Reveillac, E., Rinaldo, R., Roussel, J., Vigouroux, R., Launey, S., and Petit, E.J., 2018. A cost-and-time effective procedure to develop SNP markers for multiple species: A support for community genetics. *Methods in Ecology and Evolution*, 9 (9), 1959–1974.

Demesure, B., Comps, B., and Petit, R.J., 1996. CHLOROPLAST DNA PHYLOGEOGRAPHY OF THE COMMON BEECH (*FAGUS SYLVATICA* L.) IN EUROPE. *Evolution*, 50 (6), 2515–2520.

Denk, T., Grimm, G., Stögerer, K., Langer, M., and Hemleben, V., 2002. The evolutionary history of Fagus in western Eurasia: Evidence from genes, morphology and the fossil record. *Plant Systematics and Evolution*, 232 (3–4), 213–236.

Diniz-Filho, J.A.F., Barbosa, A.C.O.F., Collevatti, R.G., Chaves, L.J., Terribile, L.C., Lima-Ribeiro, M.S., and Telles, M.P.C., 2016. Spatial autocorrelation analysis and ecological niche modelling allows inference of range dynamics driving the population genetic structure of a Neotropical savanna tree. *Journal of Biogeography*, 43 (1), 167–177.

Dorren, L.K.A., Berger, F., le Hir, C., Mermin, E., and Tardif, P., 2005. Mechanisms, effects and management implications of rockfall in forests. *Forest Ecology and Management*, 215 (1–3), 183–195.

Dray, S. and Dufour, A.-B., 2007. The **ade4** Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, 22 (4).

Durrant, T.H., de Rigo, D., and Caudullo, G., 2016. Fagus sylvatica in Europe: distribution, habitat, usage and threats. *In*: J. San-Miguel-Ayanz, D. de Rigo, G. Caudullo, T. Houston Durrant, and A. Mauri, eds. *European Atlas of Forest Tree Species*. Publication Office of the European Union, Luxembourg, 2.

Durrant, T.H., de Rigo, D., and Caudullo, G., n.d. Fagus sylvatica in Europe: distribution, habitat, usage and threats, 2.

Dynesius, M. and Jansson, R., 2000. Evolutionary consequences of changes in species' geographical distributions driven by Milankovitch climate oscillations. *Proceedings of the National Academy of Sciences*, 97 (16), 9115–9120.

Earl, D.A. and vonHoldt, B.M., 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, 4 (2), 359–361.

Edmonds, C.A., Lillie, A.S., and Cavalli-Sforza, L.L., 2004. Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences*, 101 (4), 975–979.

Endels, P., Adriaens, D., Verheyen, K., and Hermy, M., 2004. Population structure and adult plant performance of forest herbs in three contrasting habitats. *Ecography*, 27 (2), 225–241.

Evanno, G., Regnaut, S., and Goudet, J., 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, 14 (8), 2611–2620.

Excoffier, L. and Ray, N., 2008. Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, 23 (7), 347–351.

Excoffier, L., Foll, M., and Petit, R.J., 2009. Genetic Consequences of Range Expansions. *Annual Review of Ecology, Evolution, and Systematics*, 40 (1), 481–501.

Fick, S.E. and Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37 (12), 4302–4315.

Foll, M. and Gaggiotti, O., 2008. A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics*, 180 (2), 977–993.

Freeman, B.G., Scholer, M.N., Ruiz-Gutierrez, V., and Fitzpatrick, J.W., 2018. Climate change causes upslope shifts and mountaintop extirpations in a tropical bird community. *Proceedings of the National Academy of Sciences*, 115 (47), 11982–11987.

Friesen, V.L., Burg, T.M., and McCOY, K.D., 2007. Mechanisms of population differentiation in seabirds: POPULATION DIFFERENTIATION IN SEABIRDS. *Molecular Ecology*, 16 (9), 1765–1785.

García, J.T., Domínguez-Villaseñor, J., Alda, F., Calero-Riestra, M., Pérez Olea, P., Fargallo, J.A., Martínez-Padilla, J., Herranz, J., Oñate, J.J., Santamaría, A., Motro, Y., Attie, C., Bretagnolle, V., Delibes, J., and Viñuela, J., 2020. A complex scenario of glacial survival in Mediterranean and continental refugia of a temperate continental vole species (*Microtus arvalis*) in Europe. *Journal of Zoological Systematics and Evolutionary Research*, 58 (1), 459–474.

Garnier, J. and Lewis, M.A., 2016. Expansion Under Climate Change: The Genetic Consequences. *Bulletin of Mathematical Biology*, 78 (11), 2165–2185.

Geppert, C., Perazza, G., Wilson, R.J., Bertolli, A., Prosser, F., Melchiori, G., and Marini, L., 2020. Consistent population declines but idiosyncratic range shifts in Alpine orchids under global change. *Nature Communications*, 11 (1).

Gómez, A. and Lunt, D.H., 2007. Refugia within Refugia: Patterns of Phylogeographic Concordance in the Iberian Peninsula. *In*: S. Weiss and N. Ferrand, eds. *Phylogeography of Southern European Refugia*. Dordrecht: Springer Netherlands, 155–188.

Goudet, J., Jombart, T. (2015). hierfstat: Estimation and Tests of Hierarchical F-Statistics. R package version 0.04-22. https://CRAN.R-project.org/package=hierfstat

Green, A.J., 2016. The importance of waterbirds as an overlooked pathway of invasion for alien species. *Diversity and Distributions*, 22 (2), 239–247.

Grime, J.P. and Mowforth, M.A., 1982. Variation in genome size—an ecological interpretation. *Nature*, 299 (5879), 151–153.

Grimmett, L., Whitsed, R., and Horta, A., 2020. Presence-only species distribution models are sensitive to sample prevalence: Evaluating models using spatial prediction stability and accuracy metrics. *Ecological Modelling*, 431, 109194.

Grivet, D. and Petit, R.J., 2002. Phylogeography of the common ivy (Hedera sp.) in Europe: genetic differentiation through space and time. *Molecular Ecology*, 11 (8), 1351–1362.

Guisan, A. and Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8 (9), 993–1009.

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B.A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H.P., and Buckley, Y.M., 2013. Predicting species distributions for conservation decisions. *Ecology Letters*, 16 (12), 1424–1435.

Guisan, A., Petitpierre, B., Broennimann, O., Daehler, C., and Kueffer, C., 2014. Unifying niche shift studies: insights from biological invasions. *Trends in Ecology & Evolution*, 29 (5), 260–269.

Guisan, A., Thuiller, W., Zimmermann, N. (2017). Habitat suitability and distribution models with applications in R. Cambridge, Cambridge University Press.

Günther, T. and Coop, G., 2013. Robust Identification of Local Adaptation from Allele Frequencies. *Genetics*, 195 (1), 205–220.

Günther, T., Coop, G. (2018). A short manual for Bayenv2.0. https://bitbucket.org/tguenther/bayenv2_public/src/master/

Gurevich, V.V. and Gurevich, E.V., 2019. GPCR Signaling Regulation: The Role of GRKs and Arrestins. *Frontiers in Pharmacology*, 10.

Gutiérrez-Rodríguez, J., Barbosa, A.M., and Martínez-Solano, Í., 2017. Integrative inference of population history in the Ibero-Maghrebian endemic Pleurodeles waltl (Salamandridae). *Molecular Phylogenetics and Evolution*, 112, 122–137.

Hancock, A.M., Witonsky, D.B., Alkorta-Aranburu, G., Beall, C.M., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J.K., Coop, G., and Di Rienzo, A., 2011. Adaptations to Climate-Mediated Selective Pressures in Humans. *PLoS Genetics*, 7 (4), e1001375.

Hannah, L., Midgley, G., Andelman, S., Araújo, M., Hughes, G., Martinez-Meyer, E., Pearson, R., and Williams, P., 2007. Protected area needs in a changing climate. *Frontiers in Ecology and the Environment*, 5 (3), 131–138.

Hartfield, M., Bataillon, T., and Glémin, S., 2017. The Evolutionary Interplay between Adaptation and Self-Fertilization. *Trends in Genetics*, 33 (6), 420–431.

Hattab, T., Garzón-López, C.X., Ewald, M., Skowronek, S., Aerts, R., Horen, H., Brasseur, B., Gallet-Moron, E., Spicher, F., Decocq, G., Feilhauer, H., Honnay, O., Kempeneers, P., Schmidtlein, S., Somers, B., Van De Kerchove, R., Rocchini, D., and Lenoir, J., 2017. A unified framework to model the potential and realized distributions of invasive species within the invaded range. *Diversity and Distributions*, 23 (7), 806–819.

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

Hewitt, G.M., 2000. The genetic legacy of the Quaternary ice ages. *Nature*, 405 (6789), 907–913.

Hewitt, G.M., 2004. Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359 (1442), 183–195.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., and Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25 (15), 1965–1978.

Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C., and Guisan, A., 2006. Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, 199 (2), 142–152.

Hofer, T., Ray, N., Wegmann, D., and Excoffier, L., 2009. Large Allele Frequency Differences between Human Continental Groups are more Likely to have Occurred by Drift During range Expansions than by Selection. *Annals of Human Genetics*, 73 (1), 95–108.

Holsinger, K.E. and Weir, B.S., 2009. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics*, 10 (9), 639–650.

Hvidsten, D., Frafjord, K., Gray, J.S., Henningsson, A.J., Jenkins, A., Kristiansen, B.E., Lager, M., Rognerud, B., Slåtsve, A.M., Stordal, F., Stuen, S., and Wilhelmsson, P., 2020. The distribution limit of the common tick, Ixodes ricinus, and some associated pathogens in north-western Europe. *Ticks and Tick-borne Diseases*, 11 (4), 101388.

Ives, A.R. and Helmus, M.R., 2011. Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs*, 81 (3), 511–525.

Janes, J.K., Miller, J.M., Dupuis, J.R., Malenfant, R.M., Gorrell, J.C., Cullingham, C.I., and Andrew, R.L., 2017. The *K* = 2 conundrum. *Molecular Ecology*, 26 (14), 3594–3602.

Jansson, R. and Dynesius, M., 2002. The Fate of Clades in a World of Recurrent Climatic Change: Milankovitch Oscillations and Evolution. *Annual Review of Ecology and Systematics*, 33 (1), 741–777.

Jin, D.-P., Lee, J.-H., Xu, B., and Choi, B.-H., 2016. Phylogeography of East Asian Lespedeza buergeri (Fabaceae) based on chloroplast and nuclear ribosomal DNA sequence variations. *Journal of Plant Research*, 129 (5), 793–805.

Jordan, C.Y., Lohse, K., Turner, F., Thomson, M., Gharbi, K., and Ennos, R.A., 2018. Maintaining their genetic distance: Little evidence for introgression between widely hybridizing species of *Geum* with contrasting mating systems. *Molecular Ecology*, 27 (5), 1214–1228.

Jore, S., Viljugrein, H., Hofshagen, M., Brun-Hansen, H., Kristoffersen, A.B., Nygård, K., Brun, E., Ottesen, P., Sævik, B.K., and Ytrehus, B., 2011. Multi-source analysis reveals latitudinal and altitudinal shifts in range of Ixodes ricinus at its northern distribution limit. *Parasites & Vectors*, 4 (1).

Kass, R.E. and Raftery, A.E., 1995. Bayes Factors. *Journal of the American Statistical Association*, 90 (430), 773.

Kissling, W.D. and Carl, G., 2007. Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecology and Biogeography*, 0 (0), 070618060123007-???

Klopfstein, S., Currat, M., and Excoffier, L., 2006. The Fate of Mutations Surfing on the Wave of a Range Expansion. *Molecular Biology and Evolution*, 23 (3), 482–490.

Kozakiewicz, C.P., Burridge, C.P., Funk, W.C., VandeWoude, S., Craft, M.E., Crooks, K.R., Ernest, H.B., Fountain-Jones, N.M., and Carver, S., 2018. Pathogens in space: Advancing understanding of pathogen dynamics and disease ecology through landscape genetics. *Evolutionary Applications*, 11 (10), 1763–1778.

Kuhn, E., Lenoir, J., Piedallu, C., and Gégout, J.-C., 2016. Early signs of range disjunction of submountainous plant species: an unexplored consequence of future and contemporary climate changes. *Global Change Biology*, 22 (6), 2094–2105.

Kühne, G., Kosuch, J., Hochkirch, A., and Schmitt, T., 2017. Extra-Mediterranean glacial refugia in a Mediterranean faunal element: the phylogeography of the chalk-hill blue Polyommatus coridon (Lepidoptera, Lycaenidae). *Scientific Reports*, 7 (1).

Lamprecht, A., Semenchuk, P.R., Steinbauer, K., Winkler, M., and Pauli, H., 2018. Climate change leads to accelerated transformation of high-elevation vegetation in the central Alps. *New Phytologist*, 220 (2), 447–459.

Lecocq, T., Harpke, A., Rasmont, P., and Schweiger, O., 2019. Integrating intraspecific differentiation in species distribution models: Consequences on projections of current and future climatically suitable areas of species. *Diversity and Distributions*, 25 (7), 1088–1100.

Legendre, P. and Legendre, L., 1998. *Numerical Ecology*. Amsterdam, The Neatherlands: Elsevier Science B.V.

Lenoir, J., Gegout, J.C., Marquet, P.A., de Ruffray, P., and Brisse, H., 2008. A Significant Upward Shift in Plant Species Optimum Elevation During the 20th Century. *Science*, 320 (5884), 1768–1771.

Lenoir, J., Bertrand, R., Comte, L., Bourgeaud, L., Hattab, T., Murienne, J., and Grenouillet, G., 2020. Species better track climate warming in the oceans than on land. *Nature Ecology & Evolution*, 4 (8), 1044–1059.

Lindgren, E. and Gustafson, R., 2001a. Tick-borne encephalitis in Sweden and climate change. *The Lancet*, 358 (9275), 16–18.

Lindgren, E. and Gustafson, R., 2001b. Tick-borne encephalitis in Sweden and climate change. *The Lancet*, 358 (9275), 16–18.

Liu, C., Newell, G., and White, M., 2018. The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. *Ecography*, 42 (3), 535–548.

Lotterhos, K.E. and Whitlock, M.C., 2014. Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. *Molecular Ecology*, 23 (9), 2178–2192.

Magri, D., Vendramin, G.G., Comps, B., Dupanloup, I., Geburek, T., Gomory, D., Latalowa, M., Litt, T., Paule, L., Roure, J.M., Tantau, I., van der Knaap, W.O., Petit, R.J., and de Beaulieu, J.-L., 2006. A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New Phytologist*, 171 (1), 199–221.

Magri, D., 2008. Patterns of post-glacial spread and the extent of glacial refugia of European beech (Fagus sylvatica). *Journal of Biogeography*, 35 (3), 450–463.

Maiorano, L., Cheddadi, R., Zimmermann, N.E., Pellissier, L., Petitpierre, B., Pottier, J., Laborde, H., Hurdu, B.I., Pearman, P.B., Psomas, A., Singarayer, J.S., Broennimann, O., Vittoz, P., Dubuis, A., Edwards, M.E., Binney, H.A., and Guisan, A., 2013. Building the niche through time: using 13,000 years of data to predict the effects of climate change on three tree species in Europe: Multi-temporal niche and species potential distribution. *Global Ecology and Biogeography*, 22 (3), 302–317.

Mallory, C.D. and Boyce, M.S., 2018. Observed and predicted effects of climate change on Arctic caribou and reindeer. *Environmental Reviews*, 26 (1), 13–25.

Manel, S. and Holderegger, R., 2013. Ten years of landscape genetics. *Trends in Ecology & Evolution*, 28 (10), 614–621.

Marks, G.E., 1956. CHROMOSOME NUMBERS IN THE GENUS OXALIS. *New Phytologist*, 55 (1), 120–129.

Marshall, L., Perdijk, F., Dendoncker, N., Kunin, W., Roberts, S., and Biesmeijer, J.C., 2020. Bumblebees moving up: shifts in elevation ranges in the Pyrenees over 115 years, 10.

Mauri, A., Strona, G., and San-Miguel-Ayanz, J., 2017. EU-Forest, a high-resolution tree occurrence dataset for Europe. *Scientific Data*, 4 (1).

McCoy, K.D., 2008. The population genetic structure of vectors and our understanding of disease epidemiology. *Parasite*, 15 (3), 444–448.

Medlock, J.M., Hansford, K.M., Bormane, A., Derdakova, M., Estrada-Peña, A., George, J.-C., Golovljova, I., Jaenson, T.G., Jensen, J.-K., and Jensen, P.M., 2013. Driving forces for changes in geographical distribution of Ixodes ricinus ticks in Europe. *Parasites & vectors*, 6 (1), 1.

Meier, E.S., Edwards Jr, T.C., Kienast, F., Dobbertin, M., and Zimmermann, N.E., 2011. Co-occurrence patterns of trees along macro-climatic gradients and their potential influence on the present

and future distribution of Fagus sylvatica L.: Influence of co-occurrence patterns on Fagus sylvatica. *Journal of Biogeography*, 38 (2), 371–382.

Mellado, A. and Zamora, R., 2014. Generalist birds govern the seed dispersal of a parasitic plant with strong recruitment constraints. *Oecologia*, 176 (1), 139–147.

Meynard, C.N., Gay, P.-E., Lecoq, M., Foucart, A., Piou, C., and Chapuis, M.-P., 2017. Climate-driven geographic distribution of the desert locust during recession periods: Subspecies' niche differentiation and relative risks under scenarios of climate change. *Global Change Biology*, 23 (11), 4739–4749.

Nei, M., Maruyama, T., and Chakraborty, R., 1975. THE BOTTLENECK EFFECT AND GENETIC VARIABILITY IN POPULATIONS. *Evolution*, 29 (1), 1–10.

Neve, G., Pavlicko, A., and Konvicka, M., 2009. Loss of genetic diversity through spontaneous colonization in the bog fritillary butterfly, Proclossiana eunomia (Lepidoptera: Nymphalidae) in the Czech Republic. *European Journal of Entomology*, 106 (1), 11–19.

Niedziałkowska, M., Jędrzejewska, B., Honnen, A.-C., Otto, T., Sidorovich, V.E., Perzanowski, K., Skog, A., Hartl, G.B., Borowik, T., Bunevich, A.N., Lang, J., and Zachos, F.E., 2011. Molecular biogeography of red deer Cervus elaphus from eastern Europe: insights from mitochondrial DNA sequences. *Acta Theriologica*, 56 (1), 1–12.

Nogués-Bravo, D., Rodríguez-Sánchez, F., Orsini, L., de Boer, E., Jansson, R., Morlon, H., Fordham, D.A., and Jackson, S.T., 2018. Cracking the Code of Biodiversity Responses to Past Climate Change. *Trends in Ecology & Evolution*, 33 (10), 765–776.

Nordborg, M. and Donnelly, P., 1997. The Coalescent Process With Selfing. *Genetics*, (146), 1185–1195.

Olivier Broennimann, Valeria Di Cola and Antoine Guisan (2020). ecospat: Spatial Ecology Miscellaneous Methods. R package version 3.1. https://CRAN.R-project.org/package=ecospat

Packham, J.R., Thomas, P.A., Atkinson, M.D., and Degen, T., 2012. Biological Flora of the British Isles: *Fagus sylvatica*. *Journal of Ecology*, 100 (6), 1557–1608.

Palma, R.E., Gutiérrez-Tapia, P., González, J.F., Boric-Bargetto, D., and Torres-Pérez, F., 2017. Mountaintops phylogeography: A case study using small mammals from the Andes and the coast of central Chile. *PLOS ONE*, 12 (7), e0180231.

Papageorgiou, A.C., Vidalis, A., Gailing, O., Tsiripidis, I., Hatziskakis, S., Boutsios, S., Galatsidas, S., and Finkeldey, R., 2008. Genetic variation of beech (Fagus sylvatica L.) in Rodopi (N.E. Greece). *European Journal of Forest Research*, 127 (1), 81–88.

Parchman, T.L., Jahner, J.P., Uckele, K.A., Galland, L.M., and Eckert, A.J., 2018. RADseq approaches and applications for forest tree genetics. *Tree Genetics & Genomes*, 14 (3).

Patterson, N., Price, A.L., and Reich, D., 2006. Population Structure and Eigenanalysis. *PLoS Genetics*, 2 (12), e190.

Pearman, P.B., D'Amen, M., Graham, C.H., Thuiller, W., and Zimmermann, N.E., 2010. Within-taxon niche structure: niche conservatism, divergence and predicted effects of climate change. *Ecography*, 33 (6), 990–1003.

Pearse, D.E. and Crandall, K.A., 2004. Beyond FST: Analysis of population genetic data for conservation. *Conservation Genetics*, 5 (5), 585–602.

Pecl, G.T., Araújo, M.B., Bell, J.D., Blanchard, J., Bonebrake, T.C., Chen, I.-C., Clark, T.D., Colwell, R.K., Danielsen, F., Evengård, B., Falconi, L., Ferrier, S., Frusher, S., Garcia, R.A., Griffis, R.B., Hobday, A.J., Janion-Scheepers, C., Jarzyna, M.A., Jennings, S., Lenoir, J., Linnetved, H.I., Martin, V.Y., McCormack, P.C., McDonald, J., Mitchell, N.J., Mustonen, T., Pandolfi, J.M., Pettorelli, N., Popova, E., Robinson, S.A., Scheffers, B.R., Shaw, J.D., Sorte, C.J.B., Strugnell, J.M., Sunday, J.M., Tuanmu, M.-N., Vergés, A., Villanueva, C., Wernberg, T., Wapstra, E., and Williams, S.E., 2017. Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. *Science*, 355 (6332), eaai9214.

Pedreschi, D., García-Rodríguez, O., Yannic, G., Cantarello, E., Diaz, A., Golicher, D., Korstjens, A.H., Heckel, G., Searle, J.B., Gillingham, P., Hardouin, E.A., and Stewart, J.R., 2019. Challenging the European southern refugium hypothesis: Species-specific structures versus general patterns

of genetic diversity and differentiation among small mammals. *Global Ecology and Biogeography*, 28 (2), 262–274.

Pelini, S.L., Dzurisin, J.D.K., Prior, K.M., Williams, C.M., Marsico, T.D., Sinclair, B.J., and Hellmann, J.J., 2009. Translocation experiments with butterflies reveal limits to enhancement of poleward populations under climate change. *Proceedings of the National Academy of Sciences*, 106 (27), 11160–11165.

Peterson, M.L., Doak, D.F., and Morris, W.F., 2018. Incorporating local adaptation into forecasts of species' distribution and abundance under climate change. *Global Change Biology*, 25 (3), 775–793.

Pinsky, M.L., Worm, B., Fogarty, M.J., Sarmiento, J.L., and Levin, S.A., 2013. Marine Taxa Track Local Climate Velocities. *Science*, 341 (6151), 1239–1242.

Poli, P., Lenoir, J., Plantard, O., Ehrmann, S., Røed, K.H., Leinaas, H.P., Panning, M., and Guiller, A., 2020. Strong genetic structure among populations of the tick Ixodes ricinus across its range. *Ticks and Tick-borne Diseases*, 11 (6), 101509.

