UNIVERSITÁ DEGLI STUDI DI PARMA

Dottorato di ricerca in Ecologia

Ciclo XXVI

Molecular tools, optimal sampling strategies and biogeographical investigation towards the study of adaptive gene flow in forest trees

Coordinatore:
Prof. Paolo Menozzi

Tutor:
Dr. Stefano Leonardi

Co-tutor:
Dr. Andrea Piotti

Dottoranda: Cristina Leonarduzzi

...we see beautiful adaptations everywhere and in every part of the organic world.

Charles Darwin – The Origin of Species

Ce n'est point de l'espace que je dois chercher ma dignité, mais c'est du règlement de ma pensée. Je n'aurai point d'avantage en possédant des terres. Par l'espace l'univers me comprend et m'engloutit comme un point: par la pensée je le comprends.

Blaise Pascal - Pensées

Acknowledgements

These three years of doctoral studies have been intense, including wonderful episodes and tough moments, challenging me to grow as a "researcher" but also as a person. My greatest thanks go to my supervisors Stefano Leonardi and Andrea Piotti for guiding me through this PhD, providing their scientific and human support and pushing me to do and to be the best I can. Thanks to Stefano for his immense scientific preparation, his passion, his honesty and kindness. I am particularly grateful to Andrea for trusting me, for all the nice moments spent together in the field, in the office or travelling around Italy (but also Slovenia, Croatia, Serbia and Montenegro), and for the difficulties we have overcome.

I am thankful to Beppe Vendramin for hosting me in his lab and for giving me the possibility to work at the silver fir project, and to Dragos Postolache and Ilaria Spanu for their help with the laboratory activities. I also thank Ilaria for her support and her friendship. Thanks to Elena Bianchi for the help in the field and in the lab activities.

Travelling with Andrea along the whole Italian peninsula to sample silver fir populations has been an amazing experience. I would like to express my warm gratitude to all the people that helped me during field activities: Giuliano Menguzzato, Nicola Guarino, Francesco Ripullone, Antonio Saracino, all the people of the "Comunità Montana di Vallo di Diano", the Office for Biodiversity Protection of the National Forest Service of Tarvisio, the National Park Gran Sasso-Laga, and all the people I am probably forgetting. A special thank you to Carlo Urbinati and his research group: Valeria Gallucci, Alma Piermattei, Emidia Santini and Matteo Garbarino for the time spent together in the field, for their contagious passion for nature and the discussions about how to merge genetics and dendrochronology.

My sincere gratitude goes to Steve DiFazio, Sylvie Oddou-Muratorio, Luca Bolzoni and Simone Vincenzi for their useful comments and suggestions that helped to improve the Chapters of this thesis.

I am particularly thankful to Valeria Rossi and Paolo Menozzi for supporting me and for giving me the possibility to attend some courses abroad that have been very important for me.

I am also grateful to the people working in our department and to the past/current PhD students in Ecology, especially to Gianluigi Rossi, Melissa Rosati, Marisa Rossetto and Margherita Lega.

I deeply thank Christian for staying by my side every day and for teaching me how to be a free spirited person. I thank my family for supporting me even without fully understanding my will to become a scientist. My warm gratitude also goes to the people that, in many ways, encouraged me to start this PhD, in particular Andrea Piotti, Brad Oberle, Elena Conti, Peter Linder, Marilena Meloni, Laura Granato, Alok Gupta, Elisa Belotti and Francesca Beggi.

Finally, I thank my wonderful friends that encouraged and supported me despite being physically far away, in particular Simonetta, Lorenzo, Valentina, Giusy, Maria Martina, Samuele, Andrea, Jenny, Francesca, Marco.

Table of contents

Chapter 1:
Introduction

1.1 Thesis overview and aims

My thesis aimed at developing an adequate conceptual and methodological framework for studying the dispersal of potentially adaptive genes, i.e. adaptive gene flow, among forest tree populations along ecological gradients.

According to theoretical expectations, gene flow can prevent genetic population divergence by homogenizing allele frequencies. In forest trees, this expectation has been confirmed by many experimental studies based on neutral molecular markers, showing that gene flow can be effective in maintaining genetic connectivity over large distances (Bacles *et al.* 2005, O'Connell *et al.* 2006, Lander *et al.* 2010). As a consequence, forest trees usually show high within-population genetic variation and weak geographical structure (Hamrick 2004). Nonetheless, population differentiation may arise also in forest trees, especially in widely distributed species, due to isolation by distance, abrupt geographical barriers, or as a consequence of heterogeneous environmental conditions (Gram and Sork 2001, Grivet *et al.* 2011).

Environmental heterogeneity (in particular linked to temperature, precipitation, and aridity) has been shown to influence genetic differentiation among tree populations and to create geographic genetic patterns consistent with adaptive traits (Savolainen *et al.* 2007, Eckert *et al.* 2010, Sork *et al.* 2010, Grivet *et al.* 2011), even in species with high potential for gene flow (Ortego *et al.* 2012, Temunovic *et al.* 2012). In heterogeneous environments, or along ecological gradients, different populations are likely to experience contrasting environmental challenges and selective pressures. Spatial variation in the strength and direction of selective factors may influence the distribution of genetic variation among populations, leading to population differentiation through local adaptation (Lenormand 2002, Jump and Penuelas 2005). Populations are locally adapted when they have their highest relative fitness at their home sites (Savolainen *et al.* 2007). In order for populations to respond to environmental selective pressure through local adaptation sufficient genetic variation should be available (Hamrick 2004).

In the case of species occupying a wide range of distinct ecological conditions, the homogenizing action of gene flow can have contrasting effects on the adaptive potential of

populations. Gene flow, by introducing gametes adapted to a different environment, may have a maladaptive effect and decrease the fitness of individuals. On the other hand, gene flow can also increase within-population genetic variation and spread advantageous alleles, contributing to increasing the local adaptive potential. Gene flow is expected to have a beneficial effect especially in the case of populations genetically impoverished, where the adaptive potential is particularly low. Whether gene flow enhances or counteracts adaptation of forest trees populations is an extremely complex question to answer, because it depends on many factors such as the rate of gene flow, intensity of selection, population size and structure, demographic history, and species-specific life history traits (Krutovsky *et al.* 2012).

With the currently ongoing climate warming, individuals able to perform well in warmer and drier conditions can be expected to be selectively advantaged, both in local persistence and in latitudinal and altitudinal migration/colonization. Although it is not easy task to define when gene flow has an adaptive value, the concept of 'adaptive gene flow' during forest trees' range shifts has been touched upon several times in recent literature. Davis and Shaw (2001), in their pivotal paper on the adaptive responses of forest trees to quaternary climate change, stated that "*the arrival of seeds that are somewhat "preadapted" to the novel climate (e.g., seeds from more southerly populations during periods of climate warming) may contribute to adaptation, yet selection would also promote new genetic combinations, for example, of photoperiod and temperature responses suited to the novel growing season"*. Then, it has been stressed that directional gene flow via-pollen can contribute in a "*more pervasive and faster way"* to this process (Robledo-Arnuncio 2011). If not by directly introducing advantageous genetic variants, gene flow can increase the adaptive potential of a population by incrementing its additive genetic variance (Yeaman and Jarvis 2006). Although few experiments have been aimed at quantifying gene flow along environmental gradients, Gauzere *et al.* (2013) have recently shown that high altitude forest tree populations experienced a larger gamete immigration than lower altitude ones along an altitudinal gradient (Gauzere *et al.* 2013). The Authors noted how "*directional pollen flow carrying preadapted genes to warmer conditions should thus be particularly favourable for enhancing the adaptation of population* (their upper population) *to future climatic conditions"*. Despite the relevance of potentially adaptive gene flow among forest tree populations for predicting their ability to respond to environmental changes, this is still a scarcely explored issue (Kremer *et al.* 2012) and whether the velocity of plant movement will cope with the velocity of climate change is still debated (Corlett and Wescott 2013).

Paternity and parentage analysis allow careful and direct estimation of gene flow in natural

populations. Paternity analysis aims at estimating effective gene flow via pollen, i.e. the rate of local seeds successfully pollinated by external pollen donors (Smouse *et al.* 2004). On the other hand, parentage analysis aims at assessing the rate of seed and pollen immigration by investigating the parentage of established seedlings. Paternity and parentage analysis have been generally carried out using neutral molecular markers, thus providing very useful information but restricted to the investigation of demographic patterns and mating systems. The increasing availability of potentially adaptive markers (i.e. candidate genes, SNPs linked to particular phenotypic or physiological traits) may allow investigation of the rate and direction of gene flow also for loci potentially under selection. Therefore, my thesis aimed at providing the main elements needed to set up an experiment for studying potentially adaptive gene flow in a widespread and economically important tree species, silver fir (*Abies alba* Mill.).

In order to do this, I focused on the following crucial aspects:

- First of all, when using paternity and parentage analyses it is important to choose an adequate sampling design in order to achieve robust estimates. I investigated this aspect by reviewing published paternity studies and by using computer simulations to test the effect of sampling effort on dispersal estimates (in particular on pollen immigration estimates obtained from paternity experiments). This simulation study resulted in practical guidelines for performing future paternity studies (*Chapters 2 and 3*).

- Second, an *ad hoc* set of molecular markers suitable for both biogeographical and gene flow studies should be developed. In order to study adaptive processes it is important to disentangle the effects of demography and past migration on population structure from selection signatures. For doing this, both neutral (i.e. SSRs) and potentially adaptive (i.e. SNPs) markers should be available. Since SNP markers for silver fir have just been developed in a parallel project (Roschanski *et al.* 2013), in my thesis a large set of new microsatellite markers was developed (*Chapter 4*). Until now, only few, null allele prone, molecular markers were available for silver fir.

- Third, in order to maximize the probability of observing adaptive gene flow, transects in areas characterized by steep ecological gradients should be considered. For this reason, I focused on silver fir (*Abies alba* Mill.) populations along the Apennine chain.

These populations represent an ideal system to study adaptive gene flow because they are a hotspot of genetic diversity for the species, besides being highly genetically and ecologically differentiated (Terhurne-Berson *et al.,* 2004, Liepelt *et al.* 2009, Zieghenhagen *et al.* 2005, Piovani *et al.* 2010, Carrer *et al.* 2010). Some evidences of genetic distinctiveness among groups of populations have been previously found, but the relationship among genetic structure, different eco-physiological responses, and the complex evolutionary history of these populations has not been clarified yet (*Chapter 5*).

Chapters 2 and 4 resulted in articles published in peer reviewed journals:

Leonarduzzi C, Leonardi S, Menozzi P, Piotti A (2012) Towards an optimal sampling effort for paternity analysis in forest trees: what do the raw numbers tell us? *iForest*, 5: 18-25.

Postolache D, Leonarduzzi C, Piotti A, Spanu I, Roig A, Fady B, Roschanski A, Liepelt S, Vendramin GG (2014) Transcriptome versus genomic microsatellite markers: Highly informative multiplexes for genotyping *Abies alba* Mill. and congeneric species. Plant Molecular Biology Reporter, in press.

Chapter 3 resulted in a manuscript submitted to a peer reviewed journal.

In the following paragraphs the main themes developed in my dissertation are introduced.

1.2 Dispersal and gene flow in forest trees

Dispersal is the unidirectional movement of an organism away from its place of birth (Levin *et al.* 2003) and it is one of the most important life-history traits in species persistence (Clobert *et al.* 2001). In plants, it is defined as the phase between separation from the parent until the propagules (i.e. pollen and seeds) come to the final rest (Cousens *et al.* 2008). Although dispersal is mostly a passive process in plants, plants have evolved structures to increase the chances that propagules are transported over long distances by wind (e.g. winged seeds, plumed pappus) and/or by animals (e.g. spines, hooks to attach to animals' coats) (Begon *et al.* 2006, Hintze *et al.* 2013). The probability of propagule dispersal often follows a decreasing curve, with most propagules travelling relatively short distances from the parent plant and their density decreasing more or less steeply with distance (Cousens *et al.* 2008).

Dispersal heavily affects ecological and demographic dynamics of local populations, meta-populations and communities (Silvertown 1991, Levin *et al.* 2003, Nathan *et al.* 2008, Clobert *et al.* 2012). At the population level, it determines the spatial structure of populations, and regulates the population density. In meta-populations, it affects the connectivity and distribution of individuals among different patches and the persistence of the species in the whole meta-population (Levin *et al.* 2003). At the community level, dispersal increases species richness and allows to escape from competition, and it is involved in stabilizing species coexistence.

Gene flow is the genetic outcome of the dispersal process, and it is defined as the movement and integration of alleles from one population into another that determines a change in the allele frequencies of populations (Slatkin 1987). Gene flow rate into a population depends both on the dispersal ability of the species and on the probability that migrants successfully mate with local individuals and produce viable offspring able to successfully transmit immigrant gametes. Mechanisms of reproductive isolation, such as assortative mating or lower fitness of hybrids, can determine low gene flow despite high dispersal (Endler 1977).

Pollen and seed dispersal are two critical and independent vectors of gene flow in forest trees (Ennos 1994, Tarazi *et al.* 2013). Many studies on forest species have shown that pollen can travel very long distances (hundreds of kilometres) and still be viable, especially for wind-pollinated species (e.g. Robledo-Arnuncio 2011, Williams 2010, reviewed in Kremer *et al.* 2012). Therefore, pollen dispersal is considered the main vector responsible for reproductive connectivity among populations and for maintenance of genetic variation within populations. On the other hand, seeds generally reach shorter distances due to their larger size (Bittencourt and Sebben 2007, but see Bacles *et al.* 2006 and Piotti *et al.* 2010), but their role is crucial *i*) because they are the main determinants of the recruitment pattern, *ii*) because they carry twice as genetic information compared to pollen, *iii*) for ecological processes such as the colonization of unoccupied habitats and range shifting, and *iv*) for the demographic balance (density regulation) of populations and meta-populations of forest trees (Clobert *et al.* 2012).

1.3 Genetic and evolutionary relevance of dispersal

Dispersal and gene flow have marked effects on the spatial distribution of genetic variation within and among populations. Within population, dispersal affects population structure and mating patterns. If dispersal is restricted there is an increased probability of mating among related individuals, leading to inbreeding, which in turn can reduce the genetic diversity of the

population and decreases individual fitness (inbreeding depression). Among populations, gene flow tends to homogenize allele frequencies over populations, lowering their differentiation and limiting the loss of genetic variation (Slatkin 1987).

The role of gene flow is critical in small and fragmented populations, that are usually found in peripheral areas of species' distributions and in areas heavily impacted by human exploitation. Fragmented populations are more prone to genetic diversity loss since inbreeding and genetic drift effects become stronger as the population size decreases (Ellstrand and Elam 1993). The loss of genetic variation increases the extinction risk of the populations (Spielman 2004). Genetic variation is important to ensure long-term evolutionary potential to populations and aid population persistence (Hoffmann and Sgro 2011). Gene flow can provide a "genetic rescue" by introducing new genotypes through recombination, and counteracting the loss of genetic variation by promoting the spread and establishment of new alleles in the population gene pool. In addition, seed immigration acts also as a "demographic rescue" by reducing the negative effects of small population size and peripherality (Garant *et al.* 2007, Sexton *et al.* 2011).

Similarly, gene flow can counteract the negative effects of fragmentation (Bacles *et al.* 2005, Lander *et al.* 2010, Leonardi *et al.* 2012). Habitat fragmentation disrupts populations in small separated units (Young *et al.* 1996; Fahrig 2003) that are expected to become more and more differentiated and to undergo a depauperation of genetic variation. Many studies on the effects of habitat fragmentation have highlighted a genetic diversity reduction in fragmented plant populations (reviewed by Aguilar *et al.* 2008). Nevertheless, most of these studies were based on short-lived herbaceous plant species. When considering forest trees separately, scarce empirical evidence of the expected effects of habitat fragmentation has been found (aka "the paradox of forest fragmentation genetics", Kramer *et al.* 2008). An increasing number of studies, based on indirect (F_{ST}-based) as well as direct (parentage and paternity analysis) methods, have shown that forest trees have high dispersal abilities, and that gene flow among fragmented stands can be effective in contrasting the negative consequences of habitat fragmentation (Aldrich and Hamrick 1998, Bacles *et al.* 2005, 2006; but see Jump and Penuelas 2006 for an exception). The high dispersal ability of forest trees has been suggested as the most likely explanation for their observed "genetic resilience" to perturbations (Hamrick 2004, Kramer *et al.* 2008). Due to their high potential for extensive gene flow, forest tree species have been often considered a unique reproductive unit (i.e. meta-population) over their whole bio-geographical range, rather than a set of separate populations (Nathan 2006). This could overcome, or at least limit, fragmentation effects

related to anthropogenic or selective factors (Hamrick 2004, Kremer *et al.* 2012). On the other hand, it could also be too early for the genetic effects of fragmentation to be currently visible, due to the long generation times of forest trees. It is not clear to what extent gene flow prevails over reduced population size in limiting or delaying fragmentation effects in forest trees. Ecological consequences of fragmentation, such as reduced pollen and seed production or recruitment failure, could appear earlier and be more reliable signals of the consequences of fragmentation. However, they have been rarely investigated and studies investigating both ecological and genetic effects are extremely rare (Knapp *et al.* 2001, O'Connell *et al.* 2006).

Dispersal is often described in terms of dispersal kernel, which is defined as the probability density function of dispersal distances from individual plants (Nathan and Muller-Landau 2000, Oddou-Muratorio *et al.* 2005). Most of the dispersed propagules travels short distances and lands close to their source (i.e. parental tree), but a fraction of them are able to reach extensive distances, resulting in long distance dispersal events (LDD) (Nathan 2002). LDD represents the tail of the propagules' dispersal kernel. The characterization of the tail of the dispersal kernel is a complex task as LDD events are difficult to track, due to their rarity and the spatial scale at which they occur. LDD events have great evolutionary, ecological and conservation value. In fact, LDD can keep highly distant populations connected, and determine the rate of spread of expanding populations and the species' ability to track climate change by dispersing into suitable habitats (Nathan *et al.* 2012). On the other hand, LDD can also increase the risk from invasive species and genetically modified plants. For these reasons, the accurate characterization of the dispersal kernel, and -in particular- of its tail has an extreme relevance. Many efforts have been devoted to accurately describe dispersal kernels using several modelling approaches (from empirical to mechanistic and semi-mechanistic models, Katul *et al.* 2005, Stoyan and Wagner 2001, Nathan *et al.* 2001, see Nathan *et al.* 2012 for a general overview on dispersal kernels). An increasing number of studies has shown that the kernel's tail is fatter than what was thought after early seed and pollen trapping experiments (Nathan *et al.* 2002, 2003, Nathan *et al.* 2005) and that the near- and far-components of dispersal are likely to be ruled by different factors (Clark 1998, Goto *et al.* 2006, Kuparinen *et al.* 2009, Williams 2013).

1.4 Gene flow and adaptation in changing environments

The tendency of gene flow to homogenize allele frequencies over long distances can have contrasting effects on the evolution of populations, especially when divergent selection acts in opposite directions (Garant *et al.* 2007, Robledo-Arnuncio 2011, Kremer *et al.* 2012). When the environment is largely heterogeneous, populations can become differentiated, despite having large within-population genetic variability. In these cases, extensive gene flow among populations can lead to migration load, which is defined as a change in allele frequencies in the opposite direction of natural selection (Lenormand 2002, Lopez *et al.* 2008). The introduction of genotypes adapted to different environments hampers local adaptation, and the mating among local and immigrant individuals can result in low-fitness offspring (outbreeding depression). An "intermediate level" of gene flow has been suggested to be the optimum for local adaptation because it increases additive genetic variance providing the basis for adaptive divergence without swamping allele frequencies (Alleaume-Benharira *et al.* 2006). The "optimal" gene flow level depends on many factors such as the population size, the degree of habitat heterogeneity and the strength of selection (Garant *et al.* 2007, Schiffers *et al.* 2013).

Also in this case, small and fragmented populations are the most sensitive to the effects of maladaptation. The risk is particularly intense for peripheral populations. These populations are characterized by small population size, high fragmentation and low density (Eckert *et al.* 2008), and they are threatened by highly asymmetric gene flow from larger central population (Lenormand 2002, Sork and Smouse 2006). It is important to preserve these populations because they are genetically unique since they can harbor rare alleles well suited to face changing environmental conditions (Hampe and Petit 2005). Furthermore, the rear edge populations of temperate tree species are strongly affected by climate change.

The interplay between the effects of gene flow and adaptation has a prominent role in the current global warming scenario (Kremer *et al.* 2012). Dispersal to a more suitable environment is probably the most effective mechanism to cope with large scale and rapid environmental modifications, especially for long-lived organisms (Clobert *et al.* 2012). Dispersal abilities of forest trees place maximum limits on the ability to track suitable environments. Gene flow estimates have usually been carried out using only neutral genetic markers but, in the global warming scenario, investigating how potentially adaptive genes (e.g. linked to drought or frost tolerance) move across the landscape could be crucial for

predicting the populations' fate, as dispersal along steep climatic gradients can have marked adaptive or maladaptive effect depending on the amount and direction of gene flow (Savolainen *et al.* 2007, Aitken *et al.* 2008, Robledo-Arnuncio 2011). The recent development of potentially adaptive molecular markers (SNPs linked to adaptive genes, Gonzalez-Martinez *et al.* 2006, Grivet *et al.* 2011, Neale *et al.* 20078) provides promising tools for estimating potentially adaptive gene flow from warmer to colder regions in future studies.

1.5 How to measure dispersal and gene flow

The number of empirical studies on dispersal and gene flow is increasing in recent literature. However, measuring dispersal is still a challenging task in plants, especially in natural populations, because it involves the quantitative characterization of the distribution of propagules with distance from the source plant (Ouborg 1999, Levin *et al.* 2003, Bullock *et al.* 2006, Jordano 2007)..

The first approaches to measure dispersal have generally been ecological methods. They are based on tracking (i.e. tracking the trajectories of marked propagules from known source-parents to deposition points) or trapping (i.e. inferring the propagule source from the observed pattern of trapped propagules in the study area) (Skarpaas *et al.* 2011). These methods are difficult to apply to large stands with multiple propagule sources, and have been shown to lead to LDD underestimation (Ouborg 1999).

Later on, with the advent of molecular tools, approaches based on the population genetics consequences of dispersal have started to be used (Silvertown 1991, Ouborg 1999). There are two main categories of population genetics approaches to measure dispersal: indirect and direct methods. Indirect methods estimate historical gene flow rate from the amount of differentiation among populations (e.g. F_{ST}). Direct methods (i.e. paternity and parentage analysis) estimate contemporary gene flow by reconstructing parent-offspring relationships and consequently detecting immigration events. Focusing on paternity analysis, the reconstruction of parent-offspring relationships is based on the comparison among the genotypes of all adults (potential pollen donors) in the studied population (sampled area) with a representative sample of seeds collected from selected mother trees (acting as pollen traps). For each seed, the compatibility of its genotype with the adult genotype is investigated. If the adult genotype is not compatible with the seed genotype then that adult is excluded from the potential parents (Jones and Ardren 2003). Highly polymorphic molecular markers are needed to have an exclusion probability that allows identification of one and only one true parent.

Once the parent has been identified, it is possible to estimate parent-offspring distances and to infer the dispersal curve. If no compatible parent is found within the sampled population the seed is considered pollinated from immigrant pollen. So far, microsatellites are considered the best type of molecular marker for this purpose because they are codominant, hypervariable, selectively neutral and very frequent (Ashley 2010, Jones *et al.* 2010).

Paternity and parentage analysis are probably the most used methods to estimate gene flow rates and to characterize pollen and seed dispersal kernels. These methods also provide information on mating patterns, reproductive success and population structure. The ideal situation to perform parentage/paternity analysis is to sample all potential parents in the population studied (usually easier in small and isolated populations) and to have a set of molecular markers with high exclusion probability. If this is the case the identification of the true parent can be done using the exclusion method (all candidate parents are excluded except one, which is considered the true parent, Dow and Ashley 1996). If the resolution of the marker set is too low and/or the number of candidate parents is large, maximum-likelihood (categorical and fractional allocation) or Bayesian methods can be used (Jones *et al.* 2010, Klein and Oddou-Muratorio 2011.). Alternatively, when the focus is on dispersal parameters rather than on parental identification, the so-called "full-probability methods" can be used (Hadfield *et al.* 2006). These methods use a modelling approach to estimate simultaneously parentage and other population-level variables of interest (e.g. fecundity, traits linked to reproductive success) (e.g. Neighbourhood model, SEMM, Adams and Birkes 1991, Oddou-Muratorio *et al.* 2005, Goto *et al.* 2006, Hadfield *et al.* 2006, Chybicki and Burczyk 2010). Full-probability methods can provide an accurate assessment of confidence in estimates of variables of interest (Oddou-Muratorio *et al.* 2005). Also uncertainty due to the genetic marker set can be included (Moran and Clark 2011).

As pointed out by Jones *et al.* (2010), there are mainly two key elements that make the results of a paternity/parentage study robust: an adequate sampling design and a set of highly polymorphic markers. An adequate sampling design is crucial to obtain reliable estimates, although the impact of sampling design on paternity results has received relatively little attention so far (Jones *et al.* 2010, Leonarduzzi *et al.* 2012, Nathan *et al.* 2012). Estimates of gene flow and dispersal kernel obtained through paternity analysis are dependent on the experimental design (Robledo-Arnuncio *et al.* 2007, Jones and Muller-Landau 2008, Moran and Clark 2011, Leonarduzzi *et al.* submitted). Just mentioning the most obvious cause of bias, the spatial distribution of chosen mother trees (i.e. pollen traps) sampled within a population may cause some dispersal distances to be over-represented with respect to others.

Unless the relative position of sampled mother trees with respect to potential pollen donors is taken into account, this can generate a bias in the inferred dispersal kernel and in gene flow estimates.

More effort has been devoted to uncertainty linked to genetic data. When the exclusion probability of the marker set is low the probability of misassignments increases, leading to unreliable gene flow estimates. Issues linked to the exclusion probability of the marker set and genotyping errors have been carefully investigated (e.g. Dakin and Avise 2004, Slavov *et al.* 2005, Burczyk *et al.* 2006). Some recently developed Bayesian methods can also take into account weaknesses due to the genetic data in parentage estimates (Klein and Oddou-Muratorio 2011, Moran and Clark 2011). In addition, the advent of next-generation sequencing promises to provide a virtually unlimited number of molecular markers and to open up the possibility of monitoring gene flow at the genomic level by scanning genome-wide nucleotide variation (Krutovsky *et al.* 2012).

Chapter 2:

The effect of sampling design on the estimate of pollen dispersal patterns in paternity studies: a literature survey

2.1 Introduction

Paternity analysis, based on the evaluation of the compatibility between the genotypes of adult male individuals and the male genotypic contribution to seeds, is the preferred method for estimating pollen-mediated gene flow in natural plant populations. The tracking of pollen movements that lead to the successful fertilization of ovules provides an estimate of the 'realized' gene flow. Paternity analysis requires the sampling of all males (hereafter referred to as 'potential pollen donors') within a defined area and a sample of seeds collected from fruiting trees acting as pollen traps (hereafter 'mother trees') from the same area. In monoecious species, mother trees are also potential pollen donors. After genotyping all potential pollen donors, mother trees and collected seeds, the assignment of paternity can be carried out with various analytical methods whose pros and cons have recently been reviewed by Jones *et al.* (2010). The main aim of paternity analysis is to correctly identify the true father of any collected seed (or to detect immigrant pollen when no local pollen donor is compatible with seed genotype). Results from paternity analysis (pollen immigration rate, distribution of male reproductive success and estimates of pollen dispersal kernel parameters) are thought to be affected by the resolution of the marker set used and by genotyping errors (Burczyk *et al.* 2006, Bacles and Ennos 2008).

Paternity analysis is a powerful tool for the study of within-population pollen dispersal patterns. The short distance component of the dispersal pattern has strong influence in shaping fine-scale genetic structure, that in turn determines the rate and direction of microevolutionary changes at the population level (Pluess *et al.* 2009). In isolated and low density populations, paternity analysis allows one to trace pollination events at a larger scale (the long distance component of the dispersal pattern). It is well-established that forest tree pollen is able to travel hundreds of kilometres and evidence is accumulating that after such long distance dispersal events pollen is viable and can successfully fertilize seeds (Williams 2010, Robledo-Arnuncio 2010, Buschbom *et al.* 2011). The possibility of quantifying effective pollination

over long distance has profound consequences on the study of how genes have travelled in past and ongoing tree migrations, and contributes to the sound forecasting of tree responses to anthropogenic and natural global changes (Savolainen *et al.* 2007). In addition, risk assessment of pollen escape from GM plantation and predictions on the spread of invasive alien species strongly depend on an accurate estimate of the long distance component of pollen dispersal kernel (Williams 2005).

The impact of sampling scheme on the results from paternity analysis has received relatively little attention. A few studies assessed the effects of location and number of seed traps and number of seeds collected in each trap in classical seed trapping experiments. Data from these studies are usually analyzed following a backward approach (so-called 'inverse modeling') aimed at reconstructing the dispersal kernel from the spatial location and the fecundity of potential parents and the spatial pattern of seeds collected from traps (Ribbens *et al.* 1994). Skarpaas *et al.* (2005) used a simulation approach to optimize seed trap sampling design around a point source. They showed how traps arranged in transects and sectors provide a better kernel estimation than other sampling schemes. Annuli and grid arrays outcompeted other schemes only when the anisotropy of dispersal was unknown. Pielaat *et al.* (2005) found that a trade-off between nearby and distant sampling is needed to accurately characterize the tail of the dispersal kernel. It is also well known that a random placement of traps within a rectangular or circular area determines an uneven sampling of distance classes, leading to the over-representation of the intermediated ones (Ghosh 1951).