Porretta, D., Mastrantonio, V., Mona, S., Epis, S., Montagna, M., Sassera, D., Bandi, C., and Urbanelli, S., 2013. The integration of multiple independent data reveals an unusual response to Pleistocene climatic changes in the hard tick *Ixodes ricinus*. *Molecular Ecology*, 22 (6), 1666–1682.

Portillo, A., Santibáñez, P., Palomar, A.M., Santibáñez, S., and Oteo, J.A., 2018. ' Candidatus Neoehrlichia mikurensis' in Europe. *New Microbes and New Infections*, 22, 30–36.

Pritchard, J.K., Stephens, M., and Donnelly, P., 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, (155), 945–959.

Privé, F., Luu, K., Vilhjálmsson, B.J., and Blum, M.G.B., 2020. Performing Highly Efficient Genome Scans for Local Adaptation with R Package pcadapt Version 4. *Molecular Biology and Evolution*, 37 (7), 2153–2154.

Quinzin, M.C., Normand, S., Dellicour, S., Svenning, J.-C., and Mardulyn, P., 2017. Glacial survival of trophically linked boreal species in northern Europe. *Proceedings of the Royal Society B: Biological Sciences*, 284 (1856), 20162799.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Redbo-Torstensson, P. and Berg, H., 1995. Seasonal cleistogamy: a conditional strategy to provide reproductive assurance. *Acta Botanica Neerlandica*, 44 (3), 247–256.

Rieseberg, L.H., Widmer, A., Arntz, A.M., and Burke, J.M., 2002. Directional selection is the primary cause of phenotypic diversification. *Proceedings of the National Academy of Sciences*, 99 (19), 12242–12245.

Roces-Díaz, J.V., Jiménez-Alfaro, B., Chytrý, M., Díaz-Varela, E.R., and Álvarez-Álvarez, P., 2018. Glacial refugia and mid-Holocene expansion delineate the current distribution of Castanea sativa in Europe. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 491, 152–160.

Robert J. Hijmans (2019). geosphere: Spherical Trigonometry. R package version 1.5-10. https://CRAN.R-project.org/package=geosphere

Rosselló, J.A., Lázaro, A., Cosín, R., and Molins, A., 2007. A Phylogeographic Split in Buxus balearica (Buxaceae) as Evidenced by Nuclear Ribosomal Markers: When ITS Paralogues Are Welcome. *Journal of Molecular Evolution*, 64 (2), 143–157.

Rousset, F., 1997. Genetic Differentiation and Estimation of Gene Flow from FStatistics Under Isolation by Distance. *Genetics*, 145, 1219–1228.

Sánchez-del Pino, I., Alfaro, A., Andueza-Noh, R.H., Mora-Olivo, A., Chávez-Pesqueira, M., Ibarra-Morales, A., Moore, M.J., and Flores-Olvera, H., 2020. High phylogeographic and genetic diversity of *Tidestromia lanuginosa* supports full-glacial refugia for arid-adapted plants in southern and central Coahuila, Mexico. *American Journal of Botany*, 107 (9), 1296–1308.

Sandoval-Castillo, J., Gates, K., Brauer, C.J., Smith, S., Bernatchez, L., and Beheregaray, L.B., 2020. Adaptation of plasticity to projected maximum temperatures and across climatically defined bioregions. *Proceedings of the National Academy of Sciences*, 117 (29), 17112–17121.

Schmidt, T., Arens, P., Smulders, M.J.M., Billeter, R., Liira, J., Augenstein, I., and Durka, W., 2009. Effects of landscape structure on genetic diversity of Geum urbanum L. populations in agricultural landscapes. *Flora - Morphology, Distribution, Functional Ecology of Plants*, 204 (7), 549–559.

Schmitt, T., 2007. Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in Zoology*, 4 (1), 11.

Schmitt, T. and Varga, Z., 2012. Extra-Mediterranean refugia: The rule and not the exception? *Frontiers in Zoology*, 9 (1), 22.

Schulte, U., Hochkirch, A., Lötters, S., Rödder, D., Schweiger, S., Weimann, T., and Veith, M., 2012. Cryptic niche conservatism among evolutionary lineages of an invasive lizard: Intraspecific niche conservatism. *Global Ecology and Biogeography*, 21 (2), 198–211.

Shaw, J., Lickey, E.B., Beck, J.T., Farmer, S.B., Liu, W., Miller, J., Siripun, K.C., Winder, C.T., Schilling, E.E., and Small, R.L., 2005. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany*, 92 (1), 142–166.

Shaw, J., Lickey, E.B., Schilling, E.E., and Small, R.L., 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American journal of botany*, 94 (3), 275–288.

Slatkin, M., 1993. ISOLATION BY DISTANCE IN EQUILIBRIUM AND NON-EQUILIBRIUM POPULATIONS. *Evolution*, 47 (1), 264–279.

Slatkin, M., 1996. In Defense of Founder-Flush Theories of Speciation. *The American Naturalist*, 147 (4), 493–505.

Smith, A.B., Godsoe, W., Rodríguez-Sánchez, F., Wang, H.-H., and Warren, D., 2018. Niche Estimation Above and Below the Species Level. *Trends in Ecology & Evolution*, 34 (3), 260–273.

Sokal, R.R. and Rohlf, F.J., 1962. THE COMPARISON OF DENDROGRAMS BY OBJECTIVE METHODS. *TAXON*, 11 (2), 33–40.

Stewart, J.R. and Lister, A.M., 2001. Cryptic northern refugia and the origins of the modern biota. *Trends in Ecology & Evolution*, 16 (11), 608–613.

Stockwell, D.R.B. and Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, 148 (1), 1–13.

Svenning, J.-C., Normand, S., and Kageyama, M., 2008. Glacial refugia of temperate trees in Europe: insights from species distribution modelling. *Journal of Ecology*, 96 (6), 1117–1127.

Svenning, J.-C., Fløjgaard, C., Marske, K.A., Nógues-Bravo, D., and Normand, S., 2011. Applications of species distribution modeling to paleobiology. *Quaternary Science Reviews*, 30 (21–22), 2930–2947.

Taberlet, P., Fumagalli, L., Wust-Saucy, A., and Cosson, J., 1998. Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology*, 7 (4), 453–464.

Taylor, K., 1997. Biological Flora of the British Isles: *Geum urbanum*. *Journal of Ecology*, 85 (5), 705–720.

Thuiller, W., Albert, C., Araújo, M.B., Berry, P.M., Cabeza, M., Guisan, A., Hickler, T., Midgley, G.F., Paterson, J., Schurr, F.M., Sykes, M.T., and Zimmermann, N.E., 2008. Predicting global change impacts on plant species' distributions: Future challenges. *Perspectives in Plant Ecology, Evolution and Systematics*, 9 (3–4), 137–152.

Tonione, M.A., Bi, K., and Tsutsui, N.D., 2020. Transcriptomic signatures of cold adaptation and heat stress in the winter ant (Prenolepis imparis). *PLOS ONE*, 15 (10), e0239558.

Tsyusko, O.V., Tuberville, T.D., Peters, M.B., Crawford, N., Hagen, C., Weller, S.G., Sakai, A.K., and Glenn, T.C., 2007. Microsatellite markers isolated from polyploid wood-sorrel Oxalis alpina (Oxalidaceae). *Molecular Ecology Notes*, 7 (6), 1284–1286.

Tzedakis, P.C., Emerson, B.C., and Hewitt, G.M., 2013. Cryptic or mystic? Glacial tree refugia in northern Europe. *Trends in Ecology & Evolution*, 28 (12), 696–704.

Václavík, T. and Meentemeyer, R.K., 2012. Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion: Equilibrium and invasive species distribution models. *Diversity and Distributions*, 18 (1), 73–83.

Vandepitte, K., Jacquemyn, H., Roldán-Ruiz, I., and Honnay, O., 2007. Landscape genetics of the self-compatible forest herb *Geum urbanum* : effects of habitat age, fragmentation and local environment: LANDSCAPE GENETICS OF A COMMON FOREST HERB. *Molecular Ecology*, 16 (19), 4171–4179.

Vial, L., Durand, P., Arnathau, C., Halos, L., Diatta, G., Trape, J.F., and Renaud, F., 2006. Molecular divergences of the Ornithodoros sonrai soft tick species, a vector of human relapsing fever in West Africa. *Microbes and Infection*, 8 (11), 2605–2611.

Virkkala, R. and Lehikoinen, A., 2017. Birds on the move in the face of climate change: High species turnover in northern Europe. *Ecology and Evolution*, 7 (20), 8201–8209.

Wasof, S., Lenoir, J., Gallet-Moron, E., Jamoneau, A., Brunet, J., Cousins, S.A.O., De Frenne, P., Diekmann, M., Hermy, M., Kolb, A., Liira, J., Verheyen, K., Wulf, M., and Decocq, G., 2013. Ecological niche shifts of understorey plants along a latitudinal gradient of temperate forests in north-western Europe: Species' realized-niche shifts across latitude. *Global Ecology and Biogeography*, 22 (10), 1130–1140.

Wasof, S., Lenoir, J., Aarrestad, P.A., Alsos, I.G., Armbruster, W.S., Austrheim, G., Bakkestuen, V., Birks, H.J.B., Bråthen, K.A., Broennimann, O., Brunet, J., Bruun, H.H., Dahlberg, C.J., Diekmann, M., Dullinger, S., Dynesius, M., Ejrnaes, R., Gégout, J.-C., Graae, B.J., Grytnes, J.-A., Guisan, A., Hylander, K., Jónsdóttir, I.S., Kapfer, J., Klanderud, K., Luoto, M., Milbau, A., Moora, M., Nygaard, B., Odland, A., Pauli, H., Ravolainen, V., Reinhardt, S., Sandvik, S.M., Schei, F.H., Speed, J.D.M., Svenning, J.-C., Thuiller, W., Tveraabak, L.U., Vandvik, V., Velle, L.G., Virtanen, R., Vittoz, P., Willner, W., Wohlgemuth, T., Zimmermann, N.E., Zobel, M., and Decocq, G., 2015. Disjunct populations of European vascular plant species keep the same climatic niches: Climatic niche of terrestrial vascular plants. *Global Ecology and Biogeography*, 24 (12), 1401–1412.

Wei, X., Savage, J.A., Riggs, C.E., and Cavender-Bares, J., 2017. An experimental test of fitness variation across a hydrologic gradient predicts willow and poplar species distributions. *Ecology*, 98 (5), 1311–1323.

Weir, B.S. and Cockerham, C.C., 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 36 (6), 1358–1370.

Weising, K. and Gardner, R.C., 1999. A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms, 42, 11.

Welinder-Olsson, C., Kjellin, E., Vaht, K., Jacobsson, S., and Wenneras, C., 2010. First Case of Human 'Candidatus Neoehrlichia mikurensis' Infection in a Febrile Patient with Chronic Lymphocytic Leukemia. *Journal of Clinical Microbiology*, 48 (5), 1956–1959.

Wetzels, R., Matzke, D., Lee, M.D., Rouder, J.N., Iverson, G.J., and Wagenmakers, E.-J., 2011. Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 *t* Tests. *Perspectives on Psychological Science*, 6 (3), 291–298.

White, T.J., Bruns, T., Lee, S., and Taylor, J., 1990. AMPLIFICATION AND DIRECT SEQUENCING OF FUNGAL RIBOSOMAL RNA GENES FOR PHYLOGENETICS. *In*: *PCR Protocols*. Elsevier, 315–322.

Wiens, J.A., Stralberg, D., Jongsomjit, D., Howell, C.A., and Snyder, M.A., 2009. Niches, models, and climate change: Assessing the assumptions and uncertainties. *Proceedings of the National Academy of Sciences*, 106 (Supplement_2), 19729–19736.

Wilfried Thuiller, Damien Georges, Robin Engler and Frank Breiner (2020). biomod2: Ensemble Platform for Species Distribution Modeling. R package version 3.4.12.

Williams, J.E. and Blois, J.L., 2018. Range shifts in response to past and future climate change: Can climate velocities and species' dispersal capabilities explain variation in mammalian range shifts? *Journal of Biogeography*, 45 (9), 2175–2189.

Wilson, G.A. and Rannala, B., 2003. Bayesian Inference of Recent Migration Rates Using Multilocus Genotypes. *Genetics*, (163), 1177–1191.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., and NCEAS Predicting Species Distributions Working Group†, 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, 14 (5), 763–773.

Woodward, W., B.G., 1987. Climate and plant distribution at global and local scales, 9.

Wren, C.D. and Burke, A., 2019. Habitat suitability and the genetic structure of human populations during the Last Glacial Maximum (LGM) in Western Europe. *PLOS ONE*, 14 (6), e0217996.

Wright, S., 1931. Evolution in mendelian populations. *Genetics*, (16), 97–158.

Wright, J.W., Davies, K.F., Lau, J.A., McCall, A.C., and McKay, J.K., 2006. EXPERIMENTAL VERIFICATION OF ECOLOGICAL NICHE MODELING IN A HETEROGENEOUS ENVIRONMENT. *Ecology*, 87 (10), 2433–2439.

Yang, Y., Zhang, J.-W., Sun, L., and Sun, H., 2017. Sageretia liuzhouensis (Rhamnaceae), a new species from Guangxi, China. *Phytotaxa*, 309 (3), 229.

Yousefi, N., Hassel, K., Flatberg, K.I., Kemppainen, P., Trucchi, E., Shaw, A.J., Kyrkjeeide, M.O., Szövényi, P., and Stenøien, H.K., 2017. Divergent evolution and niche differentiation within the common peatmoss *Sphagnum magellanicum*. *American Journal of Botany*, 104 (7), 1060–1072.

Zimmermann, N.E., Edwards, T.C., Graham, C.H., Pearman, P.B., and Svenning, J.-C., 2010. New trends in species distribution modelling. *Ecography*, 33 (6), 985–989.

Figures Index

Figure 1.1. Two complementary hypothesised refugia ("R") in Europe and post glaciation range shifts (arrows): (a) the classical (Hewitt, 1999) Mediterranean refugia *versus* (b) the northern refugia hypothesis. Modified from Schmitt & Varga (2012). 13

Figure 1.2. Schematic representation of the models from Garnier & Lewis (2016) of population differentiation in response to range shifts. Gradual changes in climate and consequent gradual expansion of the species' range (a) would not have important impacts in the gene diversity across populations on the leading edge. When climate is changing with high speed in comparison to the spreading speed of the species (b), an erosion of the genetic diversity will be observed on the leading edge (red broken line) compared to populations in the core of the ancient distribution. This effect will be more intense when coupled with a reduction of the trailing edge. 17

Figure 1.3. Schematic representation of range disjunction as a consequence of climate warming. In the present climate conditions (a), the species has a continuous range. Climate warming makes lowland and central areas unsuitable to the species and the range shift in both latitude and altitude causes a disjunction on the species range (b), possibly leading to the isolation of certain populations. Extracted from Kuhn et al. (2016). 18

Figure 1.4. The distribution range of *Geum urbanum* including its native Eurasian range (green) and the invaded zones in North America and East Asia (red). Adapted from Plants of the World online, Royal Botanic Garden, UK (http://www.plantsoftheworldonline.org/about). 24

Figure 1.5. The distribution range of *Oxalis acetosella* in its native range (Eurasia). Adapted from Plants of the World online, Royal Botanic Garden, UK (http://www.plantsoftheworldonline.org/about)..... 25

Figure 1.6. The distribution range of *Fagus sylvatica* spp *sylvatica* (yellow zones) and *Fagus sylvatica* spp *orientalis* (dark green zones). The light green dots represent isolated populations of *F. sylvatica* and blue triangles populations introduced and naturalised according to Euforgen (http://www.euforgen.org/species/fagus-sylvatica/ - Caudullo et al., 2017). 27

Figure 1.7. The distribution of the nine lineages of *Fagus sylvatica* based on isozymes. Modified from Figure 6 of Magri et al. (2006). 27

Figure 1.8. The distribution range of *Ixodes ricinus* in its native range (Eurasia). Adapted from the European Centre for Disease Prevention and Control – ECDC (January 2020). 28

Figure 1.9. Schematic representation of the thesis. In the second chapter, I will present the genetics analyses for *Oxalis acetosella*, *Ixodes ricinus*, and *Geum urbanum*. In the third chapter, I will apply species distribution models (SDMs) to better understand the observed genetic structure of the model species. In the fourth chapter, I will investigate the gain in SDMs' performances by the inclusion of genetic information. In this fourth chapter, I will compare genetically-informed SDMs against traditional (whole-species) SDMs for *I. ricinus* and *Fagus sylvatica*. Source of photos: *I. ricinus*: https://alchetron.com/Ixodes-ricinus; *G. urbanum*: https://en.wikipedia.org/wiki/Geum_urbanum; *O. acetosella* and *F. sylvatica*: personal collection from Dr. Jonathan Lenoir. 31

Figure 2.1. Sampling locations for *Oxalis acetosella* from 2017 (green triangles) and 2018 (black dots). Only IDs of the 2018's samples are shown. Differences in size of dots represent differences in the

number of individuals sampled per site during the 2018 campaign (5-22). The light brown colour in the background represents the species distribution range across the study region (Kwescience, Plants of the World online, 2020 - http://www.plantsoftheworldonline.org/). 37

Figure 2.2. Genetic variation explored in *Oxalis acetosella*. Migration of the ten microsatellite loci characterized by Weising and Gardner (1990). In (a) the ten loci were migrated in Agarose gel. In (b) detail of the locus CCMP-4 in Acrylamide gel (left) and Agarose gel (right). The CCPM-4 locus showed no variation when PCR products migrated in Acrylamide gel but seemed to have at least two alleles when PCR products migrated in Agarose gel. Each well was loaded with the PCR product for one individual from North Germany, Finland, North and South France, Czech Republic, West Romania, and South Sweden. 39

Figure 2.3. Allele call for the first run of 96 SNP loci for 95 samples of *O. acetosella* and one negative control (black dots). For a particular locus, samples may be homozygote for one allele (red or green dots) or heterozygote (blue dots). Most of the loci showed no variation across samples, with very few loci exhibiting one to three of the rare alleles. 43

Figure 2.4. Allele call for the second run of 96 SNP loci for 95 samples of *O. acetosella* and one negative control (black dots). For a particular locus, samples may be homozygote for one allele (red or green dots) or heterozygote (blue dots). Almost no loci were amplified, and the few loci that seem to have been amplified are most probably artefacts and it seems that they are homozygote for the same allele in all samples. 44

Figure 2.5. Sampling locations and IDs for *Geum. urbanum*. Differences in size of the points represents differences in the number of individuals per site (5-17). In green, the distribution range of the species, that cover all of the study area (Kwescience, Plants of the World online, 2020 - http://www.plantsoftheworldonline.org/). 46

Figure 2.6. Probabilities ln P(X|K) (points) and delta K (crosses, Evanno et al., 2005) for each value of K. 49

Figure 2.7. Probabilities of assignment to each of the two genetic clusters (K = 2) inferred from STRUCTURE for *Geum urbanum* in European populations. (a) Individual probabilities of assignment. (b) Geographic distribution. The pie charts in (b) in each population represents the overall probabilities for that population to be assigned to one of the two clusters (light green: northern cluster; light orange: southern cluster). 50

Figure 2.8. Genetic structure of *Geum urbanum* populations: population probabilities of assignment for each value of K from 3 (upper graphic) to 8 (lower graphic). Until K = 5, there is a general differentiation between the Northern populations (all populations on the left of LVA_1) and Southern populations (on the right of LVA_1). This pattern becomes less clear with K values of 6 or greater. For all values of K, populations from Czech Republic seem to form a somewhat concise group (with the exception of CZE_9), as do populations from Scandinavia, United Kingdom and Ireland. As the value of K becomes more important, the genetic similarity between the two Alpine populations (ITA_1 and AUT_1) becomes evident. 53

Figure 2.9. Biplot of the two first coordinates of the PCoA. The first coordinates differentiate Northern (right side) from Southern populations (left side) of *Geum. urbanum*. The results of the PCoA are very similar to those from STRUCTURE (Figs. 3 and 4). As for the STRUCTURE results, populations EST_1 and

CZE_9, situated at the geographic North and South of the study area, respectively, are grouped with populations from the opposite geographic zone. 54

Figure 2.10. Spatial patterns of the first principal coordinates of the PCoA based on the pairwise values of θ (a) and the expected heterozygosity calculated for each population of *G. urbanum* (b). The PCoA pattern in (a) closely reproduces the results from STRUCTURE (note the high and low values of CZE_9 and EST_1, respectively), while expected heterozygosity has no clear spatial pattern and is inversely correlated to the first PCoA coordinate (ρ = -0.4027, *p* = 0.0373). A Gabriel graph was draw to better represent the two-dimensions geographic relationship between populations. 54

Figure 3.1. Occurrences used for model calibration and validation during the 1970-2000 period for both *Ixodes ricinus* (a) and *Geum urbanum* (b). A total of 2,171 occurrences were used for *I. ricinus*, and 1,274 for G. urbanum. The points in red represent the location of the populations of *I. ricinus* and *G. urbanum* (25 and 27 populations, respectively) analysed in chapter 2..... 74

Figure 3.2. Schematic relationship of the different analyses carried out in this third chapter. Four main groups of analyses were conducted for both *Ixodes ricinus* and *Geum urbanum*: I. Candidate loci under selection were identified by correlational and regression methods that only take into account differences in allele frequencies across populations without testing potential links with bioclimatic variables or habitat suitability conditions; II. the relationship between allele frequencies and bioclimatic variables or habitat suitability values were investigated by correlation approaches (bioclimatic variables and habitat suitability) and ordinary least square regression (OLS) with habitat suitability only as the independent variable; III. the influence of habitat suitability values on the spatial structure of allele frequencies were investigated by autocorrelation analysis and regression analysis; and IV. the probabilities of assignment of populations to one of the two genetic clusters of each species (as inferred in chapter 2) will be regressed on the habitat suitability values during 1970-2000 as well as on the changes in habitat suitability values since the LGM. Green colours correspond to analysis with bioclimatic variables and habitat suitability values, and blue colours to spatial autocorrelation analysis. 76

Figure 3.3. Predicted habitat suitability values (0-1) for *Ixodes ricinus* during 1970-2000 (a) and the LGM (b). 85

Figure 3.4. Predicted habitat suitability values (0-1) for *Geum urbanum* during 1970-2000 (a) and the LGM (b). 86

Figure 3.5. Variation in the allele frequencies of two of the SNP loci significantly correlated to ΔS for *I. ricinus*. The letter after the locus name correspond to the nucleic acid base of the plotted allele, 'G' for guanine and "T" for thymine. The relation is inversed for the other allele. 88

Figure 3.6. SNP loci of *I. ricinus* potentially under selection according to Bayenv, pcadapt and the regression of allele frequencies on changes in habitat suitability from the LGM to the period 1970-2000 (a), and loci correlated to four bioclimatic variables (BIO 1, BIO 2, BIO 3, and BIO 4) according to the Bayenv approach (b)..... 89