So far, the most relevant study whose results on sampling effort can be easily extended to the case of paternity analysis is the one by Robledo-Arnuncio and Garcia (2007). They proposed a maximum-likelihood procedure to estimate the seed dispersal kernel from the exact identification of seed sources, as in the case of parentage assignment based on genetic compatibility. This method (Competing Sources Model, CSM) works out the problem of the uneven distribution of mother-trap potential distances by taking into account the spatial arrangement of seed traps relative to source plants. It provides better estimates of seed dispersal kernel parameters compared to standard maximum likelihood fitting used in inverse modeling (Robledo-Arnuncio and Garcia 2007). Jones and Muller-Landau (2008) compared different approaches to estimate dispersal kernel parameters and confirmed that CSM is among the most accurate and robust methods. From the results of simulations on different sampling scenarios in Robledo-Arnuncio and Garcia (2007) some practical recommendations on sampling effort emerged. For example, it was shown that fewer seeds are required to properly estimate the average dispersal distance (hereafter 'δ') with respect to the shape

parameter (hereafter '*b*') of the exponential power kernel (a widely used and flexible curve to describe seed and pollen dispersal), and that for a fixed total number of seeds increasing the number of traps is more useful than collecting more seeds per trap for reducing estimation uncertainty.

Sampling for paternity analysis differs substantially in the number of traps from the one for inverse modelling. In paternity analysis a large number of mother trees per population is rarely sampled (see for instance Schuster and Mitton 2000 and Oddou-Muratorio *et al.* 2005 for some exceptions), whereas in inverse modelling 10 to 300 traps are usually placed (Stoyan and Wagner 2001, Pairon *et al.* 2006, Jones *et al.* 2008). However, quantitative assessments or qualitative indications about the spatial arrangement of mother trees and the total number of seeds, mother trees, and seeds per mother tree can seldom be found in the literature on paternity analysis. An inadequate or insufficient sampling effort (too few sampled seeds and/or mother trees) can lead to a biased estimation of the within-population pollination patterns. In addition, the lack of standard sampling methods limits the comparison among studies to draw general conclusions.

In the present work I review 92 paternity analysis papers to provide a quantitative assessment of the sampling strategy upon which the estimate of pollen-mediated gene flow rate, the reconstruction of pollen dispersal kernel and the description of male reproductive success distribution are based. I report data on the sampling effort (the total number of sampled seeds, the number of mother trees and the number of seeds per mother tree) and discuss possible consequences of limitations in the sampling scheme on paternity analysis results.

2.2 Materials and methods

I searched for published studies that used paternity analysis to estimate pollen-mediated gene flow in forest trees. I used 3 different databases (Google scholar™, ISI Web of Science™ and Scopus™) for surveying the literature. The key-words used were: paternity analysis, tree*, pollen, genetic* and gene flow. I also tracked references within the articles found, from review papers on gene flow in forest trees (Burczyk *et al.* 2004, Ashley 2010), from Table 4 in Bittencourt and Sebbenn (2007) and Appendix A-1 in Wang *et al.* (2010a). I did not include studies based on mating models (as MLTR, Ritland 2002) and on pollen pool heterogeneity (as KINDIST, Robledo-Arnuncio *et al.* 2007) because they require only to genotype seeds (and mother trees), and studies based on indirect methods for the estimation of gene flow, in which gene flow is inferred from the spatial genetic structure or F_{ST}. I did not consider a few

studies for which I was unable to obtain the full-text. Since the focus of our paper is on the sampling strategy, studies based on previously published data were also excluded in order to avoid duplicates. I included paternity studies carried out in seed orchards and studies investigating gene flow among closely related species (e.g *Quercus* spp).

From each paper I recorded:

- information on sampling strategy: the total number of sampled seeds, the number of mother trees and the average number of seeds per mother tree;
- characteristics of the studied population: number of potential pollen donors, number of potential pollen traps (female individuals in the population for dioecious species), area, tree density;
- the studied species, its family and taxonomic group, breeding system and primary pollination vector;
- the method and molecular markers used for paternity assignment.

In monoecious species the number of male individuals is equal to the number of female individuals. Whenever life history traits could not be found in the text, I tried to gather them from other sources (e.g. on-line databases), in some cases from personal communications with the Authors. When density was not available in the text it was estimated dividing the number of individuals within the study population for the stand area. Papers that provided a poor description of the sampling design or, in general, that lacked many essential data, were excluded from our dataset.

In papers where more than one stand was sampled and/or described, as it is typically done when a system of several forest fragments is studied, I collected data for every stand when at least one seed was genotyped in order to estimate the pollen-mediated gene flow characterizing that stand (or a group to which the specific stand belongs as, for instance, in Lander *et al.* 2010). In studies where the same stand was analyzed in two or more consecutive years, if data on sampling effort were reported for each year and they differed, I considered each year as a data point.

When the studied stand was a seed orchard and when the studied species reproduced vegetatively the number of individuals was the number of ramets rather than the number of genets, since each ramet represents a spatially distinct pollen source.

2.3 Results and Discussion

General contents and sources

I collected data from 92 papers published from 1992 to September 2011. Among them 14 were also present in Burczyk *et al.* (2004) and 27 in Ashley (2010). In the latter, the literature search was only on native plants (cultivated trees and crops were excluded) and microsatellite-based studies, but experiments based on parentage analysis were also taken into account. The author specified that her search was not exhaustive, but "broadly inclusive and representative".

The number of papers published per year increased until 2008. This gradual increment was followed by a slight decrease in the following years. The general growing trend seems related to the increasing availability of microsatellite markers and to the development of methods based on maximum likelihood assignment of paternities (Fig. 2.1). The most used methods (simple exclusion, neighbourhood model, and maximum likelihood) have similar sampling requirements. The main difference is that in simple exclusion and maximum likelihood paternity studies all potential pollen donors within the sampling area must be sampled whereas in neighbourhood model-based paternity studies all potential pollen donors within circular areas of a given radius surrounding *n* mother trees must be sampled. Few studies performed paternity analysis using multiple analytical methods (n=8, 9%). The use of different approaches for the estimation of gene flow rates has been recently proposed to overcome possible drawbacks related to specific methods and/or weaknesses due to low-resolution marker sets (Bacles and Ennos 2008, Jones *et al.* 2010).

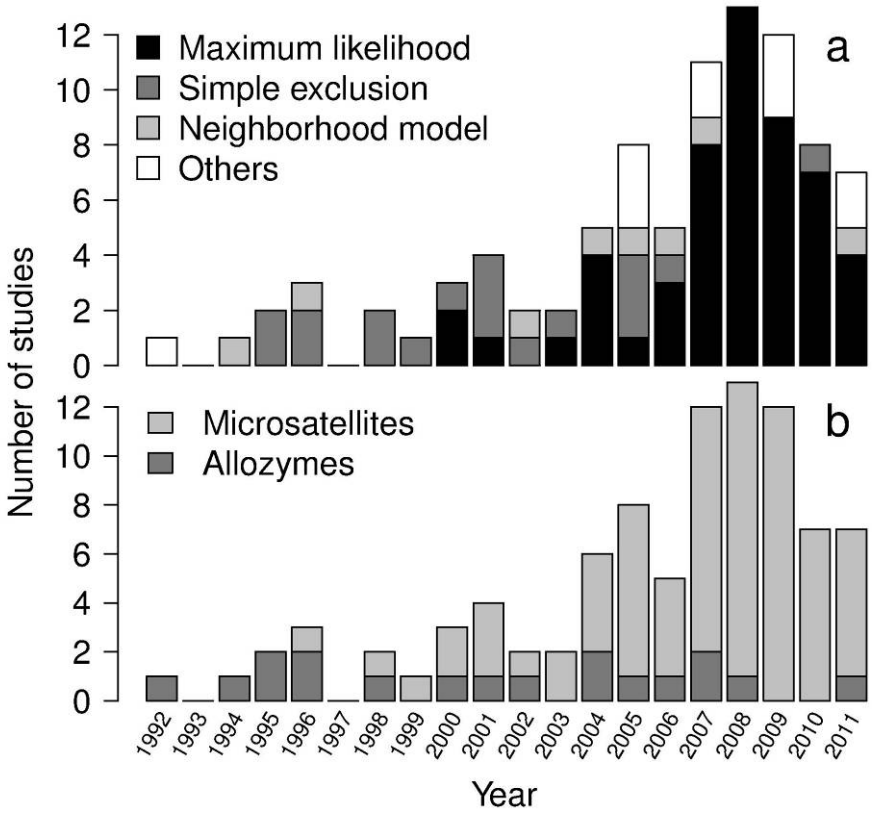

Figure 2.1: Number of published studies per year. Different colors represent (a) different methods in the analysis of paternity data, and (b) different molecular markers. In papers based on maximum likelihood methods, data analysis was usually performed by the CERVUS (Kalinowski *et al.* 2007) and FaMoz (Gerber *et al.* 2003) programs. In papers based on the neighbourhood model, data analysis was usually performed according to the methods presented in Burczyk *et al.* (2002) (implemented in the NM+ program by Chybicki and Burczyk (2010a)) and Oddou-Muratorio *et al.* (2005).

Almost half the papers (47%) were published in only 4 journals, with Heredity and Molecular Ecology clearly representing the preferred landing place for gene flow studies based on paternity analysis (respectively, 15% and 14%), followed by Forest Ecology and Management and Conservation Genetics (Fig. 2.2).

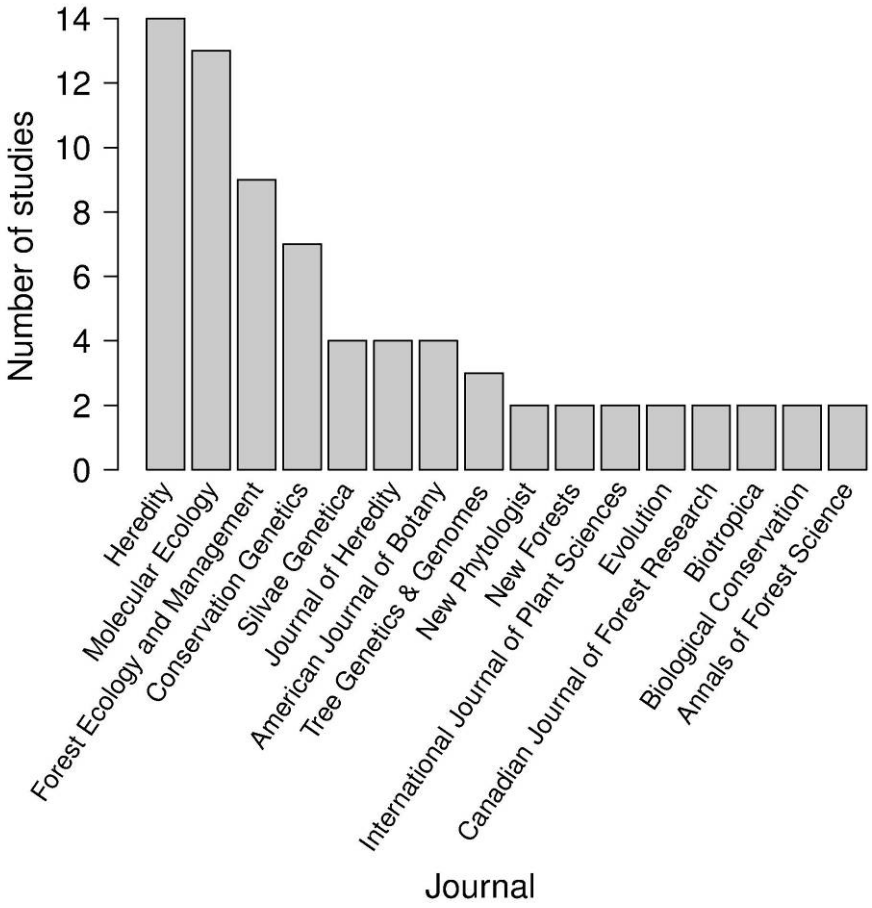

Figure 2.2: Number of studies published per journal. Journals with only 1 published study were excluded.

Studied species and sampling areas

The collected papers investigated gene flow in a total of 81 different species. Most studies (n=75, 81%) were conducted on Angiosperms whereas only 17 studies focused on Gymnosperms. Fagaceae (16) and Pinaceae (12) were the most sampled families and, together with Fabaceae and Dipterocarpaceae, represented the studied species in almost half of the collected papers. Overall, species from 30 different families were studied. The large number of gene flow papers on Fagaceae was also reported by Ashley (2010). With regard to the primary pollination vector, there were 46 studies on insect-pollinated species (50%), 38 on wind-pollinated species (41%) while in the remaining 8 articles the studied species relies on mammals, birds, or multiple vectors for pollination.

I obtained 187 data points from the 92 collected papers (see Appendix 1). Fifty-eight papers (63%) investigated a single stand providing a single data point. Papers with more than one data point were fairly frequent (n=34, 37%). In general, this is due to the analysis of multiple stands within the same article. Wang *et al.* (2010b) studied pollen-mediated gene flow in 28 fragments of *Pinus tabulaeformis* in an urban landscape. According to our criteria I retained

25 data points, the highest number of data points from a single study in our dataset. Also Lander *et al.* (2010) sampled a large number of forest fragments to estimate pollen immigration, but only a half of them matched our criteria, providing 13 data points. On the other hand, 11 papers investigated gene flow in the same stand but in multiple years (usually, in 2 or 3 consecutive years). Irwin *et al.* (2003) highlighted that single-season studies may not capture temporal variability in pollen exchange, especially in perennial plants where flowering does not occur every year. Therefore, multi-year analyses were advocated for obtaining accurate estimates of pollen-mediated gene flow patterns.

In general, our knowledge on pollen-mediated gene flow for a species is based on a single study. Only 14 species were the subject of more than one paper. *Quercus robur* was investigated in three papers, while *Araucaria angustifolia*, *Cryptomeria japonica*, *Eucalyptus grandis*, *Eurycorymbus cavaleriei*, *Fagus sylvatica*, *Picea abies*, *Populus nigra*, *Prunus avium*, *Pseudotsuga menziesii*, *Q. macrocarpa*, *Q. salicina*, *Shorea leprosula* and *Sorbus torminalis* were studied twice. The lack of independent estimates together with the usually low degree of comparability among pollen-mediated gene flow studies makes generalization on single estimates far-fetched. Studies designed for allowing meaningful comparisons among gene flow rates estimated in different ecological conditions are also rare. The importance of such comparative studies in characterizing the pollen dispersal capability of a species is discussed in Piotti *et al.* (2012).

The median area of the 118 stands for which I found sufficient information was 7.42 ha. Small stands (≤ 1 ha) are common (24%). Two thirds were smaller than 20 ha, whereas larger stands (≥ 100 ha) are rare (8%). This pattern is likely determined by the rapid increase with the studied area in sampling and genotyping effort. Studies on large areas are ideal to detect rare long distance dispersal events, but they are feasible only when species are present at low density. However, Hardy (2009) pointed out that, despite their great importance, the measures of dispersal obtained in such studies might not be representative of species with similar pollination syndrome at higher densities. Studying two natural *Populus trichocarpa* stands that dramatically differed in density (993 vs. 0.2 males/km^2) and area (19.6 vs. 31400 ha), Slavov *et al.* (2009) found large differences in pollination patterns, although the authors themselves warned about the difficulty of comparing such different areas. Piotti *et al.* (2012) compared two close *Fagus sylvatica* stands with regular densities characterized by different management regimes. They found a more skewed pollen dispersal distance distribution in the managed area whose density is about one-third of the unmanaged one (57 vs. 163 trees/ha). Data on the size of sampling areas can also be useful to understand if limits in sampling scale

may downwardly bias dispersal range estimates, even though the heterogeneity of experimental setups and methods used for data analysis as well as variation in population densities, pollination syndromes, pollen terminal velocity, stand isolation, etc. should be carefully taken into account in data collection and analysis for future works on this topic. However, methods to estimate dispersal parameters taking advantage of spatially censored data (Jones *et al.* 2005) or assuming the immigration rate to be a function of dispersal kernel (Goto *et al.* 2006) are already available but rarely applied. These approaches allow one to take into account immigration events of unknown origin in the estimation of the dispersal curve and this usually result in a substantially higher mean dispersal distance (Piotti *et al.* 2009, Chybicki and Burczyk 2010b).

Sampling strategy

Among the collected papers I found neither exhaustive justifications for the sampling strategy adopted, nor references to any guideline for sampling strategy. An exception was the paper by Oddou-Muratorio *et al.* (2005), that states that "the objectives for both years were to sample all possible distances between mother trees, and to maximize the number of mother trees in the middle part of the study site". A paragraph in their discussion focused on the methodological insights for the estimation of the dispersal kernel. Many papers reported a map of the sampling area showing the location of adult individuals with mother trees indicated by different symbols (e.g. Oddou-Muratorio *et al.* 2003, Nakanishi *et al.* 2004, Curtu *et al.* 2009). From the visual inspection of these maps no clear patterns in the choice of mother trees can be recognized. Their location varied from clustered in the centre of the stand to scattered throughout the entire sampling area. I did not find any statement about a random choice of mother trees in the collected papers. Given this lack of indications and our experience on the difficulties in sampling seeds from forest trees, I feel that such diverse sampling schemes might arise from practical constraints rather than from a thorough *a priori* evaluation. As pointed out in methodological papers on seed dispersal modeling, intuitively designed experiments are likely to lead to incomparable and non-representative results (Willson 1993, Stoyan and Wagner 2001).

Besides the spatial layout of mother trees, the other crucial factor in designing a solid sampling scheme for paternity analysis is the sampling effort, represented by the total number of sampled seeds. This, in turn, is the product of the number of mother trees and the number of seeds per mother tree. In the collected papers, I found a median number of 8 mother trees (mean = 11.3 ± 13.6 SD, Fig. 2.3), 29 seeds per mother tree (mean = 45.6 ± 74.7 SD, Fig. 2.3)

and 240 total sampled seeds (mean = 356.1 ± 423.1 SD, Fig. 2.4a). These values are lower than those usually found in classical seed trapping studies. This difference is likely to depend on the need for larger samples to infer dispersal kernel's parameters without genetic data and, on the other hand, on the significant investment of time and resources to collect genetic data (Jones and Muller-Landau 2008).

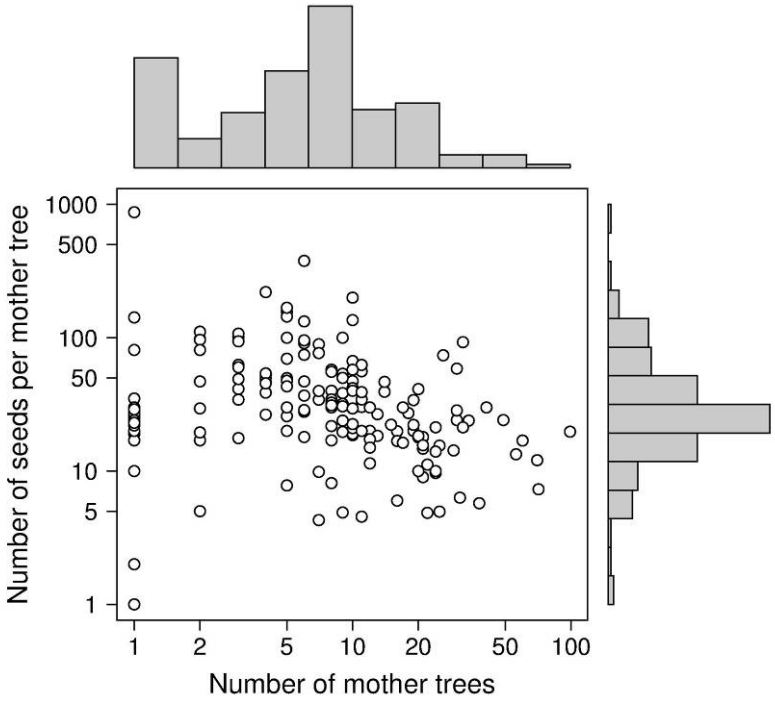

Figure 2.3: Distributions of the number of mother trees (x-axis) and number of sampled seeds per mother tree (y-axis) in log scale. White dots represent the 187 data points recorded.

As a measure of the coverage of potential pollen traps in a stand (trap coverage) the ratio of mother trees over the total number of female trees was calculated. The median value of trap coverage was 0.18 (mean = 0.28 ± 0.27 SD, Fig. 2.4b). It is not correlated with tree density, but it negatively depends on the total number of female trees, indicating that trap coverage is more exhaustive in small populations. The low trap coverage in large stands can be related to practical and economic limitations. Nevertheless, a sufficiently high trap coverage is desirable to decrease confidence interval of parameter estimates and increase comparability among results from different experiments.

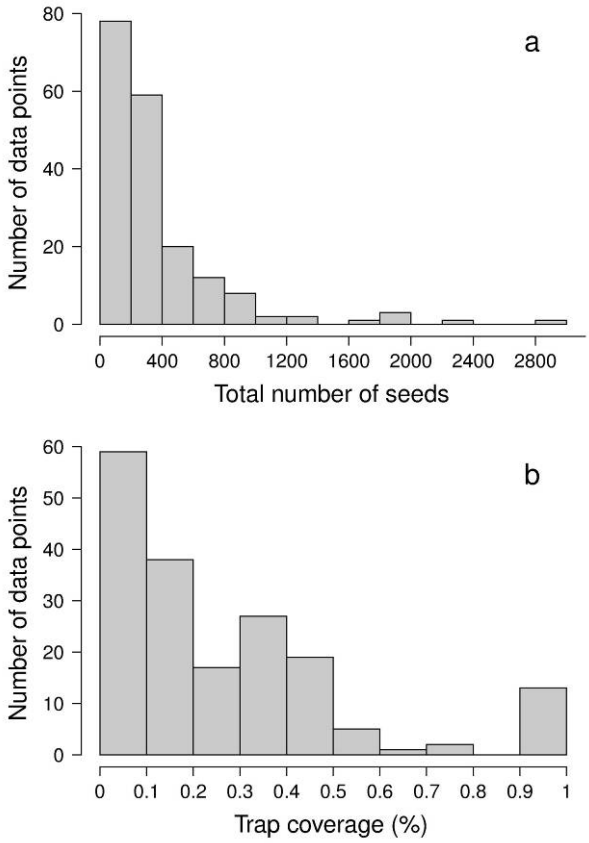

Figure 2.4: Distributions of (a) the total number of seeds (i.e. the product of the number of mother trees and the number of sampled seeds per mother tree), and (b) the percentage of trap coverage (i.e. the ratio of mother trees over the total number of female trees) characterizing the 187 data points.

My data on the distribution of sampling effort in published paternity analyses are comparable with the ones from Robledo-Arnuncio and Garcia (2007), even though it should be stressed that their work focused on studying seed dispersal with seed trapping and genetic data. By using a simulation approach, they tested the performance of the CSM by randomly placing 20, 100 and 200 traps in a squared area, and sampling 1 to 50 seeds per trap (resulting in a total number of sampled seeds between 200 and 1000). They found that the CSM performs well in estimating the δ parameter even for a relatively small number of seeds (200), whereas $\geq$ 500 seeds are needed to obtain an accurate estimate of the b parameter of the dispersal kernel. As the Authors noted, the b parameter is more sensitive than the δ parameter to decreasing number of total seeds and seed traps. The minimum number of traps they simulated was 20, that is low for classical seed trapping experiments (being therefore adequate for the aims of their paper), but almost double the mean number of mother trees in paternity studies. Among the few indications in the literature about sampling effort in paternity analysis, Oddou-Muratorio *et al.* (2005) noted that increasing the number of mother trees from 14 to 60

(in the same area in two consecutive years) sensibly reduces the confidence interval of the parameters of the dispersal kernel. On the other hand, the authors pointed out that sampling a high number of seeds per mother tree, that usually limits the number of mother trees, could be more adequate for the estimation of individual selfing rates. This implies that sampling strategy in paternity analysis should be fine-tuned to meet the specific aims of an experiment.

2.4 Conclusions

Although some results on the consequences of different sampling schemes are available for seed trapping studies (with or without genetic assignment of seeds), the case of paternity analysis, usually based on a lower sampling effort, is poorly investigated. My data collection from the literature on paternity analysis in forest trees showed a potential lack of knowledge about the effects of low numbers of mother trees, seeds per mother tree and total sampled seeds on estimates usually performed to describe within-population pollination patterns: *i*) pollen immigration, *ii*) male reproductive success, and *iii*) parameters of the pollen dispersal kernel. Only in 29 out of the 187 collected data points (15%) the number of mother trees is higher than 20, the lowest number of traps taken into account by Robledo-Arnuncio and Garcia (2007). This means that for 85% of our collected data points we have little idea about how accurate and precise the estimates from paternity analysis can be. From Table 2 in Robledo-Arnuncio and Garcia (2007) we know that the relative root mean square error (RMSE), a measurement of both accuracy and precision of an estimate normalized to the expected value, is ~0.04 for the δ parameter and ~0.10 for the b parameter when 500 seeds were sampled. RMSEs increased to ~0. 07 and ~0.17, respectively, when the total number of seeds decreased to 200. Errors roughly increase with the inverse of the square root of the total number of seeds sampled, as expected from classical statistical theory. Consequently, when the sampling effort is scarce non negligible errors in estimates can be expected, in particular for the *b* parameter. Leonarduzzi *et al.* (submitted), relying on the distribution of sampling effort presented here, explore the consequences of realistic sampling strategies on the reconstruction of different dispersal kernels to provide the basis for meaningful guidelines for paternity analysis.

Chapter 3:

The effect of sampling design on the estimate of pollen dispersal patterns in paternity studies: a simulation-based study

3.1 Introduction

Pollen dispersal is a major step of the reproductive cycle of plants and represents the main mechanism for gene flow over long distances (Kremer *et al.*, 2012). The potential for long distance dispersal (LDD) via pollen strongly influences evolutionary processes both within and among forest tree populations (Nathan, 2006). A detailed understanding of pollen dispersal dynamics is key to the comprehension of genetic structure from the local to the bio-geographic scale (Robledo-Arnuncio, 2012). The study of pollen gene flow is currently of particular importance to track gene escape from transgenic plantations (Williams, 2007), to monitor the spread of invasive species through cross-pollination with native ones (Brown and Mitchell, 2001), and to evaluate the genetic consequences of habitat fragmentation (Bacles and Jump, 2011).

Studying dispersal in plants, whether occurring by pollen or seeds, is a challenging task especially in natural populations (Jordano, 2007). One of the main goals of dispersal studies is to describe dispersal patterns by reconstructing seed and pollen dispersal kernels (Nathan and Muller-Landau, 2000). A variety of methods has been developed for this purpose, which can be grouped into two main categories: *i*) directly tracking the trajectory of each propagule from the source to its final position (e.g. Van Rossum *et al.*, 2011), *ii*) reconstructing the dispersal kernel from the pattern of dispersed propagules observed arranging a number of traps in a sampling area (e.g. Ribbens *et al.*, 1994). The former, despite providing robust estimates independent from sampling design, are logistically hard to set up in natural populations. The latter are more easily implementable, but the inferred dispersal kernel might be highly sensitive to sampling design. Scarce sampling efforts may lead to poor dispersal kernel reconstructions, while not considering the distribution of source-trap distances may cause some dispersal distances to be over- or undersampled. In particular, short and medium distances are often oversampled with respect to longer ones (Jones *et al.*, 2005). These sampling issues can indeed lead to biased dispersal kernels (Robledo-Arnuncio and Garcia,

2007). In addition, estimating dispersal kernel parameters from trapped propagules originally suffer from the difficulty of identifying their exact source.

This issue has been overcome by adopting genetic methods, where the propagule source can be identified on the basis of genetic compatibility between trapped propagules and potential parents (Jones *et al.,* 2010). Paternity analysis is the preferred method to directly estimate pollen gene flow and dispersal kernel in tree populations (Hadfield *et al.,* 2006; Ashley, 2010). Paternity analysis can be considered a trapping experiment where pollen traps are represented by mother trees from which seeds are sampled. Therefore, resulting estimates can be biased by the experimental design as well (Robledo-Arnuncio and Garcia, 2007; Jones and Muller-Landau, 2008).

Despite the increasing use of paternity analysis, little attention has been given to the effect of sampling design (Leonarduzzi *et al.,* 2012; Nathan *et al.,* 2012). Few studies have explored its consequences on gene flow and pollen dispersal kernel estimates (e.g. Klein and Laredo, 1999; Klein *et al.,* 2006; Sharma and Khanduri, 2007). More effort has been devoted to simulation-based studies investigating the effect of sampling design on the reconstruction of seed dispersal kernels (Stoyan and Wagner, 2001; Skarpaas *et al.*, 2005; Pielaat *et al.,* 2006; Robledo-Arnuncio and Garcia, 2007). The gathered information can be easily transferred to experiments on pollen dispersal after taking into account the obvious differences between the two experimental approaches (e.g. the usually lower number of traps in pollen dispersal experiments). Among simulation-based studies, the most relevant in providing sampling recommendations for genetic-based dispersal experiments is the one by Robledo-Arnuncio and Garcia (2007), where the performance of different models was assessed at different sampling efforts. They showed that the Competing Sources Model (CSM), which takes into account the relative spatial distribution of sources and traps, is effective in estimating kernel parameters. Interestingly, accuracy and precision of parameter estimates remained almost constant between 1000 and 500 sampled seeds, but rapidly decreased when reducing the total number of seeds from 500 to 200. This indicates an L-shaped relationship between errors and the number of sampled seeds, suggesting that decreasing the sampling effort after a certain threshold can produce large errors.