Figure 3.7. Frequencies of the 19 alleles from the 125 bi-allelic SNPs of *I. ricinus* positive and significantly correlated with changes in habitat suitability ΔS from the LGM to the contemporary conditions (period 1970-2000) (a) and with contemporary habitat suitability (PS) in the two Northern (violet) and Southern (yellow) genetic clusters (Mann-Whitney rank test, *p* = 0.0002 (a) and *p* = 0.6165 (b)). 89

Figure 3.8. Moran's *I* values for 125 SNP loci (a) across 25 populations of *I. ricinus* in Eurasia. Grey lines in (b) show only the loci that exhibited no autocorrelation signal in the eight distance classes; blue lines in (c) show only the loci that exhibited a positive and significant autocorrelation signal in the first distance class (0-500 km); red lines in (d) show the loci that exhibited a negative and significant autocorrelation signal in the first distance class. Triangles indicate a significance at 5% ($p < 0.05$) for one particular locus in one distance class. 92

Figure 3.9. Number of significant Moran's *I* values in each of the eight distance classes based on the correlograms of allele frequencies for 125 bi-allelic SNPs across the 25 sampled populations of *I. ricinus.* Blue triangles and the blue line show the number of significantly negative Moran's *I* values while black dots and the black line show the number of significant Moran's *I* values being either positive or negative. The number of significant values of Moran's *I* increases rapidly after the sixth distance class, mostly due to significant negative values, while the first distance classes are dominated by significantly positive values. 93

Figure 3.10. Moran's *I* values for 89 microsatellites alleles (a) across 27 populations of *G. urbanum* in Europe. Grey lines in (b) show only the loci that exhibited no autocorrelation signal in the eight distance classes; blue lines in (c) show only the loci that exhibited a positive and significant autocorrelation signal in the first distance class (0-412 km); red lines in (c) show the loci that exhibited a negative and significant autocorrelation signal in the first distance class. Triangles indicate a significance at 5% ($p < 0.05$) for one particular locus in one distance class. 94

Figure 3.11. Number of significant Moran's *I* values in each of the eight distance classes based on the correlograms of allele frequencies for the 89 microsatellites investigated across the 27 sampled populations of *G. urbanum.* Blue triangles and the blue line show the number of significantly negative Moran'*I* values while black dots and the black line show the total number of significant Moran's *I* values being either positive or negative. The number of significant values of Moran's *I* reach the maximum at the second and third distance classes. 95

Figure 3.12. Dendrograms based on the Ward's hierarchical clustering method on the genetic distances (a), changes in habitat suitability distances from the LGM to the present time (b), and the present habitat suitability distances (c) between populations of *I. ricinus*. Light blue: Northern genetic cluster. Light grey: Southern genetic cluster. 98

Figure 3.13. Dendrograms based on the Ward's hierarchical clustering method on the genetic distances (a), changes in habitat suitability distances from the LGM to the present time (b), and the present habitat suitability distances (c) between populations of *G. urbanum.* Light green: Northern genetic cluster. Light orange: Southern genetic cluster. 99

Figure 4.1. Comparison of the resulting values of the six discrimination metrics used across 20 repetitions for (a) *Fagus sylvatica* and (b) *Ixodes ricinus*. Light grey: traditional species distribution model (SDM) approach. Dark grey: genetically-informed SDM approach. Significances are indicated by asterisks (***: $p < 0.001$; **: $0.001 < p < 0.01$; *: $0.01 < p < 0.05$). 124

Figure 4.2. Predictions of the probability of presence or habitat suitability for *Fagus sylatica* (a, b) and *Ixodes ricinus* (c, d) during present day climate (1970-2000) using both the traditional species distribution model (SDM) approach (left) and the genetically-informed SDM approach (right). 126

Figure 4.3. Distribution of the probabilities of presence of *Fagus sylvatica* during the Mid-Holocene period at spatial locations where fossil records of *F. sylvatica* from the Mid Holocene period have been

found for pollen (first quartile threshold, n>=4 pollen records per site and time) (a) and macrofossil (charcoal) (b) records. Light grey: traditional species distribution model (SDM) approach. Dark grey: genetically-informed SDM approach. Stars display the significance level based on a Mann-Whitney test of difference between the two SDM approaches (*, p = 0.010). 127

Figure 4.4. Linear regression relating the mean pollen abundance (log-transformed) with a threshold of 4 pollen records (first quartile) across sites co-occurring in the same grid cell as a function of the probability of occurrence or habitat suitability according to the traditional species distribution model (SDM) approach (a) and the genetically-informed SDM approach (b) for *F. sylvatica*. For results based on summary statistics other than the mean per grid cell (e.g. median or maximum abundance per grid cell), please refer to S4. Note also that these results are based on the first quartile across all pollen records as a threshold to exclude locations with very limited pollen abundance. For results using a more restrictive threshold (median value), please refer to S6. 127

Figure 4.5. Fossil records and probability distributions during the Mid Holocene period for *Fagus sylvatica*. Probability of occurrence or habitat suitability values are based on the traditional species distribution model (SDM). Yellow triangles: macrofossil records. Green asterisks: pollen records... 128

Figure 4.6. Fossil records and probability distributions during the Mid Holocene period for *Fagus sylvatica*. Probability values are based on the genetically-informed species distribution model (SDM). Yellow triangles: macrofossil records. Green asterisks: pollen records. The genetically-informed SDM approach was capable to assign higher (> 0.5) probability values than the traditional SDM approach (**Figure 4.5**), most notably at the southern edge of the species distribution (see zooming windows), coinciding with locations where fossil records occur..... 129

Figure 5.1. Schematic organisation of the contributions of coupling the information about the spatial genetic variability of a species and SDMs to understand its evolutionary history, as developed in this thesis. Population genetics and phylogeography helps to design hypothesis concerning the evolution of species, while SDMs may be applied to test those hypotheses. Conversely, the distribution of the genetic variation within a species may improve the niche projections over space and time. 139

Figure 5.2. Arial photograph of the intensively managed openfield landscape window (OT) (a) and the low managed bocage landscape window (BT) (b) investigated in the Woodnet and FORHAIE projects, showing the studied forest patches (dark green) located in North of France (orange dot). 142

Figure 5.3. Genetic analyses of *Geum urbanum* occurring across two contrasting landscape windows (OT: Openfield Thiérache; BT: Bocage Thiérache) of 25 km^2 each in the Thièrache region (Picardy, Hauts-de-France) (source: Woodnet and FORHAIE projects) (modified from Guiller, et al., unpublished). In (a) is shown the comparison of historical (dark grey) *vs* contemporary (light grey) gene flow as estimated in Migrate v4 (Beerli & Falsenstein, 2001) and BAYESASS 1·3 (Wilson & Rannala, 2003) (** p = 0.0037). In (b) the individual assignment probabilities to one of the two clusters identified in STRUCTURE (Pritchard et al., 2000). The two clusters in (b) coincides with populations (fragments) from each of the two landscape windows. 143

Table Index

Table 2-1. Basic statistics for each population of *Geum. urbanum*. N, number of samples. Ar, Allelic richness. Ho, observed heterozygosity. He, gene diversity. Fis, inbreeding coefficient. 51

Table 3-1 *(cont.)*. SNP loci of *I. ricinus* exhibiting a significant relationship between allele frequencies across 25 populations showing: the Bayes Factor (BF) from the correlation analysis with Bayenv in columns Δs, BIO 1, BIO 2, BIO 3, and BIO 4; the X^TX statistic from Bayenv for the loci showing the 5% highest values; the *p*-values from the linear regression with present habitat suitability Ps and the differences in habitat suitability between present and LGM climatic conditions Δs; and, the candidate loci under selection according to the pcadapt method. Results of the BLAST query for those loci are shown in the last four columns. The two *p*-values in bold are the results of the lagged models. 100

Table 3-2. Significant values of the simple regressions of allele frequencies of six microsatellite loci of *G. urbanum* as a function of the five bioclimatic variables used to calibrate the SDMs, present habitat suitability (PS), and changes in habitat suitability from the LGM to the present (ΔS). Values in bold are from the lagged models. 104

Table 3-3. Cophenetic correlations between dendrograms of genetic (Θ), changes in habitat suitability from the LGM to the present (ΔS), and present habitat suitability (PS) for *Ixodes ricinus* (upper triangle) and *Geum urbanum* (lower triangle)..... 105

Table 4-1. Bioclimatic Variables from WorldClim (Hijmans et al., 2005) used for building SDMs for each species. 120

Table 4-2. Discrimination metrics used to access model performance. TP: True positives. FN: False negatives. TN: True negatives. FP: False positives. p_o: proportion of agreement. p_e: expected proportion of agreement. 121

Résumé détaillé en Français

Introduction

Le climat est l'une des forces les plus importantes qui déterminent la répartition des espèces. Alors que la Terre a subi de nombreux cycles de réchauffement et de refroidissement, certaines espèces se sont éteintes, tandis que d'autres ont connu des cycles similaires d'expansion et de contraction de leur aire de répartition. Les changements d'aires de répartition passés ont eu des conséquences importantes sur l'évolution des espèces, en Europe notamment. Ainsi, la littérature scientifique rapporte de nombreux exemples de structure génétique et phylogéographique des populations actuelles fortement influencées par la dynamique post-pléistocène.

Selon le GIEC (Groupe d'experts intergouvernemental sur l'évolution du climat), la température moyenne à la surface du globe a augmenté d'environ 0,72°C au cours du XXe siècle, et cette progression devrait se poursuivre dans le XXIème siècle avec des valeurs moyennes attendues oscillant entre 1,0°C et 3,7°C selon les scénarii du RCP (Representative Concentration Pathway ou scénarios de trajectoire du forçage radiative). Cette hausse de température s'accompagnera de variations d'autres variables climatiques (telles que les précipitations annuelles moyennes) et de la perte d'habitats terrestres par l'élévation du niveau de la mer. Ces dérèglements climatiques dans une si petite fenêtre temporelle conduiront à la redistribution des espèces capables de « suivre » (c'est-à-dire migrer) de tels changements. Selon la tolérance de l'espèce aux conditions changeantes de l'environnement, elle devra se déplacer, s'adapter localement ou s'éteindre inexorablement.

Les différents scénarii de changement d'aire de répartition peuvent avoir un impact sur l'évolution des espèces. L'expansion de l'aire de répartition d'une espèce est bien connue pour modifier les fréquences des gènes dans les populations en limite d'expansion. La croissance rapide observée dans de telles populations peut avoir pour conséquence l'augmentation de la fréquence d'allèle « rares » et la fixation de nouvelles combinaisons d'allèles ou de nouvelles mutations.

Les modèles de répartition des espèces (MDS) sont largement utilisés en écologie, en biogéographie et en biologie de la conservation. L'approche générale consiste à calibrer un modèle avec les variables climatiques contemporaines censées capturer la niche des espèces et projeter leur distribution dans le temps et/ou l'espace. La comparaison des distributions projetées à différentes époques peut ainsi nous renseigner sur la façon dont l'aire de répartition d'une espèce s'est déplacée dans le passé ou se déplacera dans le futur. Même si l'adaptation locale aux conditions climatiques a déjà été montrée pour de nombreux organismes, on peut encore débattre de la manière dont l'adaptation locale peut influer sur les performances des MDS. Théoriquement, si des groupes génétiques distincts réagissent différemment aux mêmes variables climatiques, la modélisation indépendante de ces unités génétiques peut fournir des prévisions plus fiables que celles fournies par l'entité « espèce » seulement. Dans ce contexte, certaines études récentes incorporant la variation génétique intraspécifique dans les MDS ont pu estimer le rôle de cette information dans la performance des modèles. Au vu des résultats, conclusions, et perspectives contrastés obtenus après intégration de la composante génétique dans les MDS, il est clair que des recherches supplémentaires sont nécessaires. C'est l'un des principaux objectifs de ce travail de thèse.

Les modèles biologiques étudiés

Quatre espèces modèles ont été étudiées : les deux plantes forestières *Geum urbanum* (benoîte commune) et *Oxalis acetosella* (oseille des bois) ; l'arbre *Fagus sylvatica* (hêtre européen) ; et la tique *Ixodes ricinus* (tique du mouton). Il s'agit d'espèces forestières largement répandues en Europe, mais présentant des niveaux de spécialisation écologique, des capacités de dispersion et des stratégies de reproduction différents.

Objectifs

L'objectif principal de cette thèse est d'étudier certaines des intersections entre les MDS et la génétique des populations dans un contexte de changement climatique. A partir de cet objectif principal, trois autres ont été établis :

I. Déterminer la structure des populations et la phylogéographie des espèces dont la structure phylogéographique est inconnue à l'échelle européenne, soit pour trois des espèces modèles : *Oxalis acetosella, Geum urbanum,* et *Ixodes ricinus* ;
II. Étudier par quels moyens les MDS peuvent aider à comprendre la structure génétique actuelle des espèces modèles ; et,
III. Étudier l'efficacité de l'intégration de l'information génétique pour améliorer les MDS.

Étant donné que les espèces modèles retenues dans ce travail ne disposaient pas du même niveau d'information génétique, le point de départ de l'analyse a varié selon chaque espèce. Pour *Oxalis acetosella*, la première étape a consisté à identifier des marqueurs génétiques candidats nécessaires à l'analyse de la structure génétique des populations. Présents chez *Geum urbanum* et *Ixodes ricinus,* les premières étapes du travail pour ces deux espèces ont porté directement sur la structure de leurs populations. Quant à *Fagus sylvatica,* la connaissance préalable de sa structure phylogéographique (Magri et al., 2006) m'a permis d'utiliser directement la répartition géographique des lignées pour valider l'incorporation d'informations génétiques intraspécifiques dans les MDS.

Recherche de marqueurs génétiques variables pour *Oxalis acetosella*

Les échantillons de l'oseille des analysés ont été prélevés de manière à couvrir l'aire de répartition de l'espèce en Europe. L'échantillonnage a eu lieu au cours deux années consécutives, en 2017 et 2018. Après avoir réceptionné les échantillons, toutes les feuilles individuelles ont été vérifiées avant d'être stockées. En 2018, 348 individus ont ainsi été échantillonnés dans 30 populations différentes (de 5 à 22 individus par population, moyenne = 11,6).

Pour mener à bien cette analyse génétique, j'ai dans un premier temps recherché dans la littérature des loci variables d'espèces apparentées. Deux séries de loci microsatellites ont été testées (Weising & Gardner, 1990 ; Tsyusko, 2004), mais malheureusement un seul locus était polymorphe. Dans un second temps, j'ai étudié la variabilité d'une séquence nucléaire et d'une séquence chloroplastique parmi les mêmes populations décrites dans la section précédente : la région ITS (internal transcribed spacer) de la séquence nucléaire du gène de l'ARN ribosomal et la séquence chloroplastique non codante petA-psbJ. Les résultats issus de cette analyse de séquences ont été à nouveau décourageants. Aucune variation n'a effectivement été observée au locus ITS. De la même façon, a l'exception d'un échantillon du sud-ouest de la France, aucune variation n'a été enregistrée au locus petA-psbJ.

Compte tenu de l'absence de marqueurs disponibles et variables chez *O. acetosella*, j'ai intégré le projet ASSETS - '2nd phase of BASC flagship' to identify SNP bi-allelic loci – dont le bout est d'identifier des marqueur SNP dans différentes espèces. Pour la préparation de la bibliothèque pour le RAD-seq, 10 échantillons provenant de différentes populations de la campagne 2017 ont été poolés. Ce pool a ensuite été digéré avec l'enzyme de restriction PstI. Le séquençage a été réalisé par SBS (Sequencing By Synthesis), méthode qui consiste en l'incorporation séquentielle et la détection de nucléotides. Toutes les étapes du séquençage ont été réalisées sur le séquenceur NovaSeq 6000. Une étape bioinformatique préliminaire a été réalisée, consistant à filtrer les « lectures » (*reads*) de faible qualité. À l'issue de cette étape, plus de 14 000 000 lectures d'*O. acetosella* ont été conservées pour une analyse plus approfondie. L'assemblage *de novo* a été effectué dans *STACKS* et les SNP candidats ont été filtrés selon le script bash/Python de Delord et al. (2017), ce qui a permis de sélectionner 1224 séquences bi-alléliques. Toutes les séquences dont le polymorphisme était situé à moins de 52 bases des extrémités ont été exclues de l'analyse. Après un BLAST permettant d'estimer le pourcentage similitudes entre les séquences candidats et la bibliothèque de la GenBank, toute les séquences dont la couverture de recherche était de 100 % ont également été exclues. Sur les 524 séquences restantes,

192 ont été sélectionnées au hasard pour être validées. Ces 192 SNPs ont permis de génotyper les individus des populations échantillonnés en 2018. Contre toute attente, les résultats ont là encore été très décevants. En effet, la plupart des loci n'ont montré aucune variation ou n'ont même pas pu être amplifiés. Malgré tout, ces SNP sont actuellement en cours de validation sur la base d'un petit nombre d'individus sélectionnés pour leur meilleure qualité d'ADN ; ces résultats ne sont pour l'heure pas encore disponibles.

Structure génétique des populations du *Geum urbanum*

Échantillonnage

Les échantillons analysés ont été recueillis en 2018 sur l'ensemble de l'aire de répartition de l'espèce en Europe. Tous les individus échantillonnés ont été prélevés à au moins cinq mètres les uns des autres. Tous les lieux échantillonnés ont été géoréférencés à l'aide d'un GPS. Les feuilles de chaque échantillon ont été stockées dans des enveloppes en papier individuelles (une enveloppe pour chaque groupe de feuilles d'un individu) et séchées à l'air libre.

Génotypage des microsatellites

Six loci microsatellites ont été génotypés en deux multiplex, avec des colorants fluorescents ajoutés à l'amorce avant : (i) WGU2-28 (HEX), WGU6-5 (FAM), et WGU6-7 (NED) ; (ii) WGU2-10 (FAM), WGU2-48 (HEX), et WGU6-1 (NED) (Arens et al., 2004). Le génotypage des microsatellites a été effectuée de manière semi-automatisée à l'aide de GENEMAPPER v4.0. Tous les échantillons présentant des tailles de pic ambiguës ou sans amplification pour au moins un loci ont été exclus des analyses. Le jeu de données final était composé de 302 échantillons issus de 27 populations Analyse de regroupement et différenciation de la population.

Une analyse bayésienne a été effectuée avec le logiciel STRUCTURE, en utilisant le paramètre K (c'est-à-dire le nombre optimal de groupes génétiques) variant de 1 à 15. Pour chaque population,

le nombre d'allèles, la richesse allélique (Ar), l'hétérozygotie observée (Ho), la diversité génétique (He) et le coefficient de consanguinité (Fis) ont été estimés. L'isolement par distance (IBD) a été testé au niveau continental et à l'intérieur de chaque groupe génétique principal identifié par STRUCTURE par la corrélation entre les valeurs de F_{ST} et la distance génétique entre paires de population. Les patrons spatiaux de différenciation des populations ont également été caractérisés au moyen d'une analyse de la variation spatiale de l'hétérozygotie attendue, ainsi que par une analyse des coordonnées principales (PcoA) sur une matrice de distance génétique estimée par paire de populations (

Résultats et discussion

La méthode « delta K » d'Evanno et les probabilités ln P(X|K) suggèrent deux groupes génétiques distincts. Sur la base des probabilités d'assignation pour K = 2, les résultats de STRUCTURE montrent une plus grande proximité génétique entre populations issues de mêmes régions géographiques, notamment entre celles du groupe « Sud » - composé de la plupart des populations de la République tchèque, de l'Autriche, de l'Italie et du sud de la France ; et entre celles du groupe « Nord », composé de toutes les populations scandinaves, du Royaume-Uni et d'Irlande. Ce patron de différenciation générale entre les populations du Nord et du Sud a également été observé sur le premier plan principal de la PCoA. Les IBD ont été observés dans l'analyse globale de toutes les populations, mais aucun IBD n'a été obtenu au sein de chacun de ces groupes génétiques Nord *vs* Sud.

Bien qu'étant une espèce généraliste avec une grande capacité de dispersion, *G. urbanum* présente une structure de populations marquée dans toute son aire de répartition européenne. En l'absence s de barrière géographique physique évidente entre les groupes Nord et Sud, la structure génétique continentale en deux groupes distincts pourrait s'expliquer par une expansion post-pléistocène vers le nord. La théorie de la chasse aux fondateurs (« Founder-flush theory », Slatkin, 1996) préconise qu'une population fondatrice en croissance rapide connaîtrait une sélection relâchée

en raison d'une pression écologique réduite, et donc que la dérive et la recombinaison génétiques pourraient fixer des allèles différents ou une nouvelle combinaison d'allèles de la population d'origine.

Génétique des populations de la tique *Ixodes ricinus*

Échantillonnage

Au total, 28 populations de tiques de 20 pays différents ont été échantillonnées, couvrant la majeure partie de l'aire de répartition de l'espèce, y compris des populations proches de la limite nord (Norvège, Suède, Irlande et Angleterre) et sud (Iran, Espagne et Afrique du Nord) de l'aire de répartition de *I. ricinus*.

Génotypage SNP et Contrôle de qualité

J'ai génotypé 192 SNPs comme décrit par Quillery et al. (2014). Tous les échantillons ont été amplifiés par amplification du génome entier (WGA) avant le génotypage. Le génotypage a été effectué par un système Biomark HD (Fluidigm) et des tests KASPar. Tous les SNP invariants ont été supprimés. En outre, tous les individus et loci présentant plus de 20 % de sites non amplifiés (données manquantes) ont été éliminés de l'analse. Après contrôle de qualité, 125 loci SNP et 497 individus ont été conservés pour des analyses ultérieures.

Structure génétique

Deux méthodes complémentaires de clustering ont été utilisées pour caractériser la structure génétique des populations d'*I. ricinus*, en l'occurrence la méthode DAPC (*Analyse Discriminante* des Composantes Principales) et l'approche d'inférence bayésienne implémentée dans STRUCTURE. Dans ce cas, le paramètre K, c'est-à-dire le nombre optimal de clusters, variait de 1 à 10 selon les résultats issus de la DAPC. Les paramètres de l'analyse STRUCTURE étaient les mêmes que ceux décrits pour *G. urbanum*.

Résultats et Discussion

L'analyse DAPC a identifié deux niveaux de structuration génétique, l'un suggérant l'existence de trois groupes génétiques différents, l'autre identifiant quatre groupes génétiques. La typologie en 4 clusters montre deux groupes se superposant, tandis que celle en 3 clusters montrent 3 groupes de génotypes multilocus bien distincts. Nous avons donc décidé de fixer le nombre de groupes à K = 3 avec l'approche DAPC. L'analyse bayésienne effectuée avec STRUCTURE a également identifié K = 3 groupes génétiquement homogènes dont les compositions sont très similaires aux trois groupes issus de l'approche DAPC. Dans les deux analyses, les populations nord-africaines et eurasiennes sont très différenciées.