In their review on paternity studies in forest trees, Leonarduzzi *et al.* (2012) (i.e. Chapter 2 of this dissertation) found that sampling effort and spatial layout were highly heterogeneous and that the rationale behind sampling strategy was almost never explained (but see Oddou-Muratorio *et al.*, 2005). In the reviewed studies, the median number of total sampled seeds was 240 (approximately 8 mother trees and 29 seeds per mother tree). Sampling effort

allocation ranged from 874 seeds sampled from 1 single mother tree (Lian *et al.*, 2001) to an average of 7.3 seeds sampled from 71 mother trees (Schuster and Mitton, 2000). This heterogeneity is only partially justified by the different aims that paternity studies can have. Heterogeneity in sampling design can lead to two major issues. First, studies based on different sampling designs do not produce equally accurate results. Second, comparability among studies (and among plots within a single study) is inevitably low. A large number of paternity studies is based on less than 200 seeds (41%) and 20 mother trees (84%, Leonarduzzi *et al.*, 2012) which are, respectively, the lowest number of seeds and the lowest number of mother trees taken into account by Robledo-Arnuncio and Garcia (2007). This means that accuracy can be evaluated for very few paternity studies, and that in many other studies errors are likely to be substantially larger than the ones found by Robledo-Arnuncio and Garcia (2007).

Given the relevance of obtaining adequate estimates of pollen dispersal, this work aims at evaluating by simulations the consequences of sampling effort on the reconstruction of pollen dispersal patterns in a wide range of dispersal scenarios. In particular, this study addresses the following questions: *i*) How precise and accurate are the estimates (i.e. pollen immigration rate and dispersal kernel parameters) when using commonly adopted sampling efforts? *ii*) Are the number of mother trees and the number of collected seeds per mother tree equally important? Does one prevail over the other in providing adequate estimates? *iii*) Do the same sampling requirements apply for species with different dispersal characteristics?

3.2 Materials and methods

Simulation of pollen dispersal and seed sampling

The effect of increasing sampling efforts on the reconstruction of pollen dispersal patterns was tested by using the simulation approach presented in Robledo-Arnuncio and Garcia (2007). This approach is based on the CSM and on the categorical assignment of paternity with the assumptions of no genotyping errors and unambiguous pollen sources identification. These assumptions allow to isolate the effect of sampling effort from other confounding factors such as the ones related to low exclusion probabilities and genotyping errors. The CSM is suitable for our aims because it provides accurate estimates of kernel parameters by taking into account also the effect of the relative spatial distribution of sources and traps (Jones and Muller-Landau, 2008). The simulation procedure was implemented in R (R Development Core Team, 2012).

The main steps of the simulation are:

1) *Simulating the population*. Ten thousand individuals are arranged within a 100 × 100 area (mean expected density = 1 individual/squared unit) randomly drawing X and Y coordinates from a flat distribution. A 20 × 20 central area is selected as the sampling plot, where 20 is the plot side *l* (Fig. 3.1). Simulated individuals outside the sampling plot within the 100 × 100 area represent the 'background population'.

2) *Simulating pollen traps*. According to the sampling scenario (see the *Pollen dispersal and sampling scenarios* section), a number of mother trees (*t*) acting as pollen traps is placed within the sampling plot randomly drawing X and Y coordinates (Fig. 3.1). In Robledo-Arnuncio and Garcia (2007), traps were 0.1 × 0.1 squared cells, whereas here traps are simulated as dimensionless points. Pollen traps can represent either mother trees or physical traps for pollen.

3) *Simulating pollen dispersal*. A number of seeds (*s/t*) is sampled from each pollen trap. Following the approach used in Robledo-Arnuncio and Garcia (2007), the paternal origin of seeds in each pollen trap is randomly drawn from a multinomial distribution. The multinomial distribution describes the probability of each potential pollen donor to reach a specific pollen trap according to their distance and the true dispersal kernel chosen. An isotropic dispersal process is simulated. The same pollen dispersal kernel and fecundity are assumed for all potential pollen donors.

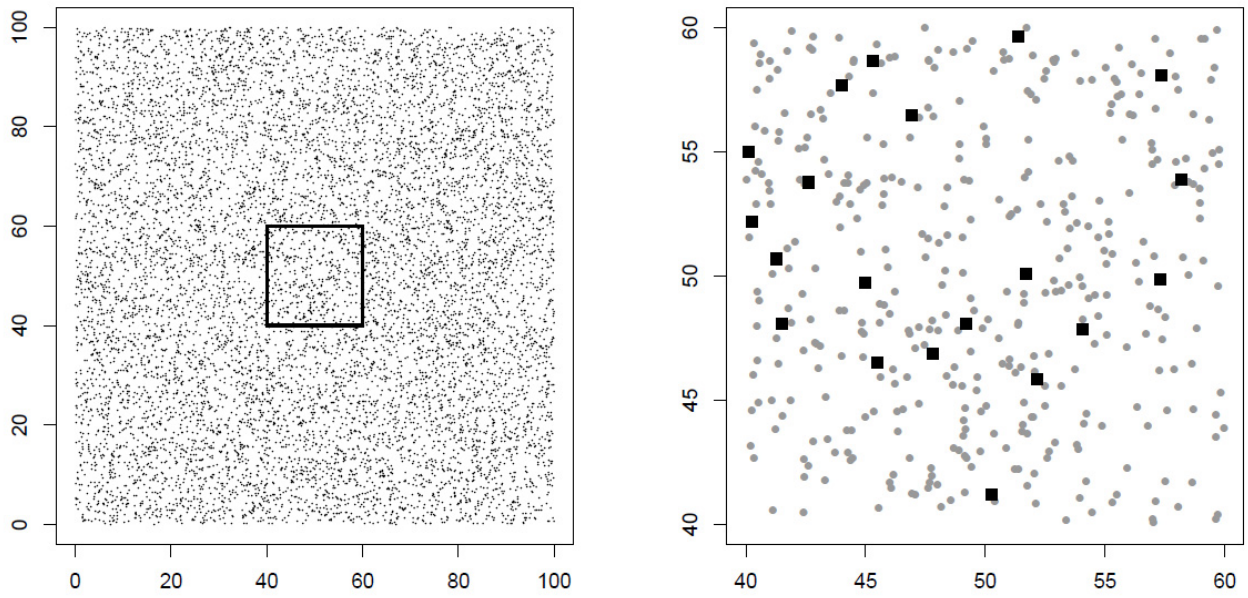

Figure 3.1: On the left, the background population (100 x 100) is shown. Individuals, represented as black dots, are uniformly distributed. The sampling plot (20 x 20) in the centre is bordered by the black line. On the right, the sampling plot is shown in detail. Mother trees (i.e. pollen traps) are represented as black squares, whereas

individuals (potential pollen donors) are represented by grey dots.

Pollen dispersal and sampling scenarios

A total of 1920 pollen dispersal and sampling scenarios, resulting from the combination of 32 pollen dispersal scenarios (PDSs) and 60 sampling scenarios (SSs), were used (Fig. 3.2). Simulations were carried out dispersing pollen according to the PDSs described below. For each PDS, the performances of all SSs in estimating pollen dispersal parameters were evaluated. For every combination of PDS and SS, 1000 independent simulations were run.

PDSs derive from pollen dispersal simulated using 3 families of two-dimensional dispersal kernels generally used to describe pollen and seed dispersal curves: exponential power and 2Dt families (as formulated by Austerlitz *et al.*, 2004), and the Weibull family (as formulated by Goto *et al.*, 2006) (Table 3.1). These dispersal kernels are characterized by two parameters: the scale parameter a and the shape parameter b. The shape parameter is usually thought to determine the fatness of the kernel tail (Austerlitz *et al.*, 2004; Nathan *et al.*, 2012). The values of a and b are specific for each family of dispersal kernel, so they are not directly comparable among different families. In contrast, the 1st moment (d, mean dispersal distance) and the 99th percentile (p_{99}) of the dispersal kernel can be compared among different families. The latter two quantities depend on both a and b (Table 3.1). The parameters a and b characterize the mathematical properties of the dispersal kernel, while d and p_{99} are directly referable to key features of the dispersal process. In particular, p_{99} can be used as a proxy for describing LDD (Nathan, 2006). For each kernel family, I used different combinations of parameters leading to 32 PDSs that span from extremely limited dispersal (i.e. $p_{99} = 2.075$, with a ratio $l/d = 20$ indicating that dispersal is extremely limited even when compared to sampling area) to long distance dispersal (i.e. $p_{99} = 81.96$, $l/d = 2$) (Table 3.1). For each family I used different b values in order to have both thin-tailed and fat-tailed curves (except for 2Dt, which is always fat-tailed). For all kernel families, specific combinations of a and b leading to the same mean dispersal distances ($d = 1, 3, 5, 10$) were used to allow comparability.

Table 3.1: The characteristics of the 32 PDSs used to simulate pollen dispersal: the formulas for each curve family, the shape parameter (*b*), the mean dispersal distance (*d*), the scale parameter (*a*), the 99th percentile of each PDS and the rate of pollen immigration are reported.

Dispersal kernel family	***b***	***d***	***a***	**99th percentile**	**Pollen immigration rate**
2d-exponential power $p_e(r;a,b)=\frac{b}{2\pi\pi^2\Gamma\left(\frac{2}{b}\right)}\exp\left(-\left(\frac{r}{a}\right)^b\right)$ $d_e(r;a,b)=a\left(\frac{\Gamma(3/b)}{\Gamma(2/b)}\right)$	0.5	1	0.05	5.045	0.094
	0.5	3	0.15	15.135	0.210
	0.5	5	0.25	25.225	0.309
	0.5	10	0.5	50.451	0.486
	1	1	0.5	3.318	0.086
	1	3	1.5	9.957	0.212
	1	5	2.5	16.595	0.324
	1	10	5	33.191	0.540
	2	1	1.12	2.42	0.083
	2	3	3.38	7.263	0.213
	2	5	5.64	12.106	0.329
	2	10	11.28	24.214	0.565
2d-Weibull $p_w(r;a,b)=\frac{b}{2\pi\pi^2}\left(\frac{r}{a}\right)^{b-2}\exp\left(-\left(\frac{r}{a}\right)^b\right)$ $d_w(r;a,b)=a(\Gamma(1+1/b))$	0.75	1	0.5	5.734	0.086
	0.75	3	1.5	18.276	0.186
	0.75	5	4.20	30.97	0.269
	0.75	10	8.40	62.876	0.404
	1.5	1	1.11	3.065	0.068
	1.5	3	3.32	9.197	0.177
	1.5	5	5.54	15.33	0.283
	1.5	10	11.08	30.661	0.503
	2.5	1	1.13	2.075	0.065
	2.5	3	3.38	6.227	0.179
	2.5	5	5.63	10.379	0.289
	2.5	10	11.27	20.76	0.531
2Dt $p_{2Dt}(r;a,b)=\frac{b-1}{\pi a^2}\left(1+\frac{r^2}{a^2}\right)^{-b}$ $d_{2Dt}(r;a,b)=a\left(\frac{\Gamma(3/2)\Gamma(b-3/2)}{\Gamma(b-1)}\right)$	1.6	1	0.18	8.198	0.098
	1.6	3	0.53	24.589	0.133
	1.6	5	0.88	40.982	0.151
	1.6	10	1.77	81.965	0.277
	2.5	1	1	4.532	0.087
	2.5	3	3	13.597	0.208
	2.5	5	5	22.662	0.313
	2.5	10	10	45.325	0.514

SSs were chosen to exhaustively cover the range of sampling efforts commonly used in published paternity studies, according to the review by Leonarduzzi *et al.* (2012) (Fig. 3.2). Overall, 60 SSs characterized by different combinations of *t* and *s/t* were selected (Fig. 3.2, Table 3.2). The total number of sampled seeds (*s*) ranged between 20 and 2000, which are reliable limits for paternity studies (Leonarduzzi *et al.,* 2012). Among the selected SSs, different combinations of *t* and *s/t* leading to the same total number of sampled seeds (i.e. *s* = 40, 80, 100, 200, 500, 1000, and 2000) were present (Table 3.2). These cases are particularly useful for investigating the relative importance of the two sampling effort components (*t* and *s/t*). Eight scenarios from Robledo-Arnuncio and Garcia (2007) were also included for comparison.

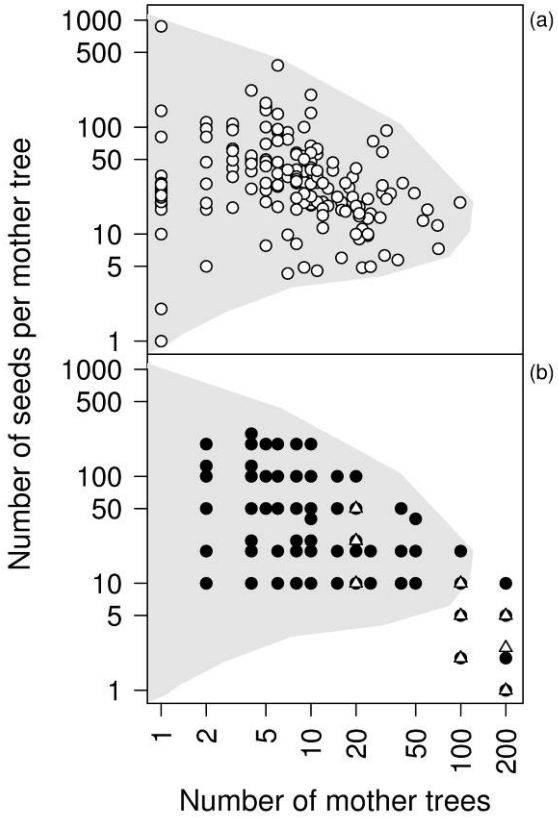

Figure 3.2: Number of mother trees (x-axis, log scale) and number of sampled seeds per mother tree (y-axis, log scale) representing: a) the 187 data points recorded in Leonarduzzi *et al.* (2012) summarizing the sampling effort in 92 paternity studies on forest trees, and b) the 60 sampling scenarios (SSs) investigated in this study (black dots). White triangles are SSs in common with Robledo-Arnuncio and Garcia (2007). In both figures, the grey area encloses sampling efforts reviewed by Leonarduzzi *et al.* (2012).

Table 3.2: The 60 sampling scenarios (SSs) used in the simulations. *s* = overall number of sampled seeds, *t* = number of mother trees (i.e. pollen traps) from which seeds are sampled, *s/ t*= average number of seeds sampled per mother tree.

s	***t***	***s/t***
20	2	10
40	2	20
40	4	10
60	6	10
80	4	20
80	8	10
100	2	50
100	4	25
100	10	10
120	6	20
150	15	10
160	8	20
200	2	100
200	4	50
200	8	25
200	10	20
200	20	10
200	100	2
200	200	1
250	2	125
250	5	50
250	10	25
250	25	10
300	6	50
300	15	20
400	2	200
400	4	100
400	8	50
400	10	40
400	20	20
400	40	10
400	200	2
500	4	125
500	5	100
500	10	50
500	20	25
500	25	20
500	50	10
500	100	5
600	6	100
750	15	50
800	4	200
800	8	100
800	40	20
1000	4	250
1000	5	200
1000	10	100
1000	20	50
1000	50	20
1000	100	10
1000	200	5
1200	6	200
1500	15	100
1600	8	200
2000	10	200
2000	20	100
2000	40	50
2000	50	40
2000	100	20
2000	200	10

Reconstruction of pollen dispersal parameters

After each simulation, kernel parameters were estimated by maximum likelihood using the CSM and the `optim` function in R. For each pair of $\hat{a}$ and $\hat{b}$ values the corresponding $\hat{d}$ and $\hat{p}_{99}$ were calculated. Pollen immigration rate was estimated as the proportion of seeds pollinated by the background population over the total number of sampled seeds.

For each PDS-SS combination, accuracy and precision of estimated parameters were assessed by computing the relative bias and root mean squared error (RMSE). To avoid possible inconsistencies introduced by outlier values I chose the winsorized mean and the median over 1000 simulations as robust statistics to represent central values. Winsorized means were calculated by trimming estimates at the 5th and 95th quantiles using the *psych* package in R. The relative mean (and median) bias was therefore computed as the winsorized mean (and median) of the differences between the point estimate and the true value, divided by the true value. Therefore, a relative bias of -0.5 is equal to an underestimation of 50% of the true value. The relative RMSE was computed as the squared root of the winsorized mean (and median) of the square difference between the point estimate and the true value, divided by the true value. Confidence intervals around median values were calculated using a bootstrap procedure with 1000 replicates. In the Results, relative median bias and RMSE instead of relative mean bias and RMSE are presented because in some PDS-SS combinations the misleading effect of outliers was not removed completely by trimming mean values (see also the "*Assessment of methodological reliability*" paragraph).

For each PDS, an additional set of simulations was carried out to precisely detect all immigrant gametes by sampling high numbers of mother trees (t = 500) and seeds per mother tree (s/t = 500). The mean pollen immigration rate over 1000 replicates was taken as the expected value and used to calculate the bias.

3.3 Results

Effect of sampling effort on kernel fitting

In the following, only results of the exponential power PDSs are reported. The main differences with other kernel families are outlined in the "*Notable differences in Weibull and 2Dt PDSs*" paragraph and in the Supporting Information.

The accuracy and precision of estimated dispersal kernels were strongly related to the total number of sampled seeds *s*. When *s* is low the global shape of the estimated dispersal kernel can be highly different from the true one and the probability of distances in the tail of the curve can be biased by several orders of magnitude (Fig. 3.3).

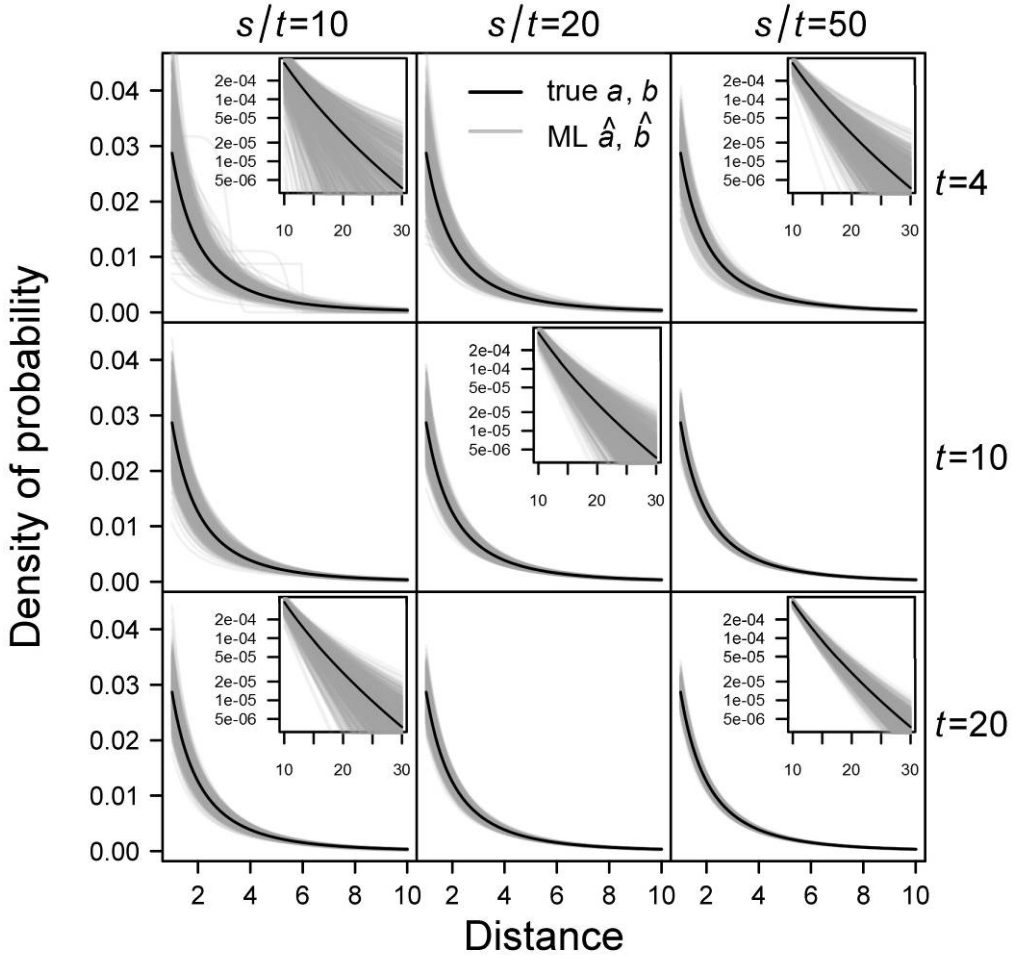

Figure 3.3: Comparison between true and inferred dispersal kernels obtained with 9 different sampling efforts characterized by increasing numbers of seeds sampled per mother tree (*s/t*, ranging from 10 to 50 along the columns) and increasing number of mother trees (*t*, ranging from 4 to 20) along the rows). In each panel, dispersal kernels inferred from 1000 simulations (grey lines) are reported and compared with the true one (black line, exponential power with $b = 0.5$ and $d = 5$). Inner windows focus on the curve tails (y-axes in log scale).

Bias and RMSE for all estimated parameters ($\hat{a}$, $\hat{b}$, $\hat{d}$, $\hat{p}_{99}$) decreased non-linearly as *s* increased (Fig. 3.4). When the bias for $\hat{a}$ and $\hat{b}$ was significantly different from 0 (i.e. confidence interval not overlapping 0: 36% of all cases for $\hat{a}$ and 48% for $\hat{b}$), bias was virtually always positive (99% for $\hat{a}$ and 100% for $\hat{b}$). In most PDSs, overestimation was low (< 0.1) until *s* drops below 200, while $\hat{a}$ and $\hat{b}$ were nearly unbiased for high values of *s* (Fig. 3.4).

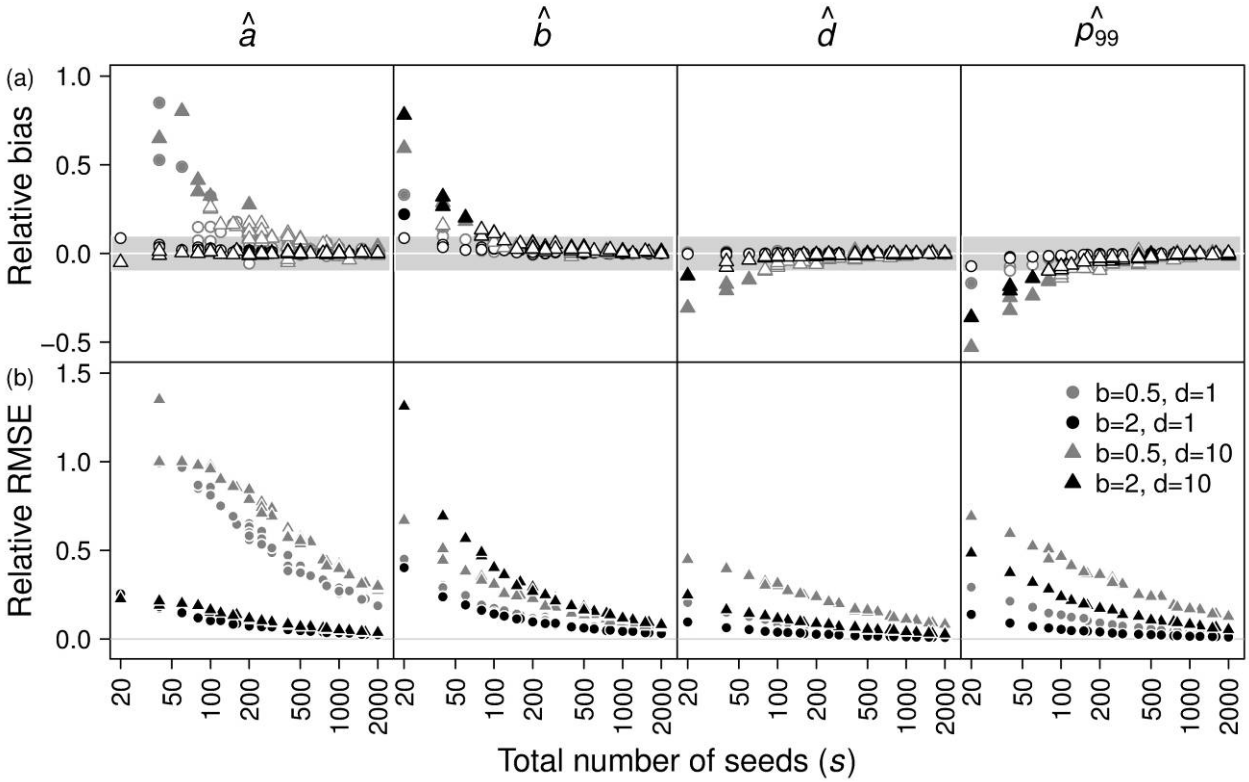

Figure 3.4: Relative bias (a) and RMSE (b) of dispersal kernel parameters ($\hat{a}$, $\hat{b}$, $\hat{d}$, and $\hat{p}_{99}$) with increasing total number of seeds (*s*, x-axis in log scale). The results for 4 different exponential power PDSs are reported with different combinations of colors and symbols as indicated in the legend. Filled-triangles and circles indicate biases significantly higher than 0.1 or lower than -0.1, whereas empty symbols correspond to values not significantly higher than 0.1 or lower than -0.1. Bias values outside the [-0.5;1] range and RMSE values outside the [0;1.5] range are not shown.

Errors in the estimation of *a* and *b* determined parameters *d* and p_{99} being underestimated in most PDS-SSs. Mean dispersal distance *d* was markedly underestimated (i.e. relative median bias significantly lower than -0.1) in 19% of cases while overestimated in only 1% of cases. The 99th percentile (p_{99}) was markedly underestimated in 49% of cases while overestimated in

only 1 PDS-SS combination. Similarly to $\hat{a}$ and $\hat{b}$, $\hat{d}$ and $\hat{p}_{99}$ were almost unbiased when $s > 100$, while the bias exponentially increased for lower sampling efforts (Fig. 3.4). Most cases where bias for $\hat{p}_{99}$ was significantly lower than -0.1 and RMSE was high (> 0.25) are characterized by low sampling effort ($s < 100$) and high mean dispersal distance ($d = 10$). Errors in $\hat{a}$ and $\hat{b}$ seem to compensate each other when calculating $\hat{d}$ and $\hat{p}_{99}$, resulting in RMSE for $\hat{d}$ always lower than 0.25 except when s was extremely low and dispersal was extensive.

Effect of simulated dispersal on kernel fitting

In general, the comparison of estimates obtained with different PDSs showed that both bias and RMSE increased with increasing d and decreasing b. Since high d and low b indicate long dispersal PDSs, relative bias and RMSE for dispersal kernel parameters were generally higher when simulated dispersal was long. Both precision and accuracy for $\hat{a}$ were more sensitive to b. On the other hand, $\hat{b}$ was better estimated when d was low. This caused $\hat{d}$ and $\hat{p}_{99}$ to be markedly underestimated (i.e. relative median bias significantly lower than -0.1) only for PDSs with large mean dispersal distances ($d = 10$) (Fig. 3.4). RMSE and bias behaved similarly, even though differences among PDSs were generally more evident for the former (Fig. 3.4).

Effect of sampling effort allocation on kernel fitting

For testing the relative importance of t and s/t on bias and RMSE, sampling scenarios with constant total number of seeds (s) were compared. For a given value of s, no clear effect of increasing t to the detriment of s/t, and *vice versa*, was detected. The error trend with increasing t keeping constant s is flat for all parameters ($\hat{a}$, $\hat{b}$, $\hat{d}$, $\hat{p}_{99}$) meaning that the same level of accuracy and precision is achieved when using different sampling allocations (Fig. 3.5).

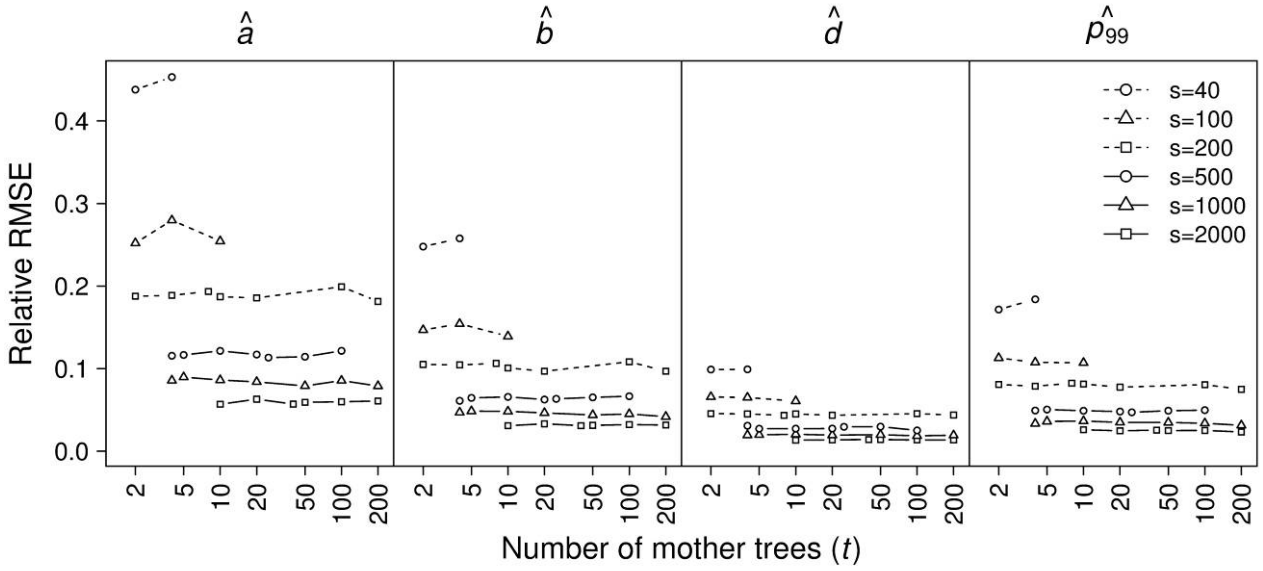

Figure 3.5: Relative RMSE of dispersal kernel parameters ($\hat{a}$, $\hat{b}$, $\hat{d}$, and $\hat{p}_{99}$) with increasing number of sampled mother trees (*t*, x-axis in log scale) in SSs characterized by constant total number of seeds (*s*). Each line links different combinations of *t* and *s/t* within SSs characterized by the same sampling effort (exponential power with $b = 1$ and $d = 3$).