Les résultats convergents des analyses DAPC et STRUCTURE indiquent avec beaucoup de certitude que : (i) tous les échantillons d'Afrique du Nord appartiennent à la même espèce et ont la même ascendance ; (ii) aucun échantillon d'Eurasie ne partage d'ascendance avec ceux d'Afrique du Nord. Bien qu'étant un ectoparasite généraliste, nos résultats mettent en évidence des populations géographiquement distinctes et génétiquement structurées chez *I. ricinus*. Des recherches supplémentaires sur la préférence de l'hôte et la capacité de dispersion sont nécessaires pour mieux comprendre ces patrons génétiques. La différenciation des populations eurasiennes en deux groupes géographiquement distincts (Europe du Nord et Eurasie du Sud) pourrait avoir des implications importantes pour la redistribution d'*I. ricinus* en réponse au changement climatique anthropique.

Vers une meilleure compréhension de la structure génétique des populations : contribution des modèles de distribution des espèces

Projections actuelles et passées de *I. ricinus*

La distribution potentielle d'*I. ricinus* a été calibrée dans des conditions bioclimatiques au cours de la période 1970-2000, en fonction de cinq variables bioclimatiques non colinéaires, à savoir : la

température moyenne annuelle ; l'amplitude thermique diurne moyenne ; l'isothermie ; la saisonnalité de la température ; et les précipitations annuelles moyennes. Toutes les variables bioclimatiques ont été téléchargées à partir de la base de données WorldClim 2 (Fick & Hijmans, 2017). Les données sur la présence de *I. ricinus* ont été extraites de Poli et al. (2020), GBIF, et VectorMap.

Projections actuelles et passées de *G. urbanum*

L'approche générale pour la modélisation de la distribution potentielle actuelle et passée de *G. urbanum* était la même que celle décrite pour *I. ricinus*, avec des différences dans les variables bioclimatiques conservées et la sélection des occurrences. Cinq variables bioclimatiques ont été maintenues : la température moyenne annuelle, la saisonnalité de la température, la saisonnalité des précipitations, les précipitations du trimestre le plus humide et les précipitations du trimestre le plus sec.

Changements dans l'adéquation de l'habitat, les loci sous sélection et la structure génétique

Pour chaque espèce, les valeurs d'adéquation de l'habitat pendant la période 1970-2000 (P_S) et le LGM (LGM_S) ont été extraites pour toutes les cellules de la grille où les populations des espèces focales se trouvaient dans le présent. De ces deux probabilités d'adéquation, il a été possible d'extraire les différences entre les deux périodes $\Delta_S = P_S - LGM_S$. Des valeurs positives de Δ_S indiquent des conditions d'adéquation de l'habitat plus favorables depuis le LGM jusqu'à la période 1970-2000, tandis que des valeurs négatives signent un habitat moins favorable pendant la période contemporaine de 1970-2000 qu'il ne l'était au cours du LGM.

J'ai testé l'influence de Δ_S sur la structure génétique par un modèle de régression linéaire avec les probabilités d'assignation comme variable dépendante et Δ_S comme variable indépendante. J'ai également identifié des loci sous sélection en utilisant une combinaison des méthodes - Bayenv, pcadapt et Bayescan.

Résultats et Discussion

Pour les deux espèces étudiées (*G. urbanum* & *I. ricinus*), les performances du modèle ont été classées de bonnes à excellentes. Pour les deux espèces, les MDS ont montré un changement dans l'adéquation de l'habitat, passant d'une répartition sud-ouest pendant le LGM à une répartition centre-nord dans les conditions actuelles.

Six loci identifiés comme étant potentiellement sélectionnés montrent des allèles variant de façon significative avec Δ_S. Ces allèles corrélés avec Δ_S étaient en outre plus fréquents dans le cluster nord. Enfin, les probabilités d'assignation de la population variaient significativement avec Δ_S pour *I. ricinu*s.

L'application des modèles de distribution des espèces a permis de mieux comprendre la structure génétique spatiale de deux espèces forestières largement répandues en Europe, la tique du mouton *Ixodes ricinus* et la benoîte commune *Geum urbanum*. Outre les différences dans la biologie des deux espèces et le type de marqueurs génétiques utilisés pour démêler leur structure génétique, les résultats obtenus suggèrent que la structure génétique continentale des deux espèces est probablement très influencée par l'expansion post-LGM de ses aires de répartition, comme le prévoit ma première hypothèse. L'adéquation actuelle de l'habitat semble avoir beaucoup moins d'importance pour la différenciation de la population de l'espèce modèle. À l'échelle continentale, les résultats des analyses génétiques et spatiales suggèrent fortement que l'expansion de l'aire de répartition des refuges du sud de l'Europe est la principale cause de la structure actuelle observée entre les groupes Nord et Sud. À l'échelle régionale et à l'échelle locale, d'autres facteurs non pris en compte ici peuvent influencer la différenciation des populations. Enfin, la qualité des informations fournies par l'approche présentée semble dépendre de la biologie de l'espèce, mais aussi du type de marqueurs génétiques utilisés dans les analyses de génétique des populations.

Comment l'intégration de l'information génétique améliore-t-elle les modèles de distribution des espèces ?

Dans cette partie, j'ai tenté de savoir si le fait d'informer les MDS de données génétiques (phylogéographie ou structure génétique des populations) pouvait ou non améliorer les performances des modèles par rapport aux approches traditionnelles, en me concentrant tout particulièrement sur *Fagus sylvatica* et *Ixodes ricinus*. Les données sur la présence actuelle de *F. sylvatica* ont été extraites de Magri et al. (2006) et de l'ensemble de données EU-Forest (Mauri et al., 2017), tandis que les données sur la présence actuelle d'*I. ricinus* ont été extraites de Poli et al. (2020), GBIF et VectorMap. Deux "espèces entières" ont été créées, une pour chaque espèce, correspondant à la combinaison des occurrences de toutes les entités intraspécifiques appartenant à une espèce donnée et de toutes les absences (*F. sylvatica*) ou pseudo-absences (*I. ricinus*) disponibles. Ces deux ensembles de données ont été utilisés pour construire les MDS traditionnels, sans intégrer aucune information au niveau intraspécifique, qu'il s'agisse de lignées ou de groupes génétiques. Au total, quatre et trois ensembles de données de présence-absence, avec une prévalence de 0,5 à chaque fois, ont été construits pour *F. sylvatica* (lignée n°1, lignée n°245, lignée n°789 et ensemble de données de l'espèce) et *I. ricinus* (groupe sud, groupe nord et ensemble de données de l'espèce), respectivement.

Sur la base des modèles à l'échelle de l'espèce que nous avons calibrés pour *F. sylvatica* dans le contexte du climat actuel, nous avons fait une estimation rétrospective de la distribution potentielle de *F. sylvatica* pendant la période de l'Holocène moyen (environ 6 000 ybp) pour chacune des trois lignées séparément ainsi qu'à l'échelle de l'espèce. Pour évaluer et comparer les performances des modèles MDS traditionnels et génétiquement informés, lesquels permettent de projeter rétrospectivement la répartition passée de l'espèce, nous avons extrait toutes les probabilités d'occurrence prédites par les modèles dans toutes les cellules de la grille qui contiennent des enregistrements de pollen ou des macrofossiles.

Résultats et Discussion

Pour les deux espèces (*F. sylvatica* & *I. ricinus*) et les deux approches de modélisation (MDS traditionnelle *vs*. MDS génétiquement modifiée), les performances prédictives dans le climat actuel ont été classées de bonnes à excellentes. Les probabilités d'occurrence prévues au cours de la période de l'Holocène aux endroits où les données polliniques ont été recueillies étaient plus élevées pour l'approche MDS génétiquement informée que pour l'approche traditionnelle, à la fois pour le premier quartile et le seuil médian (Mann-Whitney, $p = 0,010$ et $p = 0,004$ respectivement). J'ai également constaté une relation plus forte entre l'abondance du pollen (log-transformé) et la probabilité d'occurrence de *F. sylvatica* au cours de la période de l'Holocène moyen avec l'approche MDS génétiquement modifiée (R^2 allant de 0,36 à 0,42) comparativement à l'approche traditionnelle (R^2 allant de 0,21 à 0,26). D'après les résultats, l'approche MDS génétiquement informée produit des scénarios futurs plus fiables de redistribution de biodiversité, elle permet en outre d'augmenter les capacités du modèle à détecter les zones potentiellement appropriées, ce qui signifie également un potentiel de réduction des erreurs d'omission (c'est-à-dire de faux négatif : prévoir une future absence là où elle sera une présence).

Conclusion

Les modèles de distribution des espèces projettent dans l'espace géographique certaines des dimensions de l'espace écologique (c'est-à-dire la niche) de tout niveau d'organisation biologique (population, espèces, clades, communautés, etc.). La combinaison des MDS avec la génétique des populations et la phylogéographie aide à comprendre l'évolution de ces différents niveaux taxonomiqudes, en permettant une corrélation à la fois quantitative et qualitative des changements de l'adéquation de l'habitat aux changements de la structure génétique du groupe focal. En même temps, informer les MDS de la structure génétique des espèces peut, dans certains cas, aider à améliorer la performance du modèle ou l'identification d'un habitat cryptique approprié. Des

recherches supplémentaires sont néanmoins nécessaires pour déterminer exactement comment et quand incorporer ces informations génétiques dans les modèles.

Remerciements

Cette thèse a été financé par le Fond Européen pour le Développement Régional et par la Région Hauts-de-France.

Appendix 1: Supplementary Information for Poli et al., 2020. *Ticks and Tick-Borne Dis.*

Pedro Poli, Jonathan Lenoir, Olivier Dr. Plantard, Steffen Ehrmann, Knut H. Røed, Hans Petter Leinaas, Marcus Panning, Annie Guiller

Materials and Methods

Table S1. Sample coordinates. The Reference column indicates from which source samples were made available. PC = Personal collection.

Sample locality	Code	Longitude	Latitude	Number of samples	Sample Date	Reference
Morocco	MAR	4221933.21	1519759.51	10	Before 2010	Dr. Plantard, PC
Algeria	DZA	4165854.78	1520079.18	8	Before 2010	Dr. Plantard, PC
Tunisia	TUN	4287083.09	1370080.62	13	Before 2010	Dr. Plantard, PC
Spain	ESP	3292343.37	2302053.84	19	Before 2010	Dr. Plantard, PC
Iran	IRA	7920535.19	2511813.36	13	Before 2010	Dr. Plantard, PC
Turkey Istambul	TUR	5907775.11	2200447.26	9	Before 2010	Dr. Plantard, PC
North France	FRA-N	3872010.67	2994279.45	40	2013	Erhmann et al., 2018
West France	FRA-W	3465235.38	2853298.78	15	2016	Dr. Degeilh, PC
South France	FRA-S	3593881.21	2296634.56	17	2013	Erhmann et al., 2018
Ireland	IRL	3013710.61	3385835.15	20	Before 2010	Dr. Plantard, PC
England Blue Pool	GBR-BP	3470079.25	3130233.33	19	Before 2010	Dr. Plantard, PC
England Bristol	GBR-BR	3450947.31	3224484.53	19	Before 2010	Dr. Plantard, PC
Italy Domodossola	ITA-D	4188665.99	2556599.15	11	Before 2010	Dr. Plantard, PC
Italy Varese	ITA-V	4229419.76	2523525.45	10	Before 2010	Dr. Plantard, PC
Romania	ROU	5643875.12	2813096.13	9	Before 2010	Dr. Plantard, PC
Hungary	HUN	5064737.95	2796444.23	18	Before 2010	Dr. Plantard, PC

Sample locality	Code	Longitude	Latitude	Number of samples	Sample Date	Reference
Slovakia	SVK	5008087.64	2900574.08	13	Before 2010	Dr. Plantard, PC
Moldavia	MDA	5711169.6	2856440.17	10	Before 2010	Dr. Plantard, PC
West Germany	DEU-W	4257417.83	3352915.67	24	2013	Erhmann et al., 2018
East Germany	DEU-E	4462732.5	3348531.08	38	2013	Erhmann et al., 2018
South Gemany	DEU-S	4440340.3	2784710.43	14	2013	Dr. Plantard, PC
Belgium	BEL	3924610.12	3095109.35	18	2013	Erhmann et al., 2018
North Estonia	EST-N	5186688.83	4032319.73	49	2013	Erhmann et al., 2018
South Estonia	EST-S	5313297.86	3950296.69	14	Before 2010	Dr. Plantard, PC
South Swqeen	SWE-S	4533959.53	3622513.31	20	2013	Erhmann et al., 2018
Central Sweden	SWE-C	4720133.45	4047795.89	19	2013	Erhmann et al., 2018
Norway West	NOR-W	4186225.49	3886420.36	15	2006	Dr. Leinaas, PC
Norway East	NOR-E	4389275.28	4003811.98	13	2006	Dr. Leinaas, PC

Table S2. List of SNPs, variant basis and primers used in the study (from Quillery et al., 2014)

locus	variant	Primer1-specific to allele1	Primer2-specific to allele2	Primer3-common to both alleles
1133	T/C	GCTTGGCCACTTCCACTGCTTT	GCTTGGCCACTTCCACTGCTTC	ACAACAGAGAAGGCAGCCCACA
3705	A/C	AGCATGGCGCACTGTGAAAGCTC	AGCATGGCGCACTGTGAAAGCTA	TCCTAGTCGGCTGGCTGGAG
6283	T/C	AATGAGGCGTCAGTGACAGCATAAC	AATGAGGCGTCAGTGACAGCATAAT	CGTGACGTCAAGGCAGAATGCTAT
6363	A/G	TCGTCCTCCGTCACGTAGCCG	TCGTCCTCCGTCACGTAGCCA	CCATTGAACCCTGGTGGGTCATCA
10041	A/G	GTTGTTCCCTTGGCAGACG	GTTGTTCCCTTGGCAGACA	AACATACCCGAGACTGTCAAC
19998	A/G	CAGAAGTGGAGATTGTTGCGTGTG	CAGAAGTGGAGATTGTTGCGTGTA	TACATACATTGAGCATCGACCAA
21130	C/T	GCTGCTGCAACCGGTTTATCTTC	GCTGCTGCAACCGGTTTATCTTT	AGGCACGTAGATCACGAGAATTATTTC
30736	C/G	GCTAGGTGACGAGGACTGGACG	GCTAGGTGACGAGGACTGGACC	GTTGTTCCACCTTTCGCAGGAGAT
31200	A/G	CGTTCAGGTTGACCGAGAAGTAA	GTTCAGGTTGACCGAGAAGTAG	GCCTCTCGTTACTGTCGTATC
32114	C/T	GACTAATCACCAGGAAATCCATTCTGC	GACTAATCACCAGGAAATCCATTCTGT	GGCTATACTCGGACGTATGTTGA
32551	T/C	TTCGGTGGCAACAGCTCGTCCATC	TTCGGTGGCAACAGCTCGTCCATT	CCAGCCTCATAGCCGAGCACCA
34502	G/A	CGGATTCGAACCAGTTATCAATGGG	CGGATTCGAACCAGTTATCAATGGA	GCCTCTCTAGAAAACAGTTGCTCTC
42351	A/G	CTTGTAGGAATGGAGGTCATCTTCG	CTTGTAGGAATGGAGGTCATCTTCA	CTTCTGTGTCGCAGGTGGCATCAT
57206	C/G	GCACTATGAGCCATCGAAGCCAAG	GCACTATGAGCCATCGAAGCCAAC	ACGTGACAACACTTACACGGCATTTC
60684	C/T	TGCACATAGTCGCGCAATACGTTC	TGCACATAGTCGCGCAATACGTTT	CGAGCCGTTGCAACCGATCCG
61606	G/A	ACATAGGACATCTCAAGGTCATTCG	ACATAGGACATCTCAAGGTCATTCA	GAAGAAACCGAGGATGAGTGTCATG
66390	C/T	GCCGAACAGCCGTGCAACCC	GCCGAACAGCCGTGCAACCT	TCGCTGCTGTATACCCATTG

locus	variant	Primer1-specific to allele1	Primer2-specific to allele2	Primer3-common to both alleles
68328	G/C	CAGGCAGTTTGCGGTTCACAG	CAGGCAGTTTGCGGTTCACAC	TAGAGGTTTCCCAAGTATTTATCGTA
68391	A/G	CAGCGTCAAGTTGTGGTGTT	CAGCGTCAAGTTGTGGTGTC	GCATCGCGTGACATTAGTTACA
72226	G/A	GAGGTTCCTGACATGCAGGAAACG	GAGGTTCCTGACATGCAGGAAACA	GCTCTGCAGATGCAAGTTCCAA
77668	G/A	GGAACGTCGTGACAGCCGTAG	GGAACGTCGTGACAGCCGTAA	GGATGGCTTCGAGTTGGACTACTA
78934	G/C	AAAGAAGCGTTTCCCGGTCG	AAAGAAGCGTTTCCCGGTCC	TCTGGCAAAGCAAGCACTCACC
81501	T/C	GTCCTTTCGAAGGTGTATGCATTC	GTCCTTTCGAAGGTGTATGCATTT	ACGATGCTAGTTTGTCAAATAGTG
81758	G/A	ACAAATCTGAAGCAGGCGCGAAAG	ACAAATCTGAAGCAGGCGCGAAAA	AGGACGTCGCCGAGTCGTAGAT
87199	T/C	GCTGGATTGCGTCGTCGCCT	GCTGGATTGCGTCGTCGCCC	CGGCTCTGGCCAGGACCTGATG
93695	G/A	GTCCTAGCCGCTGTCCCGTG	GTCCTAGCCGCTGTCCCGTA	CTGGGACAAACTCTTTCTCGAAGTG
96296	G/A	GCATAAGCAAACTTCAAAGCTTCCACG	GCATAAGCAAACTTCAAAGCTTCCACA	ACGAGGCGGCTCTCATGTACCA
105385	T/G	CCGCGAGCATTTTTGCCACATG	CCGCGAGCATTTTTGCCACATT	TTGACGTCACGACCTATTTGACGAA
113142	A/T	GAGCTCATAGTCCTGAAGACCACA	GAGCTCATAGTCCTGAAGACCACT	TTACGTTGGTCACTATGGGAACGCT
114791	G/C	CGCTGCTAGCAGACGGGAGG	CGCTGCTAGCAGACGGGAGC	GAGAGCGTACACGATTTGCCACGA
116335	A/C	GTGCGTCGAATGTCCAGGTTTATCC	GTGCGTCGAATGTCCAGGTTTATCA	CAAGTTGCGCAAGAGGTGGCAAA
125671	C/T	GTCTGCTTCTGCTATGCTCTGTTTC	GTCTGCTTCTGCTATGCTCTGTTTT	AGCGTCTGCTGCGGAACATCGTA
129322	T/A	CAAGGCAGCGCAGTTCTGACACT	CAAGGCAGCGCAGTTCTGACACA	ATCTGCGTAGCATAAGCCGTGCC
133049	G/A	ACGGGTCGTACAGCGACAAGAG	ACGGGTCGTACAGCGACAAGAA	CGAACATTACAAACGCCGCAAGAGG
137096	T/G	GTGAATGGCAATGCCAGAGTGTAT	GTGAATGGCAATGCCAGAGTGTAG	CTCGGTATTCTGCGGAGCACAA
143089	G/A	GGCACAGGATTTGCTGGTTATAGAGG	GGCACAGGATTTGCTGGTTATAGAGA	GGTGCTATGTGTACCTCACGCC
144259	C/T	GTTGAGTGTCGTGTCCTTCGCC	GTTGAGTGTCGTGTCCTTCGCT	AACAGCTCCTCGTAGACTGCGTAC

locus	variant	Primer1-specific to allele1	Primer2-specific to allele2	Primer3-common to both alleles
145634	C/T	CGGACGCGTGGACGTGACTC	CGGACGCGTGGACGTGACTT	TGGTGACCGTGTGTTGCGCAG
150669	T/C	TGTGCACAAGATGATTCCATAATT	TGTGCACAAGATGATTCCATAATC	GTCATCGGTGATTGTGTCAGTTTAT
155043	G/A	GAATGTGATCGTGGGAGAAGATATAGG	GAATGTGATCGTGGGAGAAGATATAGA	GCTGTGGAAGCTAAGTGCTCGTTG
159151	C/G	AGACAACGTACGCGCGATTTCAC	AGACAACGTACGCGCGATTTCAG	TGCTAACTGCCAGCGCGTGG
166766	A/G	ATCGACCGGCTGGCTGGCTA	ATCGACCGGCTGGCTGGCTG	GCCTGTTCTTCTGTAAGTCGCTCTA
167418	T/A	TGTCCGATACCTGCCTCCAATTTGTT	TGTCCGATACCTGCCTCCAATTTGTA	TTACCTCCACCGGGTGTCCCAT
175115	T/C	ATGGCAGTGTCAAGAAGGCCAAGT	ATGGCAGTGTCAAGAAGGCCAAGC	CAATGGCAGTGTCAAGGTCGATCTC
176991	C/A	AGAAGCTAGACGCAGAGTTAGGGC	AGAAGCTAGACGCAGAGTTAGGGA	AGGAAGAGTCCAATGTGTGCGCAA
180239	G/T	GTCCTGTGCTGTTGCCGCCG	GTCCTGTGCTGTTGCCGCCT	TGTTCCTGGACGCAAGTCACG
189207	T/A	TGGGCGTTGCAGTAATGCAACAGTT	TGGGCGTTGCAGTAATGCAACAGTA	TCTAAGGCTCCTGGTGTAAGCACACG
197784	C/T	GTTCATTAGAAGCTGTCAGTTGACTC	GTTCATTAGAAGCTGTCAGTTGACTT	CAGTGGCGTAACACGAGAAACTAG
198227	C/T	GACAACATCCAGGGCGAGTTCTAC	GACAACATCCAGGGCGAGTTCTAT	TTGCTATAACCAGTCTTCGACGC
205578	G/A	GATGTAGCCCCAGATATACTCAAAG	GATGTAGCCCCAGATATACTCAAAA	ACAGGTACTAAACCAATTTCGGC
207995	A/C	CGAGGTAAGATTGCCACTTATCTTTCC	CGAGGTAAGATTGCCACTTATCTTTCA	ACCACCTGCCAGTGTTCGACGAT
208593	C/G	GGTCTGGTGCCTGGAAAGTGC	GGTCTGGTGCCTGGAAAGTGG	GGACGCAGTAAACAGAGCAGTCATA
209761	C/T	ACATCATAAGTCACGTGGCCTGAC	ACATCATAAGTCACGTGGCCTGAT	ACGCCGTGACGTCTCCTGAT
210654	T/C	GTGATTCTGCTGGTGATCTTTGTGATC	GTGATTCTGCTGGTGATCTTTGTGATT	AGCACGCCCAACAAGATCAACGG
212829	C/G	GGCATCTGAACGACATCGTCCACC	GGCATCTGAACGACATCGTCCACG	CGTGTGTCAGGAATGAGAGATAATC
214684	T/C	GTAACGCCGTCACACGGTAAGAC	GTAACGCCGTCACACGGTAAGAT	CTGTCTGATCCAGGCTTTACGCAA
221603	T/C	AGTCGATCATACCTTACTGCTGTGT	AGTCGATCATACCTTACTGCTGTGC	TTCGCGAGTCCGAGTTGCACAGA