Effect of sampling effort on pollen immigration rate

The expected pollen immigration rates from the background population to the sampling area were much more dependent on mean dispersal distance (*d*) than on the shape parameter (*b*) (Table 3.1). Pollen immigration increased proportionally to *d*, spanning from ~0.08 when $d = 1$ to ~0.50 when $d = 10$.

Both precision and accuracy in pollen immigration estimates generally increased together with *d* (Fig. 3.6). Contrary to what was found for dispersal kernel parameters, accuracy and precision in pollen immigration estimates were more sensitive to the number of traps (*t*) rather than total number of sampled seeds (*s*). Pollen immigration was always underestimated, and biases were significantly lower than -0.1 in 58% of cases. Bias and RMSE easily exceed -0.5 and 0.5, respectively, when pollen dispersal was spatially limited ($d = 1$) and the number of mother trees sampled was low ($t \leq 10$). For PDSs characterized by low *d*, accuracy and precision reached a plateau at $t \sim 20$, regardless of *s*. On the other hand, when pollen dispersal was extensive ($d = 10$), pollen immigration estimates were almost unaffected by sampling efforts, even when *s* was as low as 40 and *t* as low as 2 (Fig. 3.6).

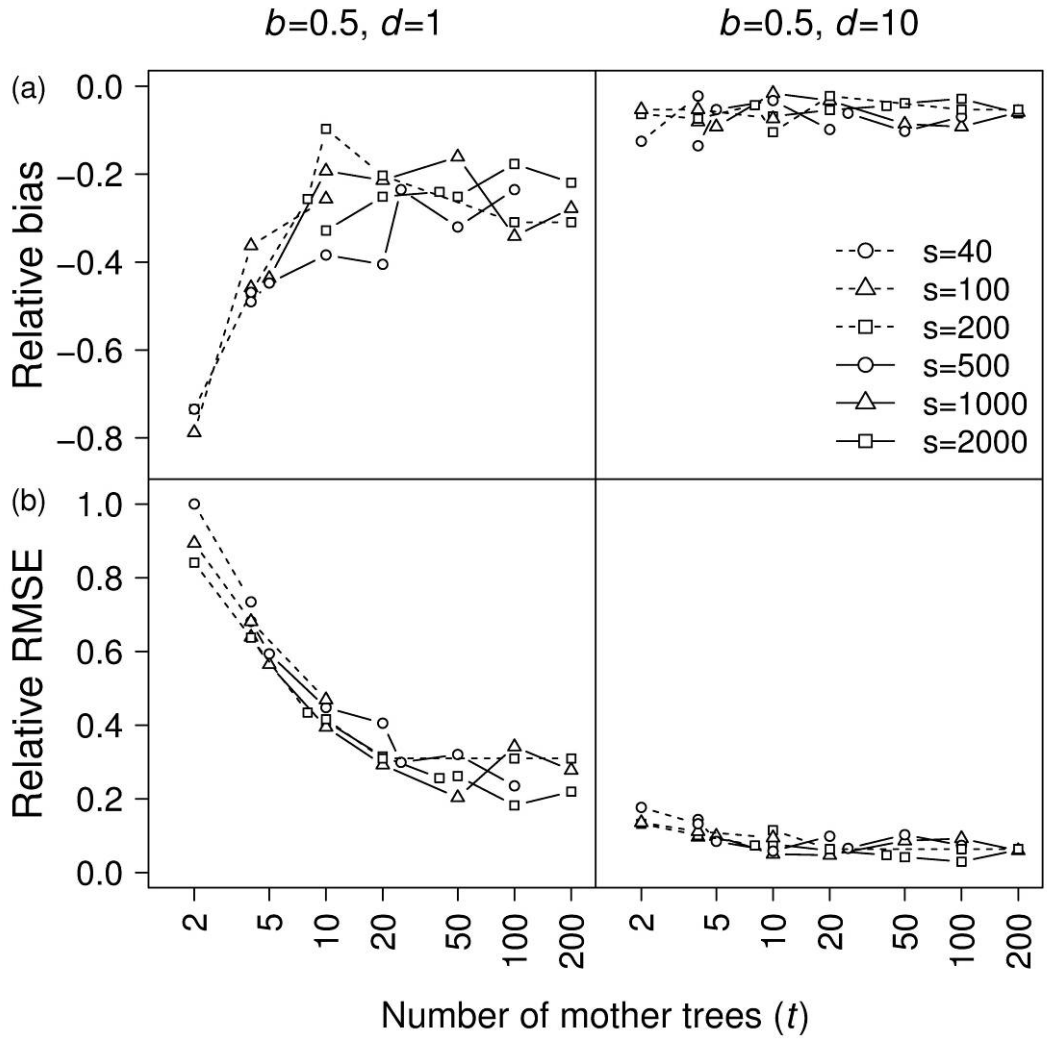

Figure 3.6: Relative bias (a) and RMSE (b) of pollen immigration estimates with increasing number of sampled mother trees (*t*, x-axis in log scale) in SSs characterized by constant total number of seeds (*s*). Each line links SSs characterized by equal *s* and different combinations of *t* and *s/t*. The first column reports results for an exponential power dispersal kernel with $b = 0.5$ and $d = 1$, the second the ones for an exponential power dispersal kernel with $b = 0.5$ and $d = 10$.

Notable differences in Weibull and 2Dt PDSs

A non-linear negative relationship between errors and *s* was also found for Weibull and 2Dt PDSs. Compared to monotonically decreasing dispersal curves (i.e. exponential power), PDSs characterized by curves with a peak (i.e. Weibull for some parameter combinations) generally required a higher sampling effort for a correct reconstruction of curve characteristics near the origin (shape, peak presence and position). As an example, when dispersal was simulated according to a peaked Weibull PDS ($d = 3$ and $b = 2.5$), in 21% of simulations the reconstructed kernel was monotonically decreasing. Such drastic change was due to an

underestimation of $\hat{b}$ larger than -0.2 (Fig. 3.7). In general, parameter biases were lower for Weibull PDSs than for exponential power ones. On the other hand, moderate RMSE for $\hat{a}$ and $\hat{b}$ produced large RMSEs for $\hat{d}$ and $\hat{p}_{99}$ when dispersal was highest (b = 0.75, d = 10) (Fig. 3.8).

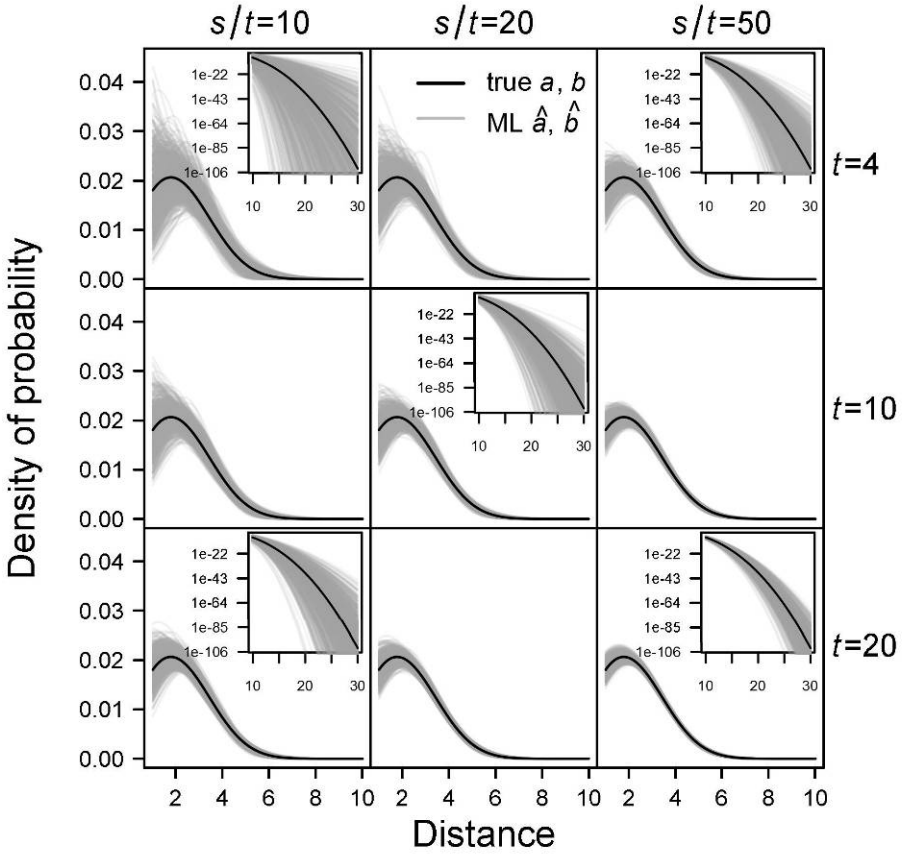

Figure 3.7: Comparison between true and inferred dispersal kernels obtained with 9 different sampling efforts characterized by increasing numbers of seeds sampled per mother tree (*s/t*, ranging from 10 to 50 along the columns) and increasing number of mother trees (*t*, ranging from 4 to 20 along the rows). In each panel, dispersal kernels inferred from 1000 simulations (grey lines) are reported and compared with the true one (black line, Weibull with b = 2.5 and d = 3). Inner windows focus on the curve tails (y-axes in log scale).

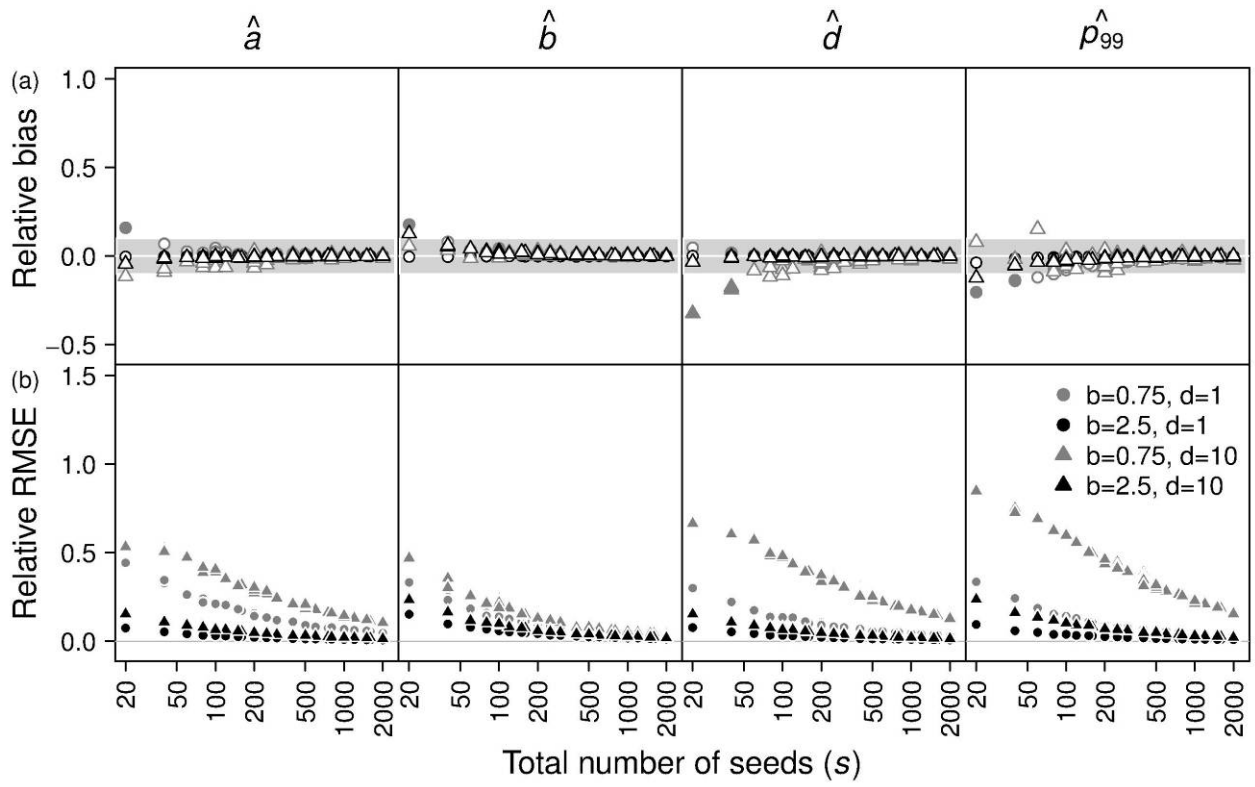

Figure 3.8: Relative bias (a) and RMSE (b) of dispersal kernel parameters ($\hat{a}$, $\hat{b}$, $\hat{d}$, and $\hat{p}_{99}$) with increasing total number of seeds (*s,* x-axis in log scale). The results for 4 different Weibull PDSs are reported with different colors and symbols as indicated in the legend. Filled-triangles and circles indicate biases significantly higher than 0.1 or lower than -0.1, whereas empty symbols correspond to values to not significantly higher than 0.1 or lower than -0.1. Bias values outside the [-0.5;1] range and RMSE values outside the [0;1.5] range are not shown.

Contrary to what was found for exponential power PDSs, errors in $\hat{a}$ and $\hat{b}$ did not compensate in 2Dt PDSs. In fact, even small biases in $\hat{a}$ and $\hat{b}$ (< 0.1) led to high biases in $\hat{d}$ and $\hat{p}_{99}$ (Fig. 3.9). As for the other families of curves, $\hat{d}$ was always underestimated. When b = 1.6, such underestimation can be severe even with our largest sampling effort (s = 2000) (Fig. 3.9).