locus	variant	Primer1-specific to allele1	Primer2-specific to allele2	Primer3-common to both alleles
224277	C/A	ACAGCTAGGAGCAAAGTCCAGTTCCC	ACAGCTAGGAGCAAAGTCCAGTTCCA	CTATTCCCCTTTCGATCGAACATCGG
225377	G/A	TAAAGAGTCGCCTTGGGGAATCTGG	TAAAGAGTCGCCTTGGGGAATCTGA	CACGGACAACAACATTGAACGAG
230247	T/G	GTTTCCAGCTCGCGGTCGATT	GTTTCCAGCTCGCGGTCGATG	GACTGCGTAGAGTGCGCTTTTCAA
233961	A/C	GTCATGCATTTGACAAACTTTGTTA	GTCATGCATTTGACAAACTTTGTTC	GACACTACTAGGGCCTCAATCAA
234508	C/T	TGCTGTCGCTACGCTCGACC	TGCTGTCGCTACGCTCGACT	GAGAGCAGCTCCTGGGAGTCCTTG
236290	T/C	GATGCAATATGTTTACTGGATTCGC	GATGCAATATGTTTACTGGATTCGT	TAGAAATCGGGGCCCCAACGG
243436	T/C	CTTGTGCCTGGCGTCATCTGT	CTTGTGCCTGGCGTCATCTGC	AGGCCCGTGCTCGCTCG
251320	T/A	AGGATCACGTTATACGAAGGCAAGT	AGGATCACGTTATACGAAGGCAAGA	CAAGGATGACAGCACCGGTACGA
255757	T/G	TTCATCGGCGTATCCTTTGAGCGAT	TTCATCGGCGTATCCTTTGAGCGAG	ATGATGGCGACGTAGAGGTAGTTCA
259770	C/G	ACCCTTTTTGAAAGATGAACGTTGTC	ACCCTTTTTGAAAGATGAACGTTGTG	CGTTGCTCAAAGTCAAATGCCAGTG
281206	T/G	GACACTACTAGGGCCTCAATCAAGCAT	GACACTACTAGGGCCTCAATCAAGCAG	CAGTCATGCATTTGACAAACTTTG
283680	T/A	GGCGAAACCTTTGAAGCGTTCTTCAT	GGCGAAACCTTTGAAGCGTTCTTCAA	GACAGCGTGATGACTGTTCTTGTG
287805	T/G	CTGCCGCCTGTAATTCCCGACT	CTGCCGCCTGTAATTCCCGACG	TAGGTTCACGACACGAGGTTGATTC
292025	C/T	AACGCCGTGAAAGCCGCGAAC	AACGCCGTGAAAGCCGCGAAT	GCACACCGTACATCACCGAAGCC
296275	C/A	CTGCGTAGAGTGCGCTTTTCAAGGTC	CTGCGTAGAGTGCGCTTTTCAAGGTA	TCGTTTGGTTTCCAGCTCGCGGT
298125	A/G	TTTGTTCAGTTGTCAGAGGTGGCAGTA	TTTGTTCAGTTGTCAGAGGTGGCAGTG	CCTTGTGGCATGCTCCAGTGATTC
299627	C/T	GGTATCCGCTCGCTCGATATGTATATC	GGTATCCGCTCGCTCGATATGTATATT	CGTGTGCAGCTATCCAAAGACTCG
300752	T/G	AGATGCTGAACTGTCAGATGACGAAT	AGATGCTGAACTGTCAGATGACGAAG	ACCACTGTAGTTGTGTCTCGCTCTG
303781	C/G	CTCCAATTAGCTTCAAATGAATGTTC	CTCCAATTAGCTTCAAATGAATGTTG	CTTGGTTAGTTTCTGCTGGCGTTTTC
305888	C/A	GTTTCCTCCACGCAGAGCGAAAGA	GTTTCCTCCACGCAGAGCGAAAGC	CATGCGCTTCGCACTGTCG

locus	variant	Primer1-specific to allele1	Primer2-specific to allele2	Primer3-common to both alleles
307361	T/C	GCGGTATTTTCGGTCAGGC	GCGGTATTTTCGGTCAGGT	GACAAATGTTCGTCGTTCTCAACAG
313057	A/C	AATAGCGGCCAGCAGTTCCTCATA	AATAGCGGCCAGCAGTTCCTCATC	CGAATCCGATAGTGCCGTGAGAGA
320000	A/C	CAAATTTCGTGTTCGTCCATGGCGTGA	CAAATTTCGTGTTCGTCCATGGCGTGC	CGTGACTTGACGTGACGTGCCA
329834	T/G	TAGAAAGCCGGCCCGGATCTT	TAGAAAGCCGGCCCGGATCTG	CTTTCCCAGTTCAAGCACTCTTTTAG
333882	T/C	GCTCCTCCATGTCTTGTCGTCGTTTCT	GCTCCTCCATGTCTTGTCGTCGTTTCC	CACGGTGGCAGCGGGAA
336267	G/T	GCGTTGTCTGTACATCCGCCAT	GCGTTGTCTGTACATCCGCCAG	GAGCGCAGCGGATACTCTGTTCA
339272	A/G	CCGCACCGGCTTTTACGACA	CCGCACCGGCTTTTACGACG	TCTCGTCGCTGGAGGCGTCAT
340581	C/T	CTGAACCCAACGTTGGCTGAACT	CTGAACCCAACGTTGGCTGAACC	ACTGAGTGGTTCTAGTAACGATGGCT
356074	G/A	AAGTATGGGGGAACCCGTGTGA	AAGTATGGGGGAACCCGTGTGG	TAGGAGTTGGAACACTGCGACG
356395	G/A	CATTTGCGATAGGTCGATCACGATATG	CATTTGCGATAGGTCGATCACGATATA	CCGACTTCCGACGCATGTAAAATG
371093	A/G	AGCGATGGCGTCTACCAGCGGA	AGCGATGGCGTCTACCAGCGGG	TTCTGGACTAGCAGCGAGCGAC
374382	T/C	CATGCTTTGTCAACTTTCGAGAT	CATGCTTTGTCAACTTTCGAGAC	TTATGCTGTCAGCTGAGTCCCG
376474	T/C	AGGTGGCCACTCTGACATGGATC	AGGTGGCCACTCTGACATGGATT	TGTAGAGTGTAGATGCCAGCTTCCTC
380487	C/T	CAGCCGTTCGACGGGATC	CAGCCGTTCGACGGGATT	TCGCTCGTGTCCCTCGTGT
393248	T/G	CTGCATGTCTTGGCGTCTGATGTCTTCT	CTGCATGTCTTGGCGTCTGATGTCTTCG	GGTTCACTGGCCAAACGCTCCTCTAC
399212	A/G	GTTCAATGGGGCTTCTGCTATCA	GTTCAATGGGGCTTCTGCTATCG	GCGTGAATTCAACGTTCGCTAAG
411541	G/A	AGTCGTTGTGGGCGCGCATGGG	AGTCGTTGTGGGCGCGCATGGA	GTCAGGCTGTTCGGCTTGACGTATG
419658	T/G	TGTCCTCGTACGTGCTCGTTGTGACT	TGTCCTCGTACGTGCTCGTTGTGACG	AGCAGATGGCCTGGTAGCGGTCC
428503	G/A	CATGCAGGATACCGTGTGAGTTCAG	CATGCAGGATACCGTGTGAGTTCAA	GATGCTGTGCGCGTTGGACTG

locus	variant	Primer1-specific to allele1	Primer2-specific to allele2	Primer3-common to both alleles
438644	A/G	GCACTGCAAACACCTCTGCTCAAGTATG	GCACTGCAAACACCTCTGCTCAAGTATA	CTATGAATGCTCTTGCTAGCAGGCTTTA
441042	A/G	GAATTCCAAACGCGGTTCATAAACCACG	GAATTCCAAACGCGGTTCATAAACCACA	TCGAAGATAGTGTGCTCAATGGCGGTTA
446758	T/A	TTGTTGCGAACATAGAGTACAGAGGAGCA	TTGTTGCGAACATAGAGTACAGAGGAGCT	GCTACAACGTGGGAATTGCCGAGGA
450975	T/G	TGCGGTTACGCAGTCGAAGCTATT	TGCGGTTACGCAGTCGAAGCTATG	ATGGGCACTCAAGGTGCGCACG
465604	A/T	CCTAAACGTCTCGGCGCTAATA	CCTAAACGTCTCGGCGCTAATT	AACTAAGACCACATTCCCGACATTG
465892	G/A	CCCACTGACGAGCGTGCTGAAGA	CCCACTGACGAGCGTGCTGAAGG	CATGCTCTTTCCTGTTGTCCGGTTCA
468480	A/G	CATAACGCTGAATTATCTTCGCCGACTA	CATAACGCTGAATTATCTTCGCCGACTG	GTAAGGGGCCCACAAGCCTGG
480915	A/G	CTAATTCTCGTTCTACTGCCGCATG	CTAATTCTCGTTCTACTGCCGCATA	GGACACATCTCAGAACCAGATTG
487540	T/C	CACGGGAACGACGGGCACT	CACGGGAACGACGGGCACC	GGCACGTGAAGCTCCGAGATTTCAT
493429	A/G	TAGTGGGTTCGCTGAAGAACTACAAGAA	TAGTGGGTTCGCTGAAGAACTACAAGAG	CGCGCAGCTTTCTGAAGTAGTTGT
552113	T/A	TCATAGTTGGTTCACAGGCGACCT	TCATAGTTGGTTCACAGGCGACCA	GTTCTGGACTAAGTATGATTCGCTCCA
558063	A/G	CAGCTCCTGGGAGTCCTTGAGA	CAGCTCCTGGGAGTCCTTGAGG	AGTGGCTGCTGTCGCTACGCT
561492	T/C	ATCTTGCGACTGCTCGAT	ATCTTGCGACTGCTCGAC	TTCTCGCCCAGGAATGCCAT
580716	T/C	TCGGCGTTCAGCAGGCTTGAC	TCGGCGTTCAGCAGGCTTGAT	GCACCAGACCGCCGGCGA
583125	T/G	TGTTCTGAGGAAATGAGATGACTGTT	TGTTCTGAGGAAATGAGATGACTGTG	CAACACACGTCAACAGCAACAT
585284	T/A	GCTTCAGTTATCAGCTGTAAACCTA	GCTTCAGTTATCAGCTGTAAACCTT	TTCGGTAATGCGTGTATTACTCA
585318	A/G	GTACATCACCGAAGCCGAACAG	GTACATCACCGAAGCCGAACAA	TTAGCCGCAACGCCGTGAAA

locus	variant	Primer1-specific to allele1	Primer2-specific to allele2	Primer3-common to both alleles
589219	C/T	ATGCCGCACGTGCTTGAGGTC	ATGCCGCACGTGCTTGAGGTT	CAAGAAACGGCAACAGCGGACAATGAAC
627150	A/C	CAATACAGCGGTATTTGCACTA	CAATACAGCGGTATTTGCACTC	CAATGGAGCAGACGCATCT
751708	G/A	TTGAAGCACAGCTCTTAGAGAAGG	TTGAAGCACAGCTCTTAGAGAAGA	GACTCCGTCAGCTGGTTTATG
754496	C/G	GCCTCGGCGTCGGAACTCG	GCCTCGGCGTCGGAACTCC	TGGCTGAAACCAGGGACCTCAA
761047	G/A	CAACATGGACGTTTTCAAGATTGCCA	CAACATGGACGTTTTCAAGATTGCCG	GAGCCTCGCTCAGCACGGAA
763022	T/C	CACAAAGGGCACGATTTCCTCT	CACAAAGGGCACGATTTCCTCC	AGATGAGTCTGCCATCGTGTCT
764527	T/A	GGGCGTTGCAGTAATGCAACAGTA	GGGCGTTGCAGTAATGCAACAGTT	AAGGCTCCTGGTGTAAGCACACG
767569	A/G	AAACACACCTTGAACTCAGCCTCA	AAACACACCTTGAACTCAGCCTCG	GGACGACAGCTATCAACATTAGCC
768618	C/G	GAACAATTCAAAACCATGATTGAAACAC	GAACAATTCAAAACCATGATTGAAACAG	TACACTCCCAAGTGAGTTGATGC
771828	T/C	GATCCAAAGTGATCATGCCGATAGT	GATCCAAAGTGATCATGCCGATAGC	ATATCACAGTATCACGTCACGG
775381	A/G	TGTGCAGCTATCCAAAGACTCGG	TGTGCAGCTATCCAAAGACTCGA	ATGGTATCCGCTCGCTCGATATGT
777961	C/G	CTCAGCACAAGTGAATGTCAAG	CTCAGCACAAGTGAATGTCAAC	GGGCATTTGTAAGCATCTTATCGC
781023	G/T	GGCTCTATGTAGAACCAAAGATAAGTGAG	GGCTCTATGTAGAACCAAAGATAAGTGAT	ATTCTGCGGCTTCAACGAATCA
783090	G/A	ACCCGTACAGCAAACCACTACG	ACCCGTACAGCAAACCACTACA	CGACTGATTTCTCGCAACCCA
792422	T/C	TGCCACGGTAGTTTTGCTTAGT	TGCCACGGTAGTTTTGCTTAGC	ATGTTCCACGAGGCCCGTTG
43247	C/T	AGTAGACTTAAAGGCCACGCTCGAC	AGTAGACTTAAAGGCCACGCTCGAT	CCTTATATTCTCTGTCAGCGTAAG
84140	T/C	CAATCGAAATCGTGACCAATGGGATTC	CAATCGAAATCGTGACCAATGGGATTT	ACCAAGTGCCGCGCAAAGCAT
117944	C/T	CGAATTCGAAGGCGGAGATCCTC	CGAATTCGAAGGCGGAGATCCTT	CGGCTTGGCGAAGCGACG

locus	variant	Primer1-specific to allele1	Primer2-specific to allele2	Primer3-common to both alleles
316915	T/G	CGCTTCGCCGAGCACTCG	CGCTTCGCCGAGCACTCT	ACCGGTTGTGCTACGCGTAGGT
197588	T/G	CAAGCGCATCCCCATTCTGATCTT	CAAGCGCATCCCCATTCTGATCTG	CTTAGAAAGGCAAGACCTCCTTCA
2932	C/T	CTCCTACGAGGGGTGCCTGT	CTCCTACGAGGGGTGCCTGC	TTGTGACGTTCCTCGTGCTCCCT
112567	C/T	GCTCATGCGCATTGGAAGC	GCTCATGCGCATTGGAAGT	TTGCACGTACTACGTGCCTCTG
207179	T/C	CGCACGGAGATGGCATTCCTC	CGCACGGAGATGGCATTCCTT	ACACGATCTTCGGCGAGAACGTCA
165428	G/C	GTCCGCCACGTCGGTTCCAGAG	GTCCGCCACGTCGGTTCCAGAC	AAGCGGGGCTCTGCTTCCGCCT
109194	C/T	AGGCCCACAACTCCACTCTTC	AGGCCCACAACTCCACTCTTT	TACGGTAGCTATGTAACAGACACTA
139650	C/T	TACGACGGCACCGAGATC	TACGACGGCACCGAGATT	ATCTCCGGCGAGGCGTACA
56083	T/C	CCAGGCGCTCCTCCTCGGTC	CCAGGCGCTCCTCCTCGGTT	CGCCGGAGTTGGCCCAGGA
143860	A/G	ACAGGTACACGAACGATCGCAGAA	ACAGGTACACGAACGATCGCAGAG	TGCGTTCGTGCTTGTGTCATGT
152555	A/G	GCTCCAGGACAACCGTTTACCTCA	GCTCCAGGACAACCGTTTACCTCG	ATGGAAACATCGCTACACATGG
51899	A/G	GAGGTGTACGAGTGTCACTCGAAG	GAGGTGTACGAGTGTCACTCGAAA	GTATCTAGGAGGCTCGGGCGAAA
225801	C/T	GACTTCTGACATTTGATAGAATGCTC	GACTTCTGACATTTGATAGAATGCTT	TGCGGGTCAGCCATCTTACAAGTA
190468	G/A	TGAACGAAGCTGAGAGGCGCTATGA	TGAACGAAGCTGAGAGGCGCTATGG	TACGCCCAGACACTCTTGTTCAGT
31277	C/G	ATCATAGACCAACTCGCCTGCATC	ATCATAGACCAACTCGCCTGCATG	GATTCTGGAAGACAGCTTTTTCGC
455987	G/C	AATGTACGCGACGTACGCACAAG	AATGTACGCGACGTACGCACAAC	GGATTTCCGAGAGAAGCCATTTTCAG
27147	T/G	CGCAATTGTGACACCACTAG	GCGCAATTGTGACACCACTAT	CGGCTTTTGATACTCCCATCA
751588	G/T	CCGCATTTCTTCACTGCTGTTTGAAAG	CCGCATTTCTTCACTGCTGTTTGAAAT	TCGCAAATCCTGGCGCGGTAA
313642	T/A	GTGCAGTTGGCAATGGAGGTGA	GTGCAGTTGGCAATGGAGGTGT	CCGGACAACTGAAGGTGGTGC
182969	G/A	AAGACGCACTTGCCCTGGAAACATG	AAGACGCACTTGCCCTGGAAACATA	GGTCTGAGTCTTGGTTGTGTCGCAT

locus	variant	Primer1-specific to allele1	Primer2-specific to allele2	Primer3-common to both alleles
186625	A/G	GAGGAGCTGCGATGCAGAAGTGGTA	GAGGAGCTGCGATGCAGAAGTGGTG	ATGCTGATGACGCAACGCTGACTTC
191703	A/G	CCGCCGTCTTTGCAGCCTCA	CCGCCGTCTTTGCAGCCTCG	GGGGCCCCGATTTCTAGAAC
438440	A/G	GTTGAGCGCATGCGCAGGGAA	GTTGAGCGCATGCGCAGGGAG	ACTCCCTGACGTAGCCTTCGTAGGA
82163	T/C	TAAGGCTTCCAGGTGACTTC	CTAAGGCTTCCAGGTGACTTT	GGTGTGTTGCTTCTATATTG
788521	C/T	ACCCGAACTTTGCAGGCCAT	ACCCGAACTTTGCAGGCCAC	AATGAACGACCGAGCGAATCCAGA
233756	C/G	TCTACAAACCAGGCGGTTGTAAGC	TCTACAAACCAGGCGGTTGTAAGG	TCTGTTTGGGACTCCTTCCACCG
201653	G/A	GCAGTCATCAAACGTGATTTCGTCCG	GCAGTCATCAAACGTGATTTCGTCCA	AAATTGGAGAGATCACTTGACCCGC
259800	C/G	CGTGTGCCTCGCTGGCATC	CGTGTGCCTCGCTGGCATG	GCGCATTCCAGAGGCTTCC
370147	C/T	GACACCCTAGCAAAGCAAAGCGTTCTC	GACACCCTAGCAAAGCAAAGCGTTCTT	TTTCGTTCACGGCTCCCGCAA
153000	G/A	CCTACCTGCTTCCAACATTCTTTAGG	CCTACCTGCTTCCAACATTCTTTAGA	TGCACATTAGGTCAGAGATGCGGA
500950	A/T	CCACAACTCATCGCACCGAAGACT	CCACAACTCATCGCACCGAAGACA	AGACGATTATTCGGCTGTGACACATT
170547	C/G	GGTGAATACGCGTCGCGTGAGTC	GGTGAATACGCGTCGCGTGAGTG	GTGACCTTTGGTAGGACGGCAGC
466967	C/G	GAATATTTATGATGTGACCACGGCAAAC	GAATATTTATGATGTGACCACGGCAAAG	AACGCCCTGCCGCATAGTCC
246408	T/C	GGAAACAGTTATAACTATCTAGAACT	GGAAACAGTTATAACTATCTAGAACC	CACACCGAGAAATCAGACGTACC
5630	G/A	CAGCAAGCAGAGAACGTCGTCGATG	CAGCAAGCAGAGAACGTCGTCGATA	TTCAGGGTGAGACCGTCGGC
561563	A/G	TGAAGGATCTCGTACACAATACACAG	TGAAGGATCTCGTACACAATACACAA	CGAGTACTTCACGACCACGCA
338495	T/C	GGTTCTCGAAGCCGCGTTTC	GGTTCTCGAAGCCGCGTTTT	TCTGCAGCTGCTGTAGAGTCCTG
166887	A/G	TCGGCCGCCAGCAGCGTCA	CGGCCGCCAGCAGCGTCG	CCCGTCGGGAGCAATGCAG
766292	T/C	TGCCGAAGCTGGGTTTCGT	TGCCGAAGCTGGGTTTCGC	CTGGGCTGCTCCGAGGACTA
176206	G/T	ACTGCGATTGAAGTGCGTCCCG	ACTGCGATTGAAGTGCGTCCCT	ATCCTCTTGAAATTTGCTGCGGGTG

locus	variant	Primer1-specific to allele1	Primer2-specific to allele2	Primer3-common to both alleles
245496	T/C	TTCCAGCGTGCACCGTACC	TTCCAGCGTGCACCGTACT	GAAAATGCAATTTTTGTGAGCCT
199727	T/A	GGCTTCTTGTCTCGTTATTATCGT	GGCTTCTTGTCTCGTTATTATCGA	CAGTGCCACTTTATGTGAGTTG
524153	G/A	CTCTATCAAACGATGTGCTACTGTGA	CTCTATCAAACGATGTGCTACTGTGG	ACTAATTCATTGTAACCCATTTCACGAT
54140	A/G	GGTAGACACAATCTGCTCATAATGG	GGTAGACACAATCTGCTCATAATGA	ATGACTGTTACAATCTTTTGAATGC
18708	A/G	CTCCGCGTGTATGCGAGTGAA	CTCCGCGTGTATGCGAGTGAG	GGCGCGTATCATCCCAGAGC
546612	G/C	TTTCCCGCGCAGGCCGCTAG	TTTCCCGCGCAGGCCGCTAC	TCAAGGCCAACGGCGCGCA
523859	C/A	CTGGACCTGTGCTACCGTGAGTCC	CTGGACCTGTGCTACCGTGAGTCA	GCTCAGGATGTCGTACGCGCGG
160279	A/T	ATCAGCAGCGCACACGCTCA	ATCAGCAGCGCACACGCTCT	CGTCGACGGGCGATCGTGA
624322	A/G	TATCAGCTAAAGCCTCCTTCTCAGTCA	TATCAGCTAAAGCCTCCTTCTCAGTCG	GAACTGAAGCACCAGCGCCT
410904	C/G	GTCAGAGTAAGGATCTGCTAGATACCG	GTCAGAGTAAGGATCTGCTAGATACCC	TAAGAAGGTTGGCCCGAATTTGTGAA
71660	C/A	GAAATTAGAATGGTACCTGGATTACC	GAAATTAGAATGGTACCTGGATTACA	CCTTTGGGGTGCGCTTATGTAAT
87165	G/A	GAATCCACGTGTCAGAGCCCTGG	GAATCCACGTGTCAGAGCCCTGA	GGTTGTATTTACAACTGACTCCTCGG
61479	A/G	GGCTAATCCTGCTTCTTGGCCTT	GGCTAATCCTGCTTCTTGGCCTC	CGATCCTGAAATCGAGCAAAGCC
571455	T/A	GTTCTGCCAGCAATTCTATCACT	GTTCTGCCAGCAATTCTATCACA	GGATGGATGCAAAGTGATATTTTAG
270863	C/T	GCAATTATAGGATCTCCGTAAACTCT	GCAATTATAGGATCTCCGTAAACTCC	CCTTTTTACGGACACTCACTTTCCTG
185472	C/G	ATTCGCCAGACCACTTGGATTCTC	ATTCGCCAGACCACTTGGATTCTG	CGTTTTCAATGAGTCTTGATTCTCG
200386	T/C	GATGGAATTAGGTACGGTCATTTCAT	GATGGAATTAGGTACGGTCATTTCAC	GTTCAGCGCATACTATGACTGACAA
40367	A/G	CACATGTGGCAAGCATTCAA	CACATGTGGCAAGCATTCAG	GCAGCAACGTTTGCTTCAGA
494898	T/C	AGCGTTGCACGCCATACATTCTCT	AGCGTTGCACGCCATACATTCTCC	TCCACAGGGTCACGTGACGCA

locus	variant	Primer1-specific to allele1	Primer2-specific to allele2	Primer3-common to both alleles
14134	C/T	CATACATTCCCTGAATACCTAGAGC	CATACATTCCCTGAATACCTAGAGT	ATTAGCCAAGCGCCCCG
361495	T/G	ATAACACAGGCAGACATTGGAGGCAG	ATAACACAGGCAGACATTGGAGGCAT	GCTCACATGCATTGAAACTGATGTC

Results

Table S3. Basic statistics per locus.

Locus	Observed heterozygosity	Gene diversity	Fst	Fis
1133	0.1101	0.2324	-0.0040	0.5262
31200	0.4319	0.467	0.0629	0.075
66390	0.1269	0.291	0.0539	0.5639
129322	0.0700	0.0873	0.0061	0.1974
159151	0.2649	0.3446	0.1504	0.2312
198227	0.5741	0.4987	-0.0093	-0.1512
221603	0.2760	0.3557	0.0932	0.2243
251320	0.1778	0.3271	0.1478	0.4564
298125	0.5543	0.4636	0.0127	-0.1956
329834	0.0918	0.2123	-0.0187	0.5675
374382	0.176	0.4705	0.0317	0.626
3705	0.1782	0.4146	0.0673	0.5704
32114	0.0759	0.1787	-0.0022	0.5753
68328	0.2477	0.3736	0.0396	0.3369
93695	0.2975	0.4519	0.0335	0.3418
133049	0.1991	0.3770	0.1375	0.4718
255757	0.2072	0.2468	0.2394	0.1604
299627	0.1109	0.2400	-0.0107	0.5381
376474	0.2141	0.3006	0.0491	0.288
6283	0.3140	0.4164	0.0634	0.2459
32551	0.262	0.343	0.0092	0.2361
96296	0.2678	0.3734	0.2144	0.2829
137096	0.5322	0.4652	0.0382	-0.1438
207995	0.8246	0.45	0.0726	-0.8323

Locus	Observed heterozygosity	Gene diversity	Fst	Fis
225377	0.2191	0.3855	0.097	0.4316
259770	0.1676	0.2408	0.0235	0.3041
300752	0.4814	0.3724	0.1811	-0.2928
336267	0.2115	0.2254	0.0116	0.0618
380487	0.3168	0.4498	0.0807	0.2957
6363	0.3144	0.4011	0.1326	0.2162
34502	0.1219	0.3181	0.0031	0.6166
105385	0.3326	0.3783	0.0686	0.1208
143089	0.7837	0.4849	0.0132	-0.6163
208593	0.1799	0.2333	0.0843	0.2288
230247	0.2052	0.3581	0.223	0.427
281206	0.2277	0.3757	0.2359	0.394
303781	0.1121	0.2912	-0.027	0.6152
339272	0.1745	0.1785	0.0237	0.0222
393248	0.0722	0.1882	-0.0019	0.6163
176991	0.1966	0.3406	0.029	0.4228
144259	0.1649	0.2687	0.0127	0.3863
113142	0.2391	0.3714	0.2421	0.3563
77668	0.4137	0.4424	0.0191	0.0648
42351	0.128	0.1783	0.439	0.2821
10041	0.1733	0.4534	0.0484	0.6178
399212	0.0839	0.3201	0.1127	0.7377
340581	0.0682	0.0911	0.0379	0.2513
305888	0.1105	0.1722	0.0151	0.3583
283680	0.6714	0.4485	0.0759	-0.497
233961	0.2212	0.3625	0.2481	0.3898

Locus	Observed heterozygosity	Gene diversity	Fst	Fis
209761	0.0828	0.3481	0.0662	0.7621
180239	0.094	0.151	0.0109	0.3773
145634	0.3787	0.4585	0.018	0.1741
114791	0.2429	0.3708	0.0262	0.3449
57206	0.1859	0.3233	0.0253	0.425
19998	0.0914	0.1196	0.0199	0.2359
411541	0.227	0.3092	0.0348	0.2658
356074	0.2633	0.3343	0.0754	0.2123
307361	0.0887	0.1801	0.0333	0.5074
287805	0.0435	0.0744	0.0317	0.4154
234508	0.2516	0.3103	0.059	0.1891
210654	0.228	0.2674	0.4378	0.1475
189207	0.1252	0.3037	0.0042	0.5879
150669	0.2338	0.4764	0.0025	0.5092
116335	0.1823	0.3998	0.0415	0.544
81501	0.2806	0.4505	0.101	0.377
60684	0.2221	0.442	0.0547	0.4974
21130	0.1723	0.412	0.0976	0.5818
356395	0.3341	0.4604	0.0791	0.2745
313057	0.1085	0.388	0.0497	0.7203
292025	0.1001	0.1208	-0.0134	0.1714
236290	0.0798	0.1775	-0.0126	0.5505
212829	0.2648	0.4684	0.0637	0.4347
197784	0.3144	0.2259	0.1932	-0.392
155043	0.1033	0.1762	0.5208	0.4136
125671	0.3398	0.4652	0.0539	0.2695

Locus	Observed heterozygosity	Gene diversity	Fst	Fis
81758	0.1436	0.2886	0.1153	0.5023
61606	0.2479	0.3716	0.2166	0.3331
428503	0.3744	0.395	0.0206	0.0521
320000	0.2643	0.2579	0.1614	-0.0248
296275	0.1384	0.3744	0.2015	0.6303
243436	0.1853	0.2832	0.3311	0.3455
214684	0.2222	0.4963	-0.0018	0.5522
438644	0.1799	0.4129	0.149	0.5644
487540	0.0981	0.1404	0.288	0.3014
767569	0.1635	0.2148	0.0246	0.2389
165428	0.211	0.2398	0.0199	0.1199
191703	0.1966	0.3729	0.2489	0.4729
153000	0.0919	0.4793	0.0403	0.8082
166887	0.1921	0.3481	0.0653	0.4483
441042	0.3243	0.4491	0.0996	0.2777
84140	0.1825	0.3554	0.0587	0.4865
438440	0.0944	0.1089	0.1972	0.1325
766292	0.738	0.4745	0.0398	-0.5553
523859	0.3518	0.4397	0.1064	0.1999
270863	0.0571	0.23	0.0146	0.7516
446758	0.0327	0.1022	0.0336	0.6802
552113	0.144	0.219	0.0062	0.3423
627150	0.1918	0.4255	0.074	0.5491
117944	0.8044	0.4649	0.0642	-0.7304
139650	0.2503	0.2708	0.1388	0.0757
176206	0.3211	0.5036	-0.0175	0.3624

Locus	Observed heterozygosity	Gene diversity	Fst	Fis
185472	0.1899	0.3965	0.0348	0.5211
450975	0.2154	0.3343	0.1091	0.3557
558063	0.2358	0.2929	0.0731	0.195
751708	0.0754	0.1006	-0.026	0.2506
775381	0.0911	0.2703	-0.0258	0.6629
27147	0.1022	0.3726	0.0636	0.7257
200386	0.1562	0.4361	0.1004	0.6418
777961	0.0977	0.3904	0.1195	0.7499
754496	0.3408	0.4748	0.051	0.2822
561492	0.0813	0.3268	0.0409	0.7512
465604	0.0536	0.2293	0.0085	0.7664
410904	0.1761	0.2763	-0.0034	0.3628
199727	0.1067	0.4443	-0.0179	0.7599
751588	0.1149	0.3961	0.0719	0.7099
152555	0.1713	0.4355	0.058	0.6067
2932	0.0525	0.0604	0.0063	0.1315
781023	0.1682	0.4324	0.0954	0.6109
761047	0.1318	0.173	-0.0271	0.2378
580716	0.316	0.3822	0.0479	0.1731
465892	0.113	0.1621	-0.0162	0.3033
5630	0.7425	0.4783	0.0441	-0.5525
313642	0.1283	0.1442	0.0071	0.1102
783090	0.2976	0.3471	0.0268	0.1426
763022	0.1247	0.302	0.0606	0.5872
583125	0.534	0.4398	0.097	-0.2141
468480	0.3318	0.4527	0.0932	0.267

Locus	Observed heterozygosity	Gene diversity	Fst	Fis
14134	0.1556	0.345	0.0144	0.549
259800	0.125	0.1671	0.0074	0.2517
182969	0.2349	0.2092	0.005	-0.1228
225801	0.1255	0.254	0.3865	0.5058
792422	0.402	0.4864	-0.0019	0.1736
764527	0.1107	0.2841	0.0287	0.6103
585284	0.1384	0.2335	-0.0063	0.4072
480915	0.0849	0.2483	0.032	0.6581
338495	0.2448	0.3455	0.0157	0.2915
186625	0.5284	0.4536	0.0852	-0.1649

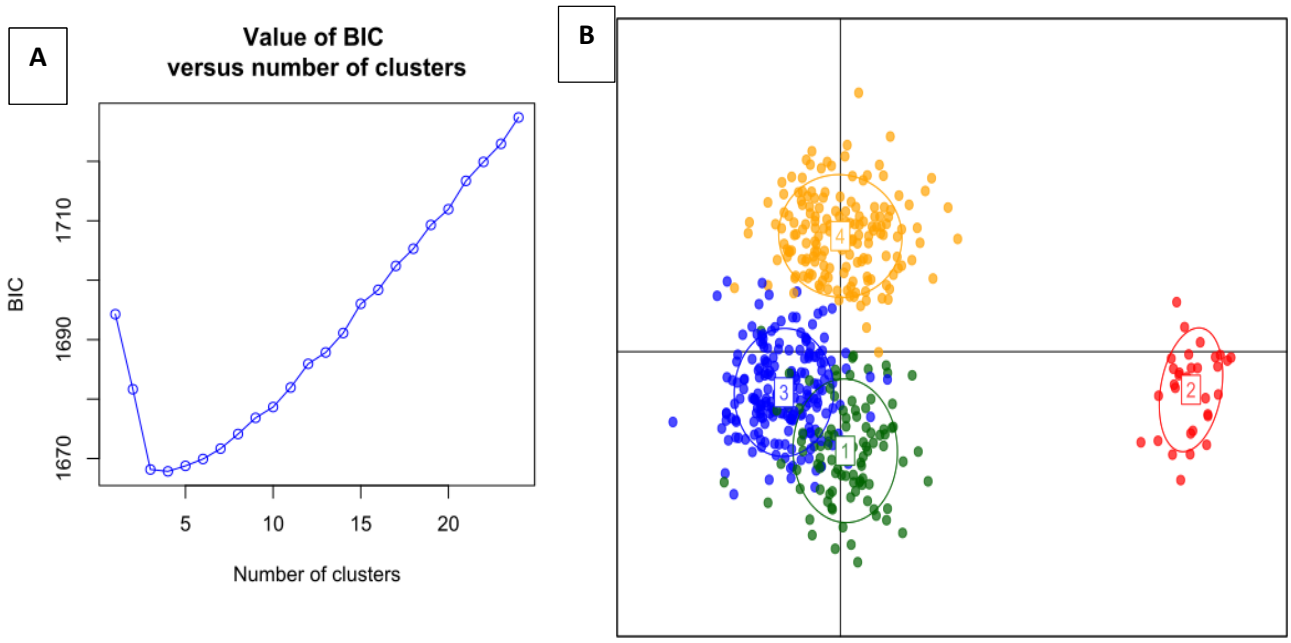


Figure S1. Discriminant analysis of principal component (DAPC) of *Ixodes ricinus* based on 497 individuals using 125 SNPs. A. BIC values as a function of the number of clusters *k*. The difference in BIC values between *k* = 3 and *k* = 4 is 0.842. B. Scatterplot of individuals on the two principal components of DAPC. The graph represents the individuals as dots and the groups as inertia ellipses. Two of the clusters overlap, while when *k* = 3 we identify 3 well separated groups (figure 3). Red : North African cluster; yellow : only individuals from southern Eurasian cluster; green : only individuals from the Northern European cluster; blue: admixture cluster with mainly individuals from the northern European cluster in fugure 3.

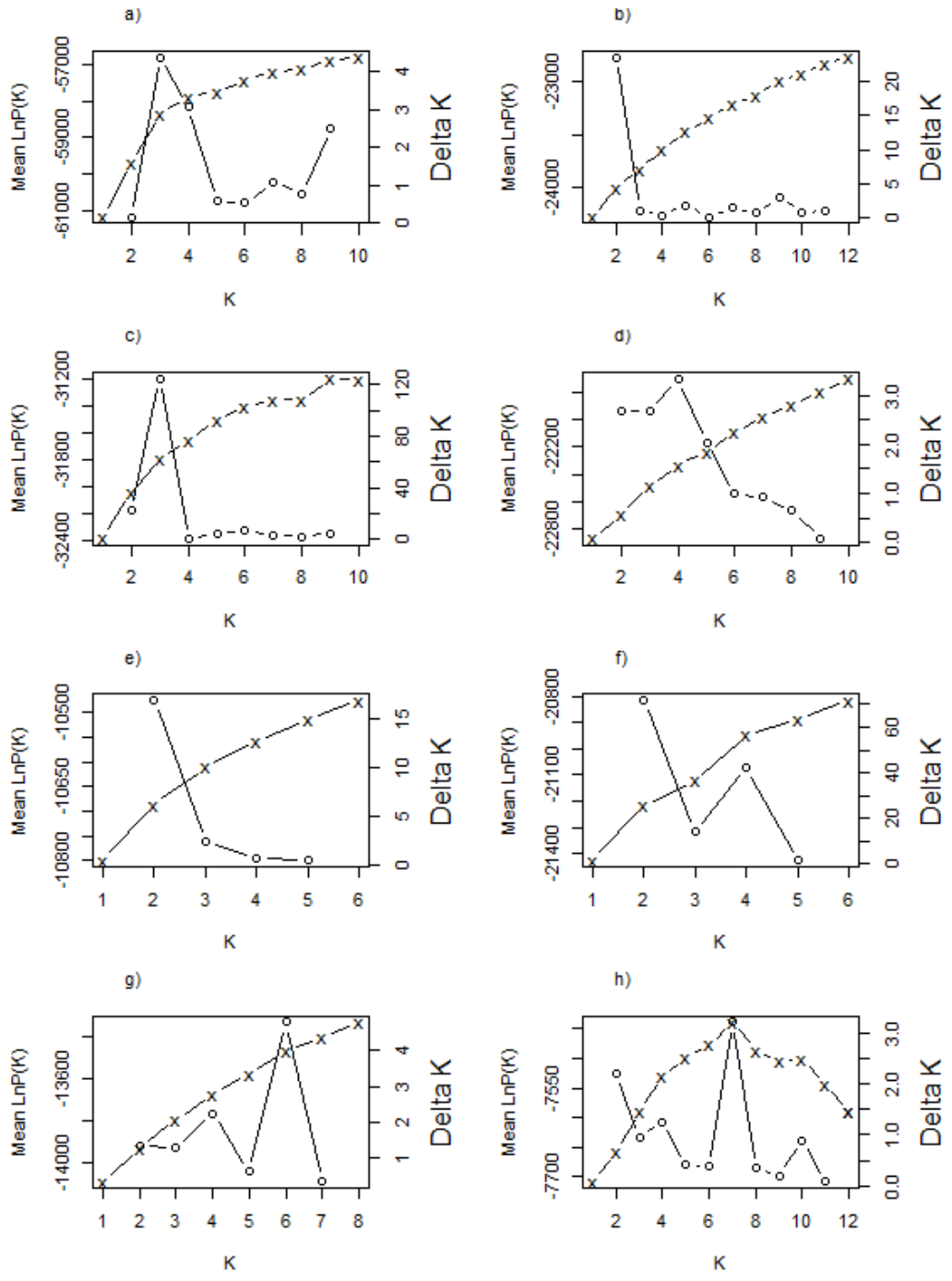


Figure S2. Probabilities ln P(*X*|*K*) for each level of hierarchical analysis. First round of analysis: a); Second round: b) southern Eurasian cluster and c) northern European clusters; Third round: d) Southern European cluster without Iran, e) Central Sweden, Norwegian West and East and North Estonia, f) Moldavia, North France, West German, Belgium, South Estonia, East German, South German and South Sweden; Forth round: g) Atlantic samples (Spain, South and West France, Ireland and England, h) South-west samples (Italy, Romania, Slovakia and Hungary), i) and i): fourth round of analysis. Details of each level of Hierarchical analysis are present in the corresponding session.

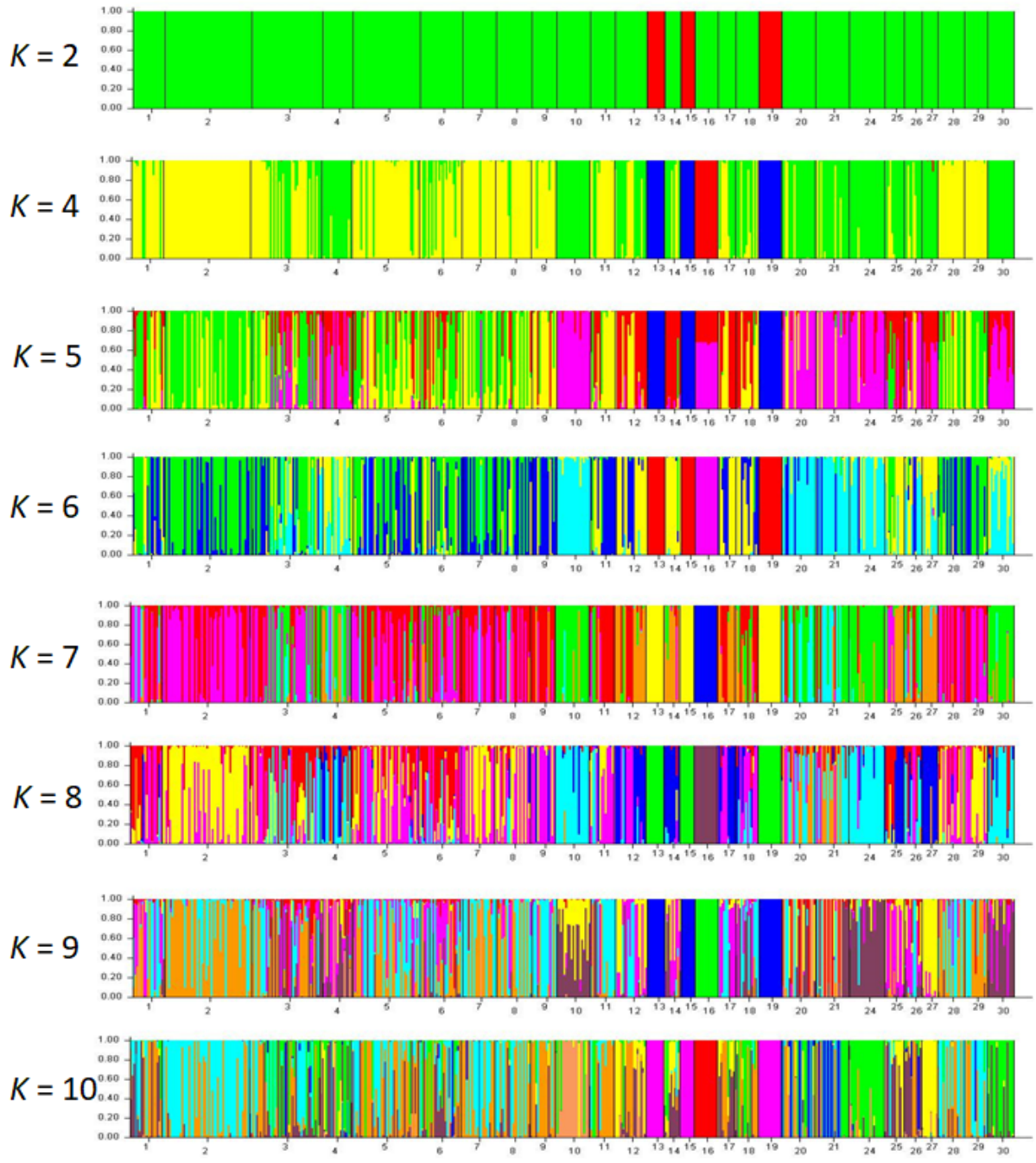


Figure S3. STRUCTURE Individual probabilities for each value of *K* from 2 to 10.

Hierarchical analysis

Finer genetic structure was identified from hierarchical analysis (Figure S6 and S7 for STRUCTURE and DAPC analysis, respectively). The southern Eurasian cluster was further separated into two differentiated clusters, irrespective of the approach used (STRUCTURE or DAPC). The STRUCTURE approach separated Iran from the remaining samples, while the DAPC approach assigned most individuals from both Iran and Turkey samples to the same cluster (violet). The northern European cluster was further separated into two to three clusters depending on the methods, DAPC and STRUCTURE, respectively. Clusters identified by the DAPC approach were distributed almost equally among the different sampled locations. Of the three clusters identified by STRUCTURE, the orange and green ones showed a clear affinity to certain sample locations, while the grey cluster was represented in all sampled locations. No further structure was identified for the African cluster in both methods.

The DAPC's third round of analysis was unable to identify further genetic structure in the northern European cluster. It did however identify two groups inside the southern Eurasian cluster (without Turkey and Iran as a result of previous analyse). It appears that individuals from Spain, Western France and Ireland were mainly assigned to one (light blue) cluster. No other cluster was identified by the DAPC approach regarding refined hierarchical analysis. The STRUCTURE's third round of analysis was able to identify a K = 4 in the southern European cluster. Individuals from Turkey were assigned to an exclusive cluster (grey). Individuals from south-western Europe and from Italy were mainly assigned to one cluster (orange), while those from Spain, West France and Ireland were grouped in a different cluster (blue). The fourth cluster (green) was distributed across all sampling locations with few individuals (11 out of 179) exhibiting more than 50% of assigning probability. In the northern European cluster, for this third round of hierarchical analysis, individuals were regrouped according to population probabilities *of the two almost exclusive clusters from last step, green and orange ones. From this third round until the last one, Evanno's method (Evanno et al. 2005) always identified two clusters, but the analyses of ln[Pr(X|K)] was not clear in identifying those clusters (Figure SX). Also, individual probabilities of inside those K = 2 clusters show very mixed populations. The results*

for those subsequent rounds with a K = 2 are presented in Supplementary Information (figure SXX). We did a fourth and last round of hierarchical analysis for the two main southern Eurasian clusters identified in the previous round: (i) one cluster composed of Spain, West and South France, Ireland and England samples and (ii) the other cluster composed of Italy, Romania, Hungary and Slovakia. For the first one, Evanno's method identified K = 6, but the analysis of ln[Pr(X|K)] does not indicate any structure. For the later, both methods clearly identified a K = 7 structuring. In both cases, clusters are mainly distributed in all sample sites and very rarely a single individual had ~100% probability of being assigned to a particular cluster. The exceptions were individuals from West France and Ireland for which probability values to be assigned to the same cluster reached one.

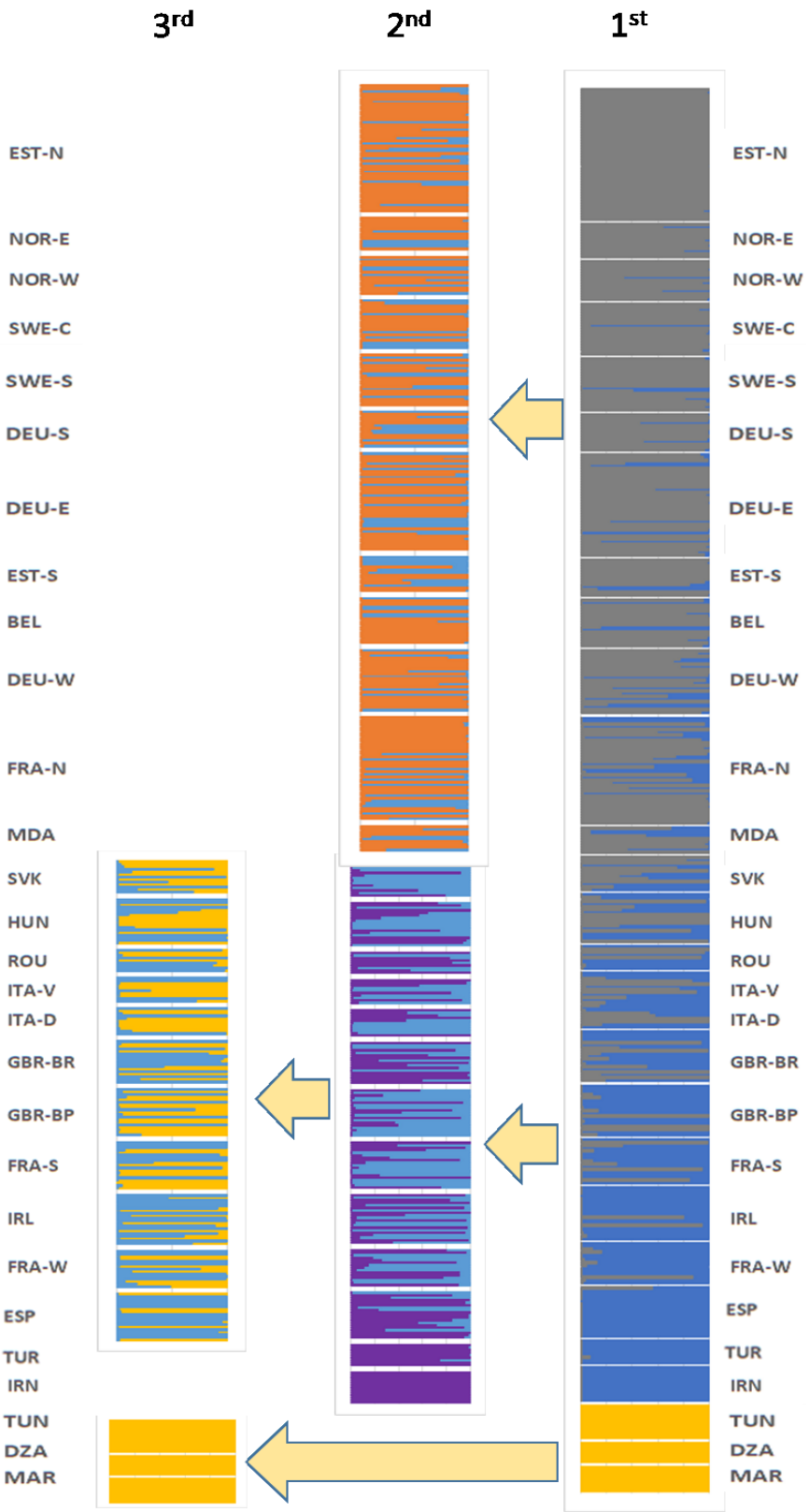


Figure S4. DAPC Hierarchical analysis. Each column corresponds to one level of analysis.

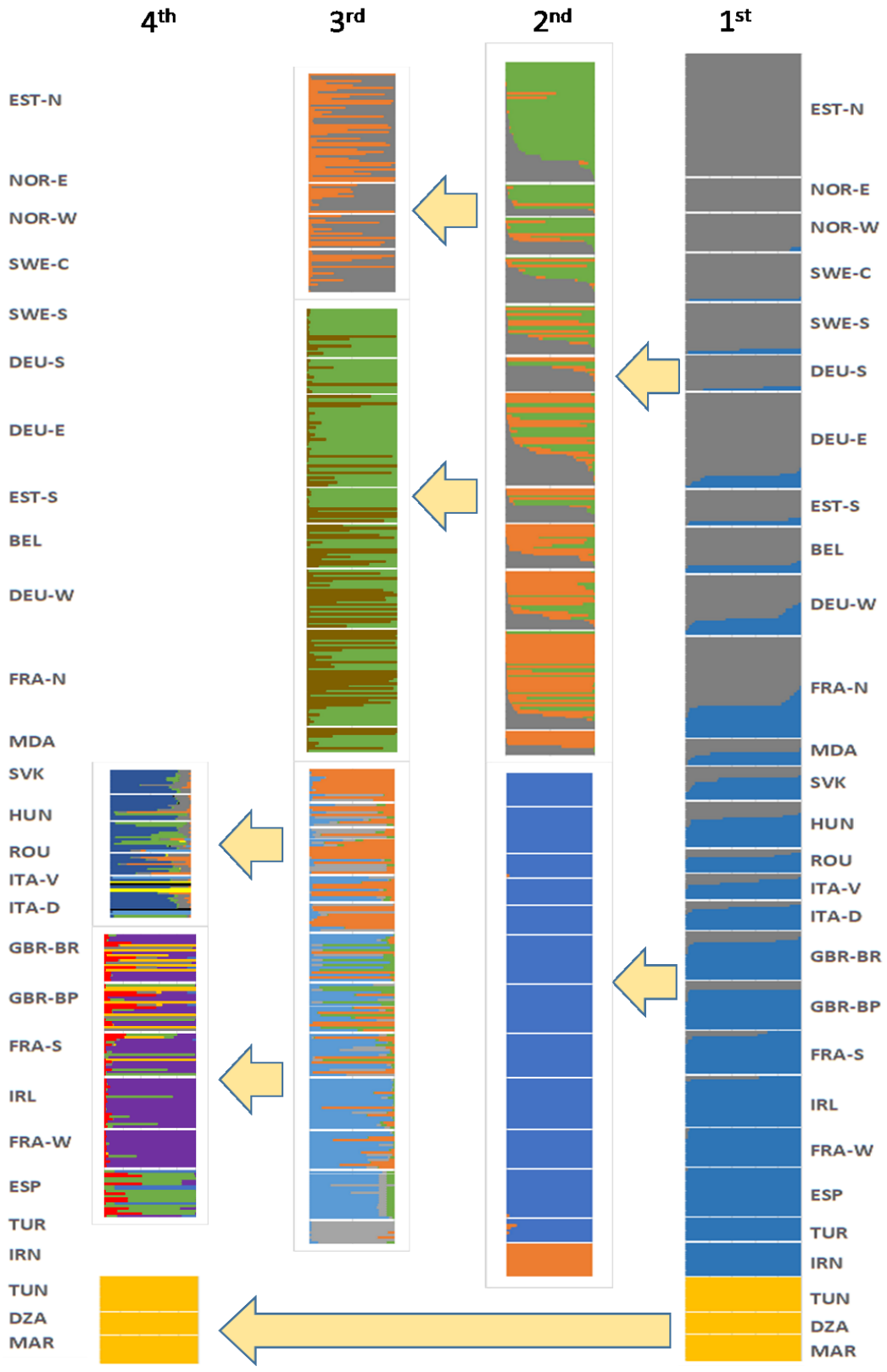


Figure S5. STRUCTURE Hierarchical analysis. Each column corresponds to one level of analysis.

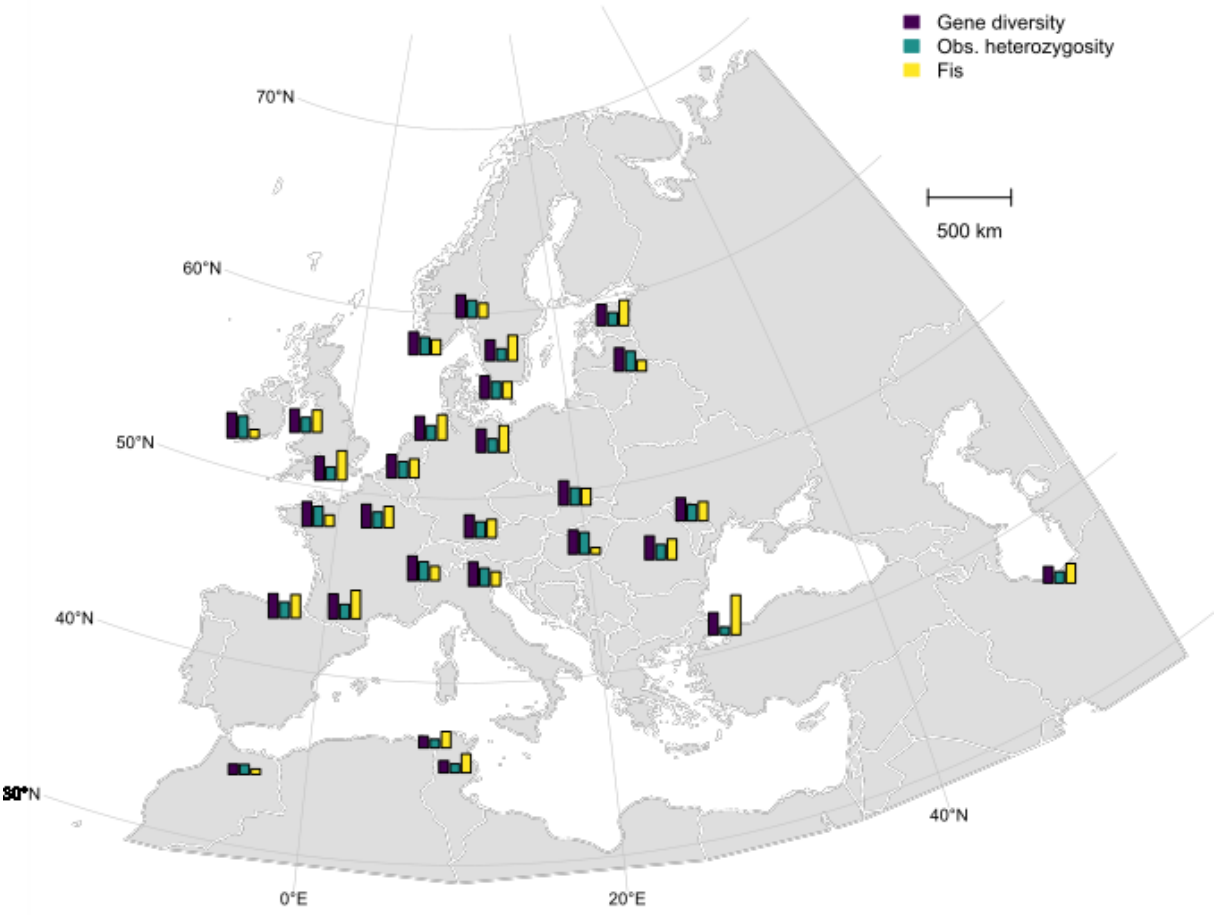


Figure S6. Mean gene diversity, observed heterozygosity, and *F*is per population. Mean population gene diversity was always greater than the observed heterozygosity and *F*is was always positive.

Appendix 2: Supplementary Information for the submitted article: How does incorporating genetic information improve species distribution models?

Supporting Information for:

How does incorporating genetic information improve species distribution models?

Pedro Poli, Annie Guiller, Jonathan Lenoir

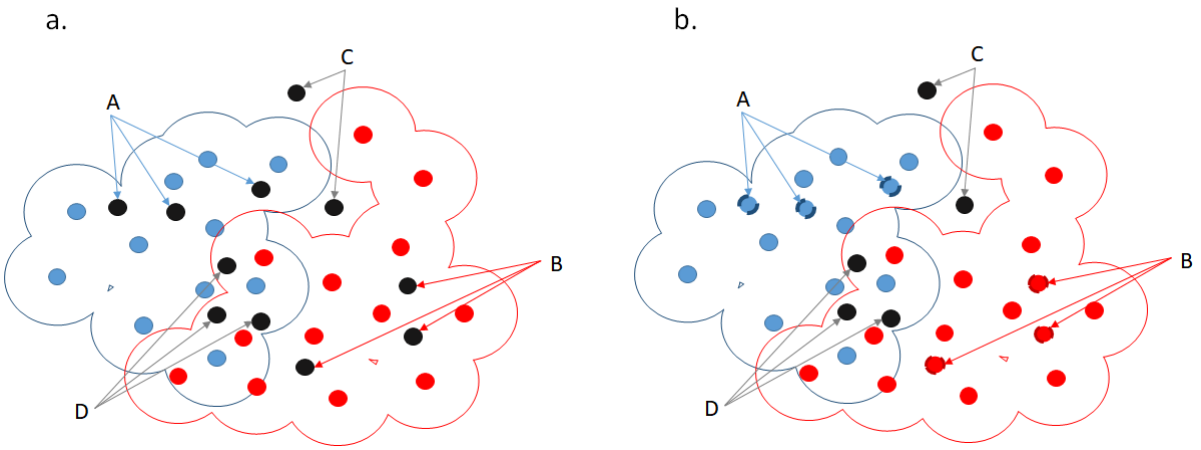


Fig. S1. Schematic representation of presence selection from the European Forest database for *Fagus sylvatica*. First we built and dissolve buffers around each presence points from Magri et al. (2006) for which lineages were already assigned (a.). Whenever a presence from the database was situated exclusively inside a buffer of one lineage, that presence was assigned for the corresponding lineage (b). In this representation, the three points identified as A were assigned to the blue lineages, while points B were assigned to the red one. Points C were situated outside all buffer zones, and were not incorporated to the final dataset. Points D were situated in the interface of the buffer zones of both lineages, and so were also excluded from the final dataset.

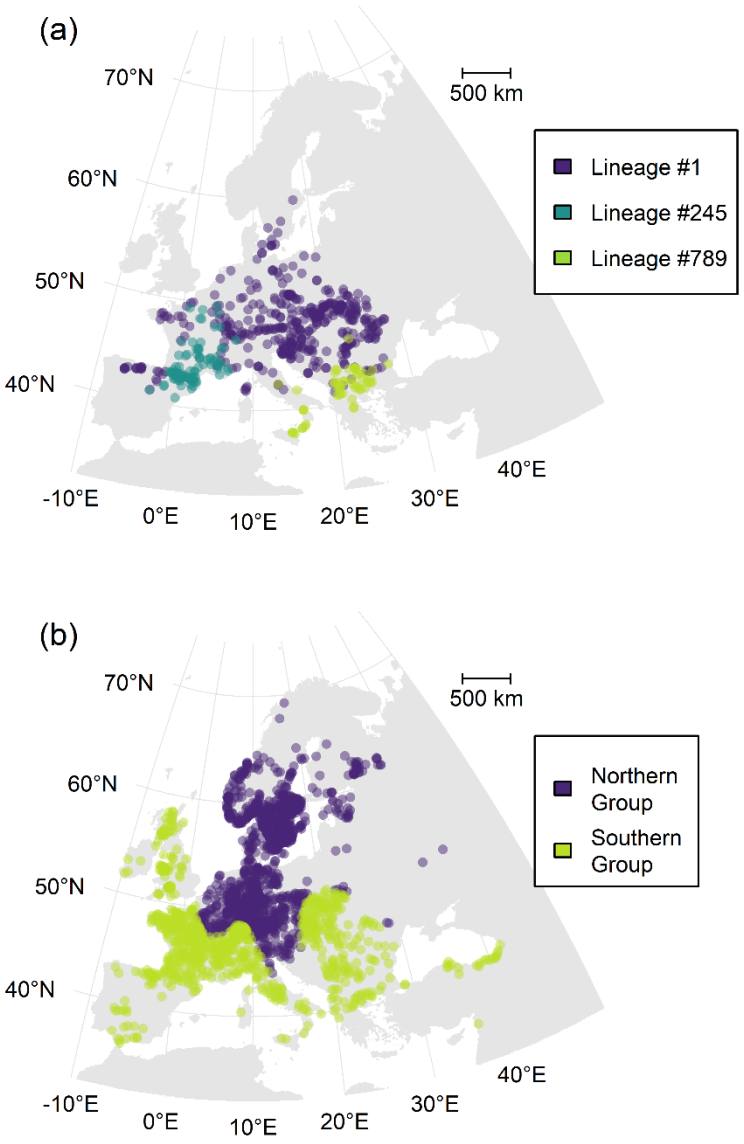


Fig. S2. Presences of each genetic unit of *Fagus sylvatica* (a) and *Ixodes ricinus* (b) used to build genetic-based and traditional species distribution models (SDMs).

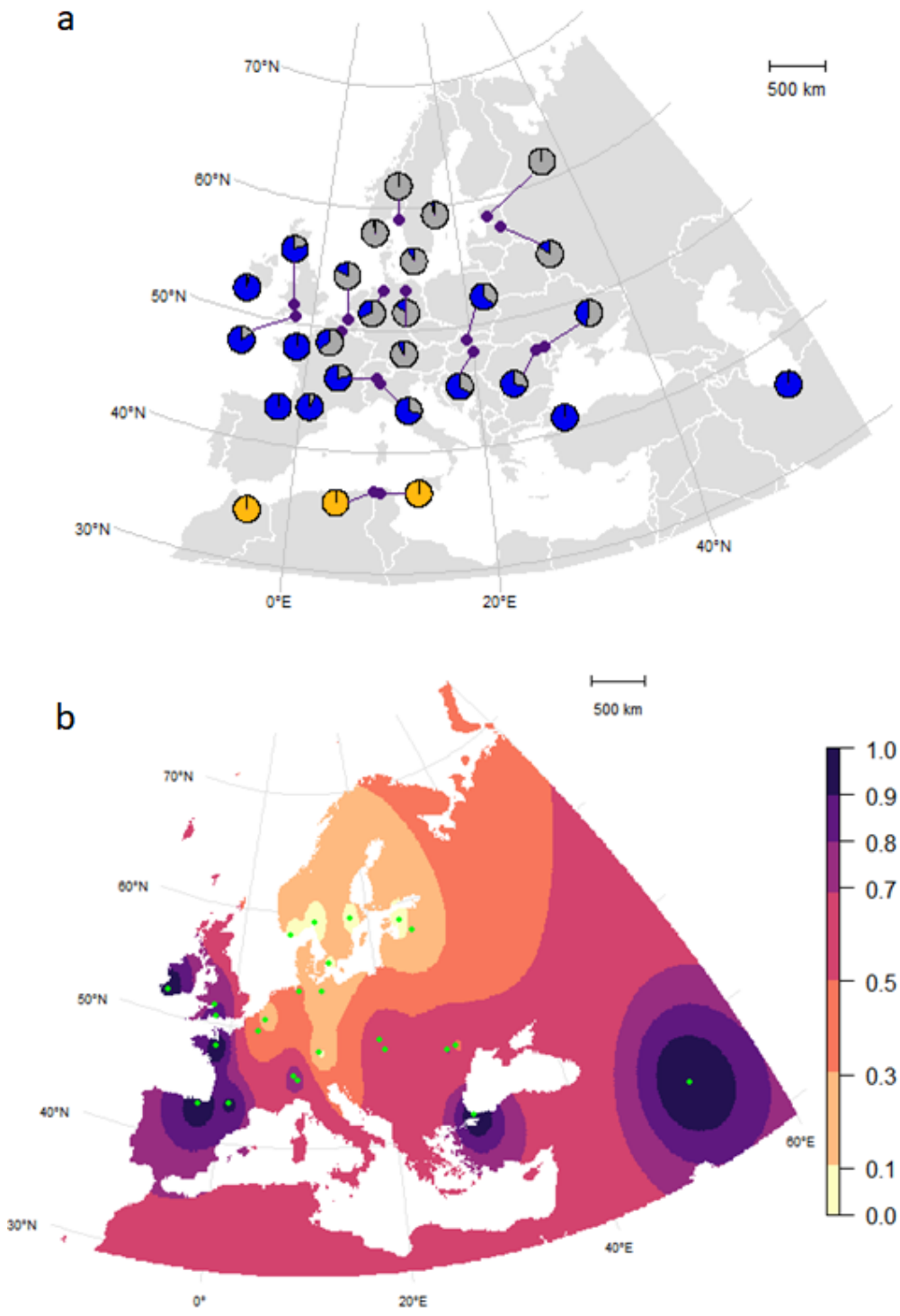


Fig. S3. Probabilities of assignment to one of two genetic clusters identified in Poli et al. (2020) according to STRUCTURE (Pritchard et al., 2000) results. In figure **a** we reproduce figure 3b from Poli et al. (2020). In figure **b** the population probabilities were interpolated to assign a cluster identity for each presence point from GBIF and Vector Map. The probabilities in the interpolated map are for assignment to the southern cluster: higher probabilities represent zones most likely occupied by individuals from the southern cluster while small probabilities represent zones most likely occupied by the northern cluster.

Results

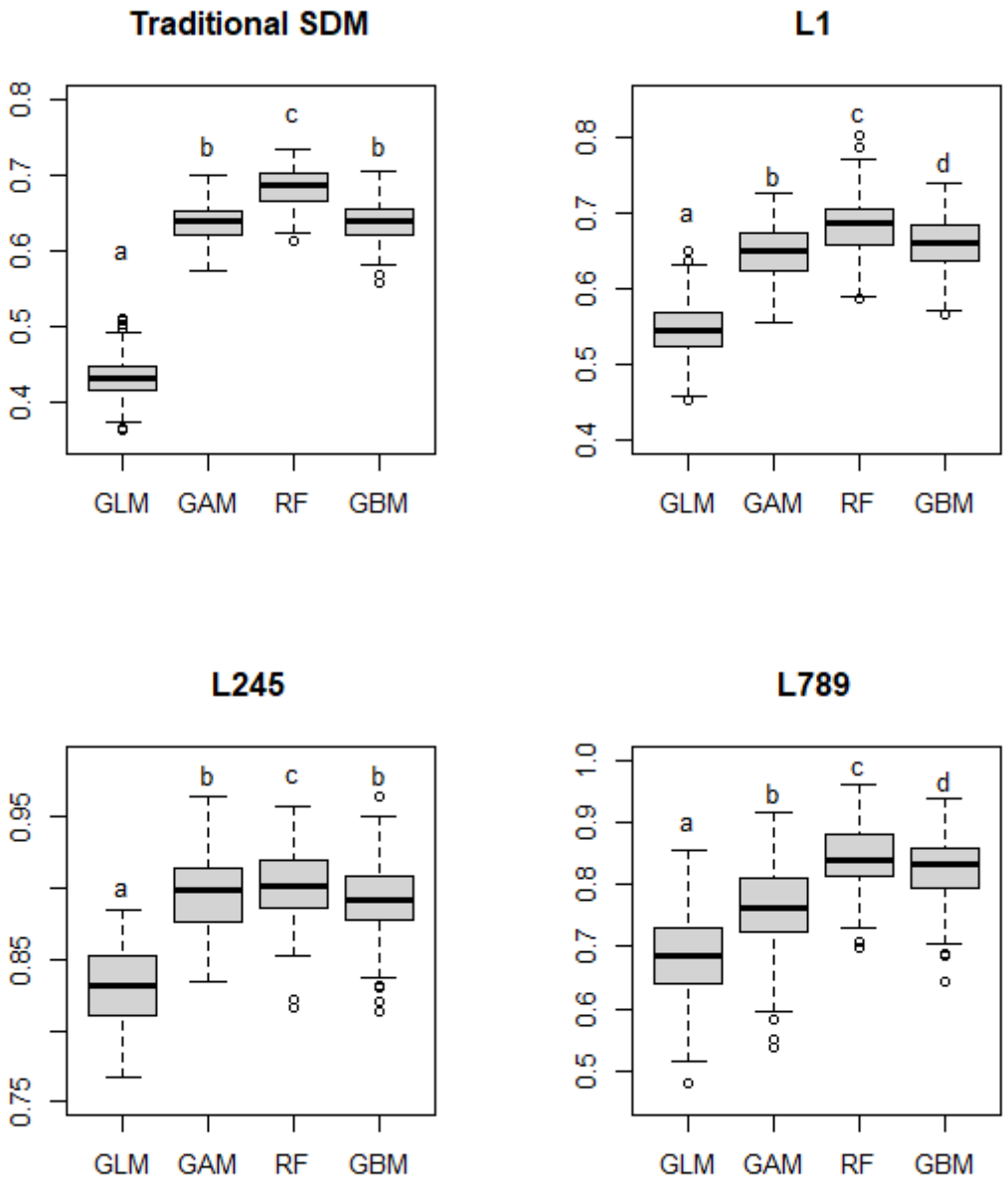


Fig. S4. AUC values for each algorithm across 20 repetitions of each one of the models for *F. sylvatica*: complete model, lineages #1, #245, and # 789.

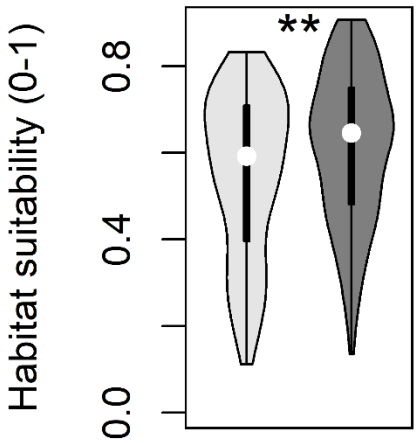


Fig. S5. Distribution of the probabilities of presence of *Fagus sylvatica* during the Mid-Holocene period at spatial locations where pollen records of *F. sylvatica* from the Mid Holocene period have been found (median threshold, n>=37 pollen records per site and time). Light grey: traditional species distribution model (SDM) approach. Dark grey: genetically-informed SDM approach. Stars display the significance level based on a Mann-Whitney test of difference between the two SDM approaches (**, p = 0.004).

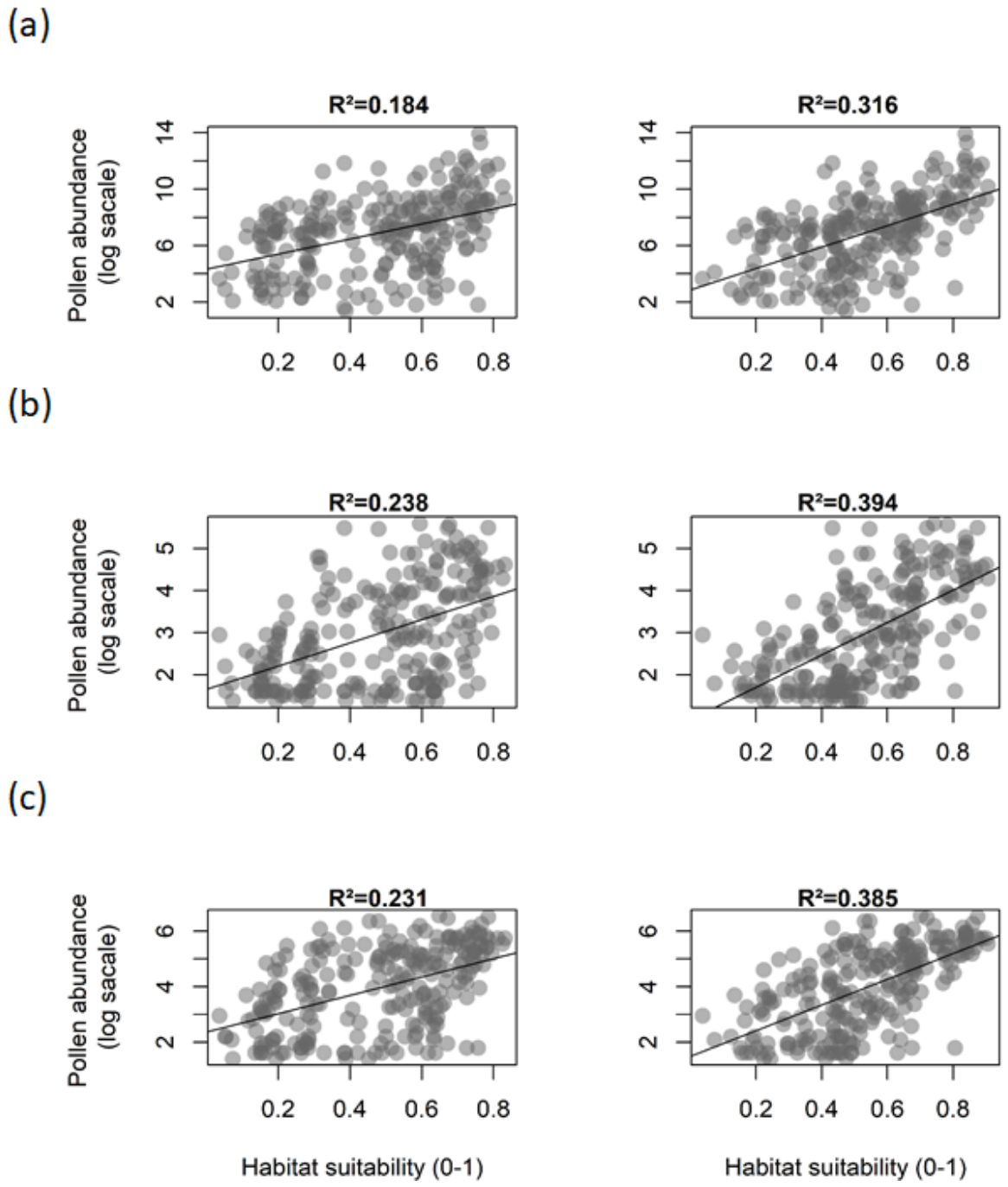


Fig. S6. Linear Regression of the logarithm of the sum (a), median (b) and maximum (c) of pollen abundance per grid cell (log-transformed) with a threshold of 4 pollen records (first quartile) as a function of the probability of occurrence according to the traditional SDM approach (first column) and the genetically-informed approach (second column).

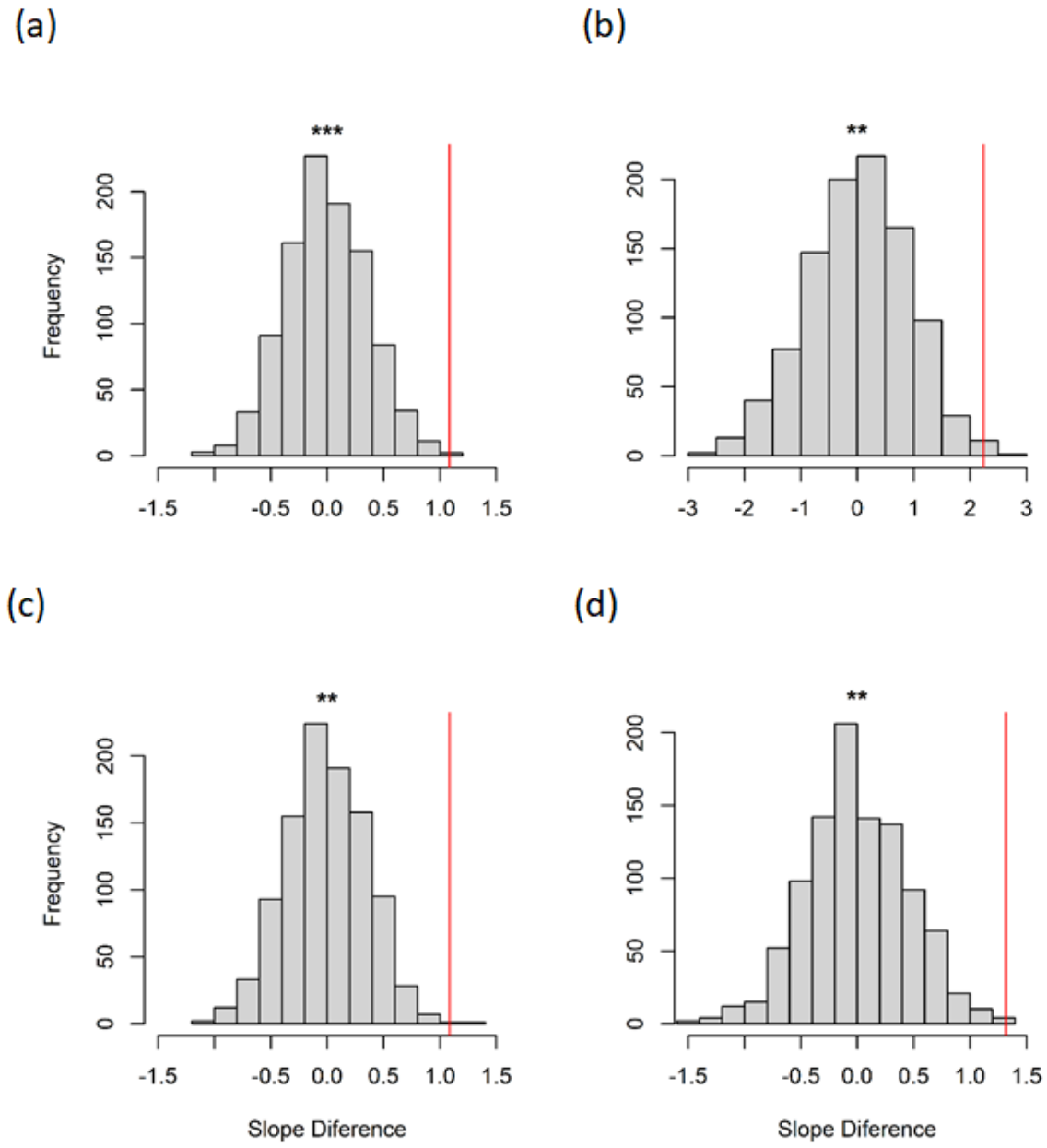


Fig. S7. Differences in the slope of the regressions of the pollen abundance as a function of the probability of occurrence from the traditional and genetic-based SDM approaches were tested by a 1000 permutations. Here, the histograms of the simulated differences are presented for the mean (a), sum (b), median (c) and maximum (c) of the pollen abundance per grid cell with a threshold of 4 pollen records (first quartile), and the red line indicates the observed difference. For all metrics, *p*-values were smaller than 0.01 (but see Table S2 for details on the significance *p*-values).

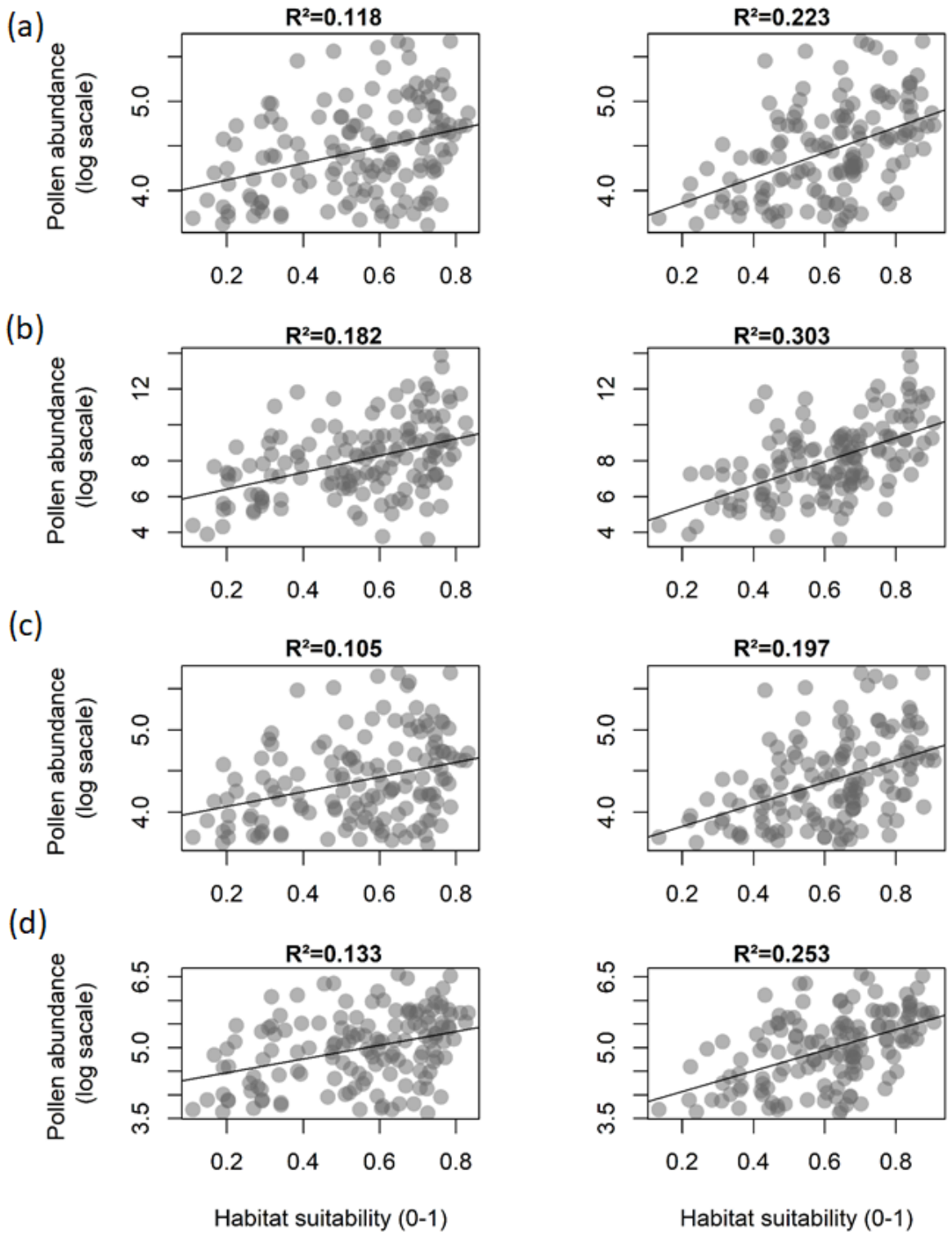


Fig. S8. Linear Regression of the logarithm of the mean (a), sum (b), medium (c) and maximum (d) of pollen abundance per grid cell with a threshold of 37 pollen records (medium) as a function of the probability of occurrence according to the traditional SDM approach (first column) and the genetically-informed approach (second column).

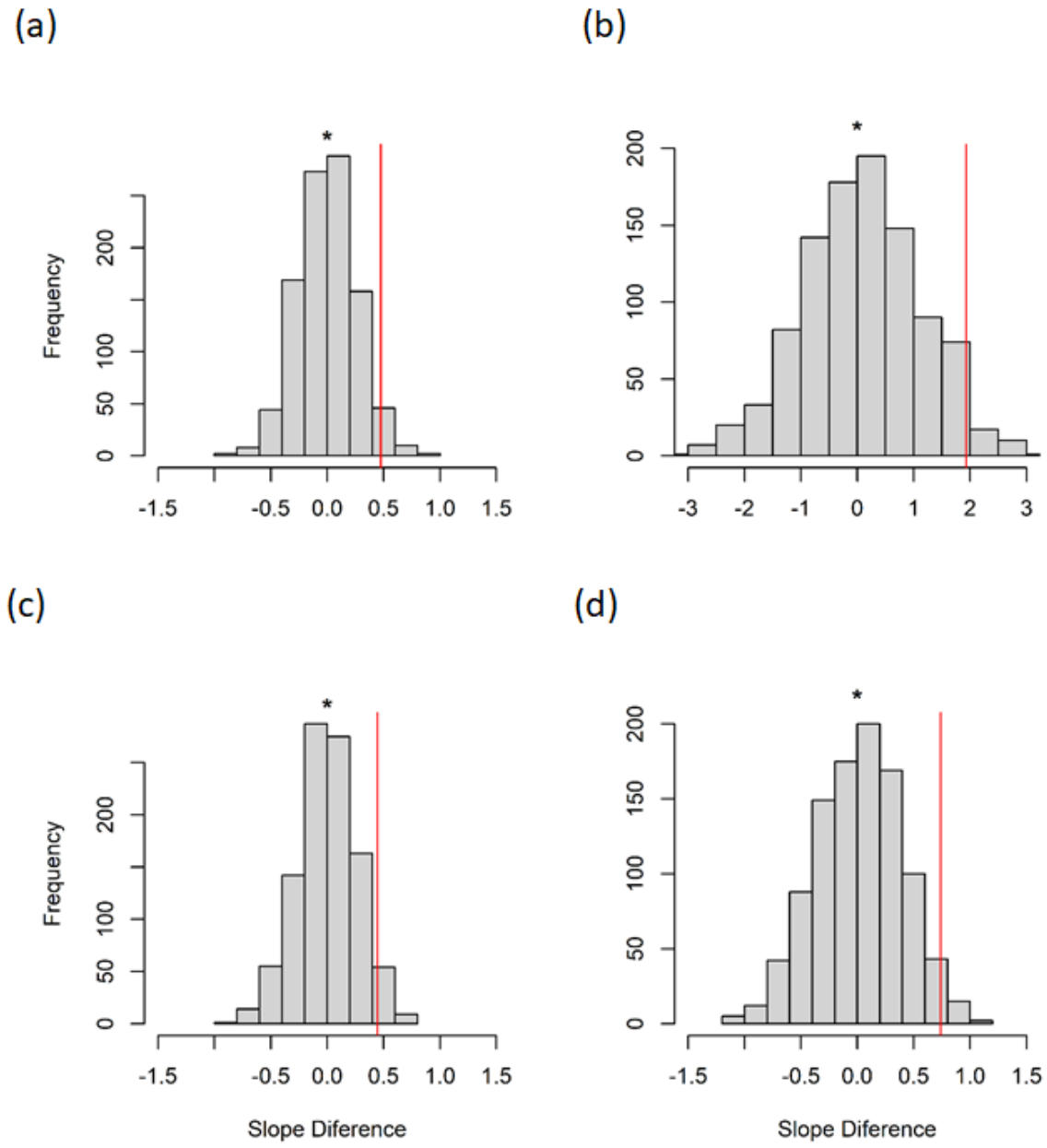


Fig. S9. Differences in the slope of the regressions of the pollen abundance as a function of the probability of occurrence from the traditional and genetic-based SDM approaches were tested by a 1000 permutations. Here, the histograms of the simulated differences are presented for the mean (a), sum (b), median (c) and maximum (d) of the pollen abundance per grid cell with a threshold of 30 pollen records (median), and the red line indicates the observed difference. For all metrics, *p*-values were smaller than 0.05 (but see Table S2 for details on the significance *p*-values).

Table S1. Values (mean±sd) for the six metrics applied to the evaluation of model performance for both traditional and genetically-informed approaches. Significance of the Mann-Whitney test is represented by asterisks: ***: $p<0.001$; **:$0.001<<0.01$; *:$0.01<p<0.05$.

Species	Metric	Traditional SDM	Genetic-based SDM	Significance
Ixodes ricinus	AUC	0.9173±0.0079	0.9133±0.0092	
	Sørensen	0.8439±0.0092	0.839±0.008	
	Sensitivity	0.8907±0.0108	0.9584±0.0095	***
	Specificity	0.7811±0.0172	0.6759±0.0195	***
	TSS	0.6718±0.0205	0.6343±0.0202	
	OPR	0.1982±0.0126	0.2538±0.0114	*
Fagus sylvatica	AUC	0.8263±0.01	0.8215±0.0109	
	Sørensen	0.8325±0.0091	0.8348±0.0075	
	Sensitivity	0.8873±0.0107	0.9081±0.0108	***
	Specificity	0.7555±0.0204	0.7324±0.0184	***
	TSS	0.6427±0.0213	0.6405±0.0179	**
	OPR	0.2158±0.0141	0.2274±0.0117	***

Table S2. Significance values of the statistical tests comparing pollen abundance of *F. sylvatica* on the Mid-Holocene period and probability of occurrence according to the traditional and the genetic-based SDM approaches. The permutation test was applied to test the difference in the slope of the regression of the logarithm of the abundance of pollen as a function of the probability of occurrence between the two approaches. Two threshold of abundance across records per grid cell and time were considered: the first quartile (n>=4) and the medium (n>=30). Four metrics of pollen abundance across sites on a grid cell were considered: mean, sum (total), median and maximum.

	Mean	**Sum**	**Median**	**Max**
First quartile (n>=4)	<0.001	0.007	0.002	0.002
Median (n>=30)	0.038	0.032	0.044	0.025

References

Poli, P., Lenoir, J., Plantard, O., Ehrmann, S., Røed, K.H., Leinaas, H.P., Panning, M. & Guiller, A. (2020) Strong genetic structure among populations of the tick Ixodes ricinus across its range. *Ticks and Tick-borne Diseases*, **11**, 101509.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945-959.