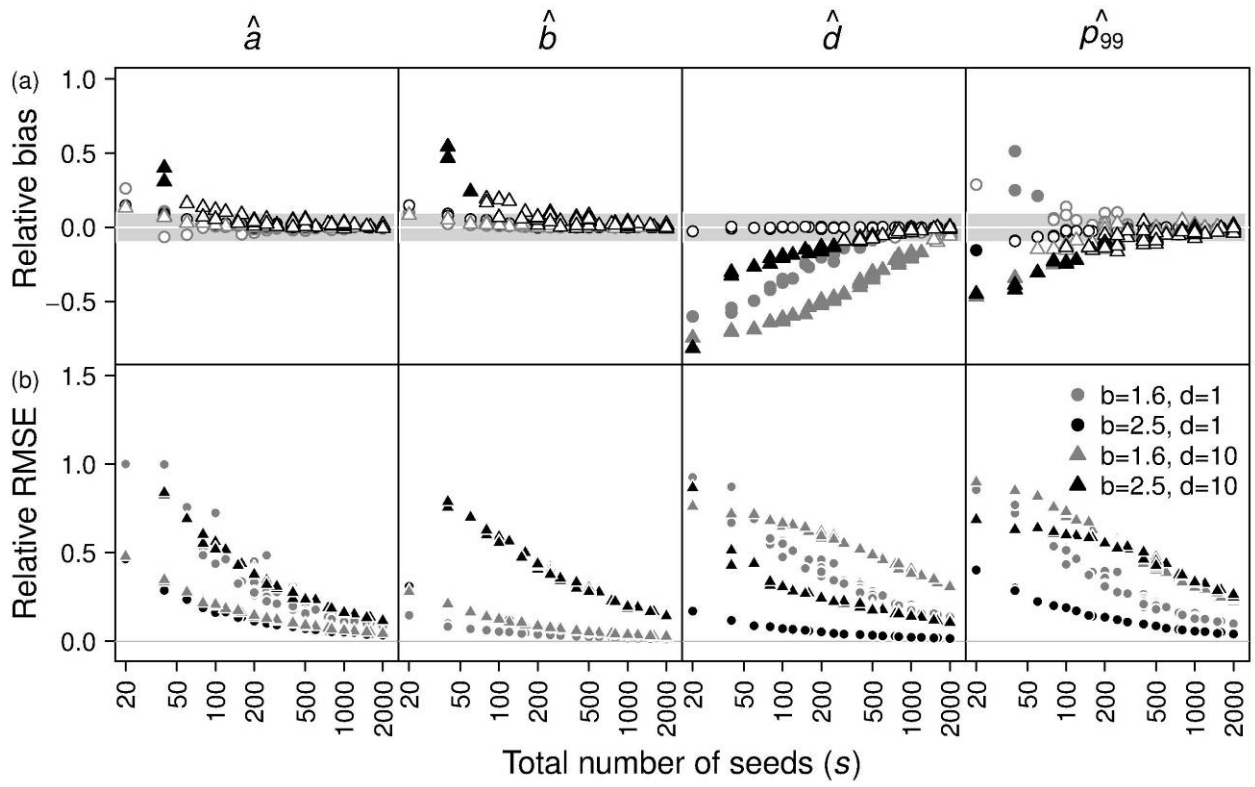

Figure 3.9: Relative Bias (a) and RMSE (b) of dispersal kernel parameters ($\hat{a}$, $\hat{b}$, $\hat{d}$, and $\hat{p}_{99}$) with increasing total number of seeds (*s,* x-axis in log scale). The results for 4 different 2Dt PDSs are reported with different colors and symbols as indicated in the legend. Filled-triangles and circles indicate biases significantly higher than 0.1 or lower than -0.1, whereas empty symbols correspond to values to not significantly higher than 0.1 or lower than -0.1. Bias values outside the [-0.85;1] range and RMSE values outside the [0;1.5] range are not shown.

Unique among the investigated PDSs, the 2Dt characterized by short mean dispersal distance but an extremely fat tail (b = 1.6, d = 1) determined a substantial overestimation of $\hat{p}_{99}$ when the sampling effort was low (s < 100). On the contrary, $\hat{p}_{99}$ was markedly underestimated when mean dispersal distance was long (d = 10). As for exponential power and Weibull, RMSE in 2Dt was generally high in long distance scenarios, though a non negligible RMSE was also found when d = 1. This caused $\hat{d}$ and $\hat{p}_{99}$ to be imprecisely estimated even when dispersal was relatively limited.

In all PDSs belonging to Weibull and 2Dt curve families, different sampling allocations provided the same level of accuracy and precision confirming what was found for exponential power PDSs (Fig. 3.10, Fig. 3.11, Fig. 3.12, Fig. 3.13).

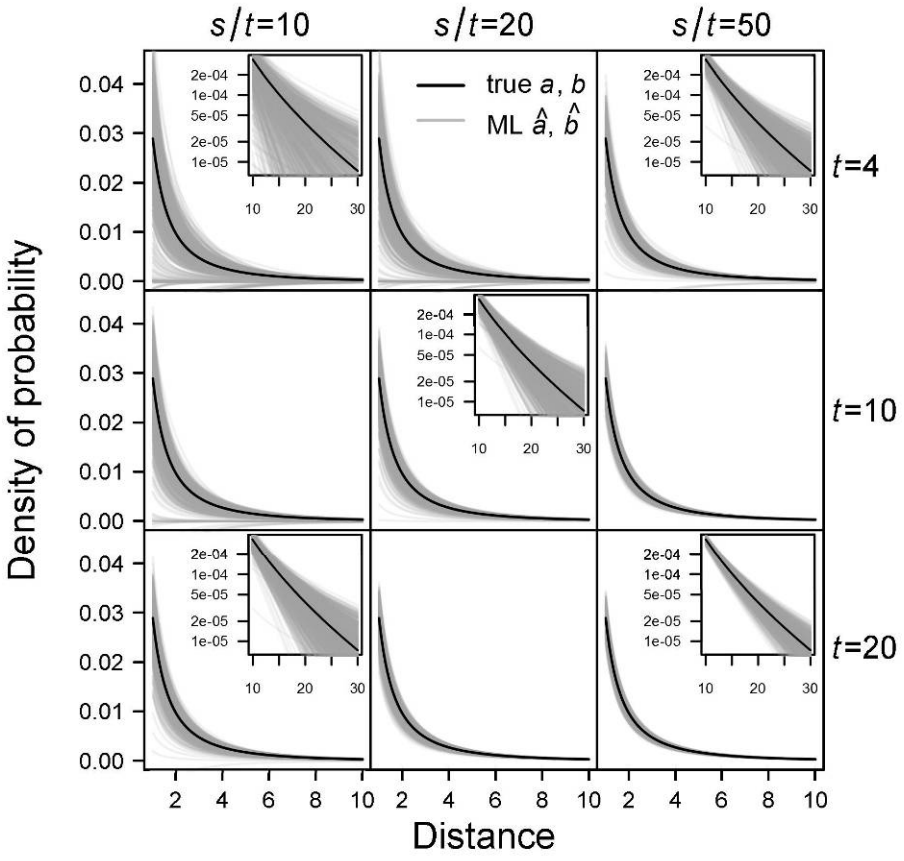

Figure 3.10: Comparison between true and inferred dispersal kernels obtained with 9 different sampling efforts characterized by increasing numbers of seeds sampled per mother tree (*s/t*, ranging from 10 to 50 along the columns) and increasing number of mother trees (*t*, ranging from 4 to 20 along the rows). In each panel, dispersal kernels inferred from 1000 simulations (grey lines) are reported and compared with the true one (black line, Weibull with $b = 0.75$ and $d = 5$). Inner windows focus on the curve tails (y-axes in log scale).

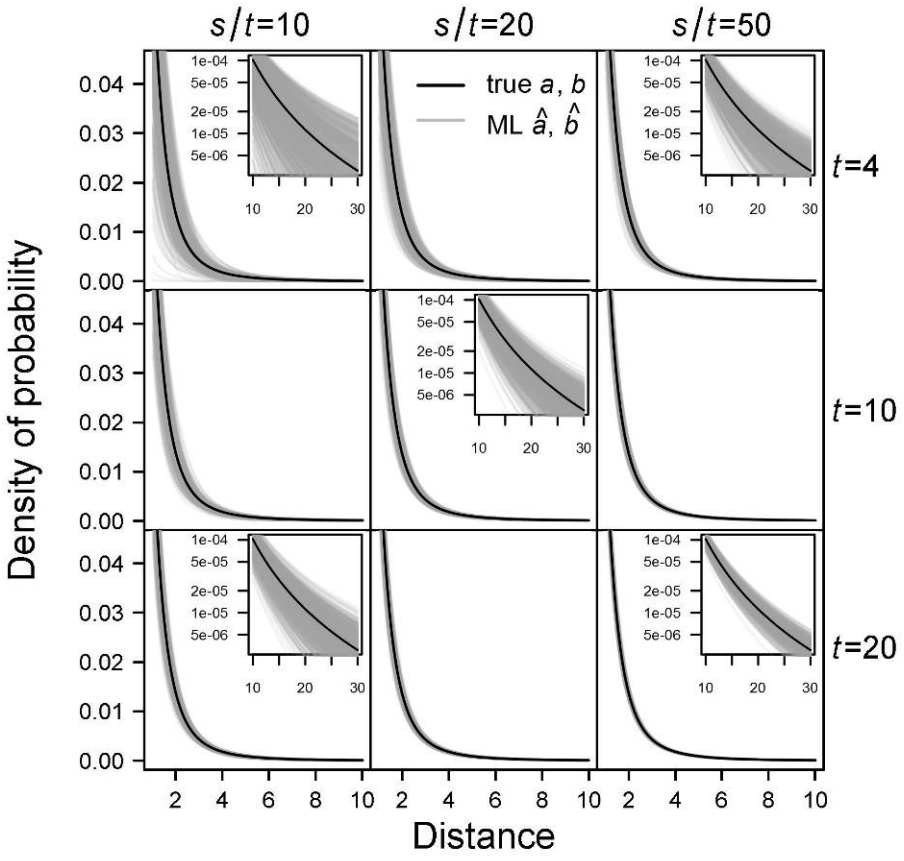

Figure 3.11: Comparison between true and inferred dispersal kernels obtained with 9 different sampling efforts characterized by increasing numbers of seeds sampled per mother tree (*s/t*, ranging from 10 to 50 along the columns) and increasing number of mother trees (*t*, ranging from 4 to 20 along the rows). In each panel, dispersal kernels inferred from 1000 simulations (grey lines) are reported and compared with the true one (black line, 2Dt with $b = 1.6$ and $d = 5$). Inner windows focus on the curve tails (y-axes in log scale).

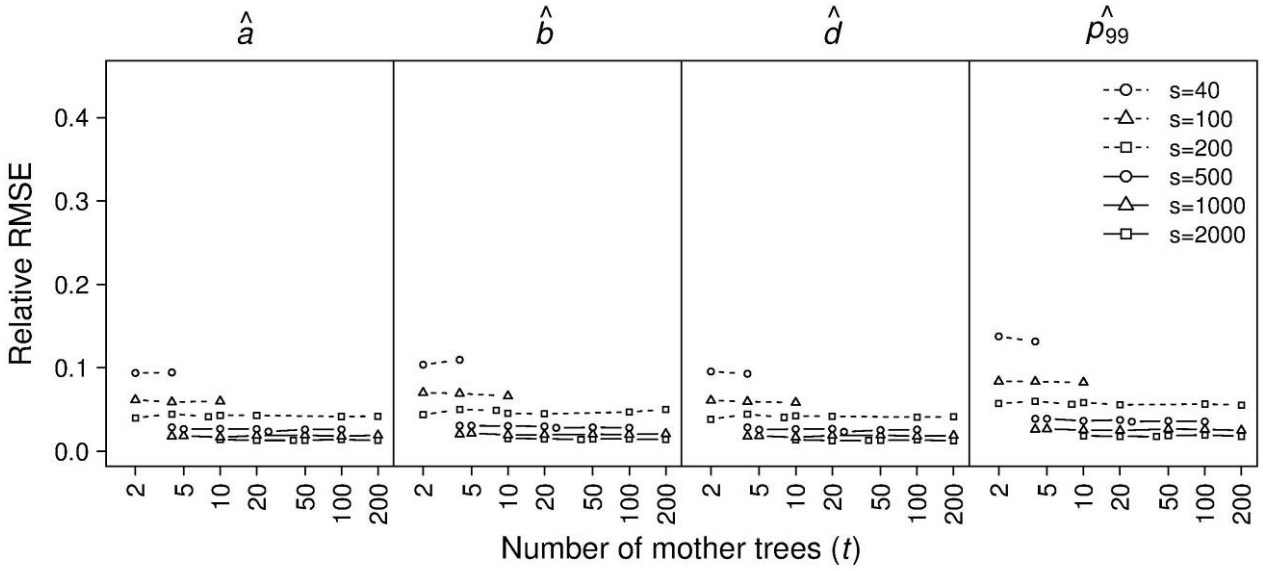

Figure 3.12: Relative RMSE of dispersal kernel parameters ($\hat{a}$, $\hat{b}$, $\hat{d}$, and $\hat{p}_{99}$) with increasing number of sampled mother trees (*t,* x-axis in log scale) in SSs characterized by constant total number of seeds (*s*). Each line links different combinations of *t* and *s/t* within SSs characterized by the same sampling effort (Weibull with b = 1.5 and d = 3).

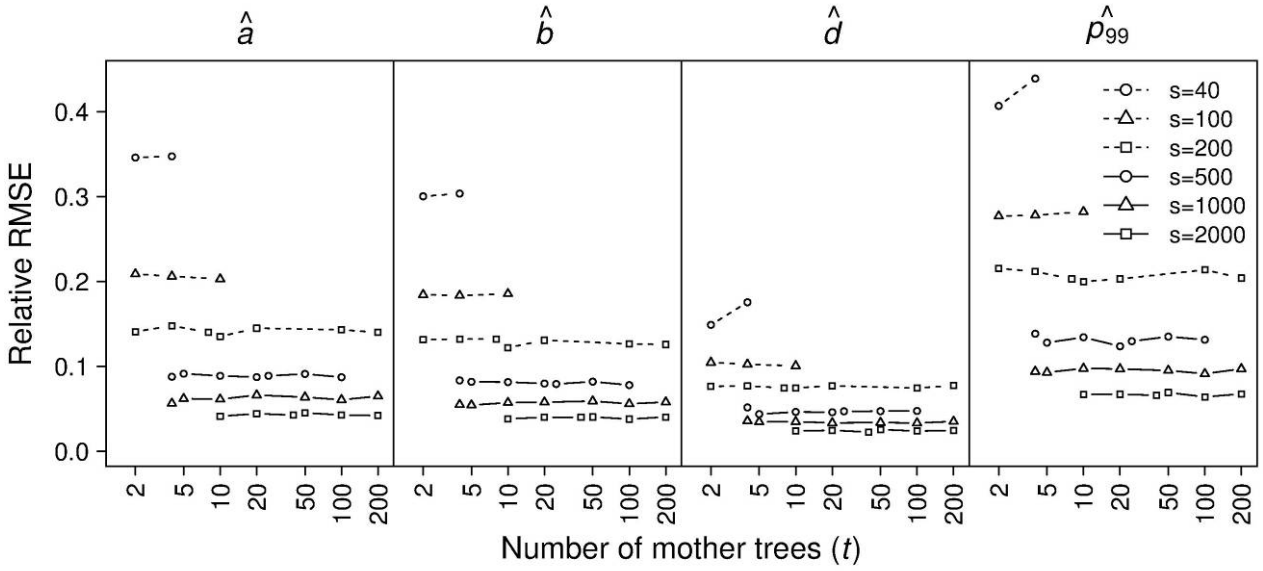

Figure 3.13: Relative RMSE of dispersal kernel parameters ($\hat{a}$, $\hat{b}$, $\hat{d}$, and $\hat{p}_{99}$) with increasing number of sampled mother trees (*t,* x-axis in log scale) in SSs characterized by constant total number of seeds (*s*). Each line links different combinations of *t* and *s/t* within SSs characterized by the same sampling effort (2Dt with $b = 2.5$ and $d = 3$).

In accordance with results from exponential power, bias and RMSE for pollen immigration estimates in Weibull and 2Dt PDSs improved with increasing *t*, and were virtually unbiased when *d* was high (Fig. 3.14, Fig. 3.15). Pollen immigration in 2Dt was generally underestimated, with 61% of cases significantly lower than -0.1 (Fig. 3.15). In Weibull PDSs, pollen immigration rate was in general accurately estimated, underestimation only occurred when $d = 1$ (Fig. 3.14). This is the only curve family where in few PDS-SS combinations (2% of cases) pollen immigration rate was slightly overestimated (up to 0.2).

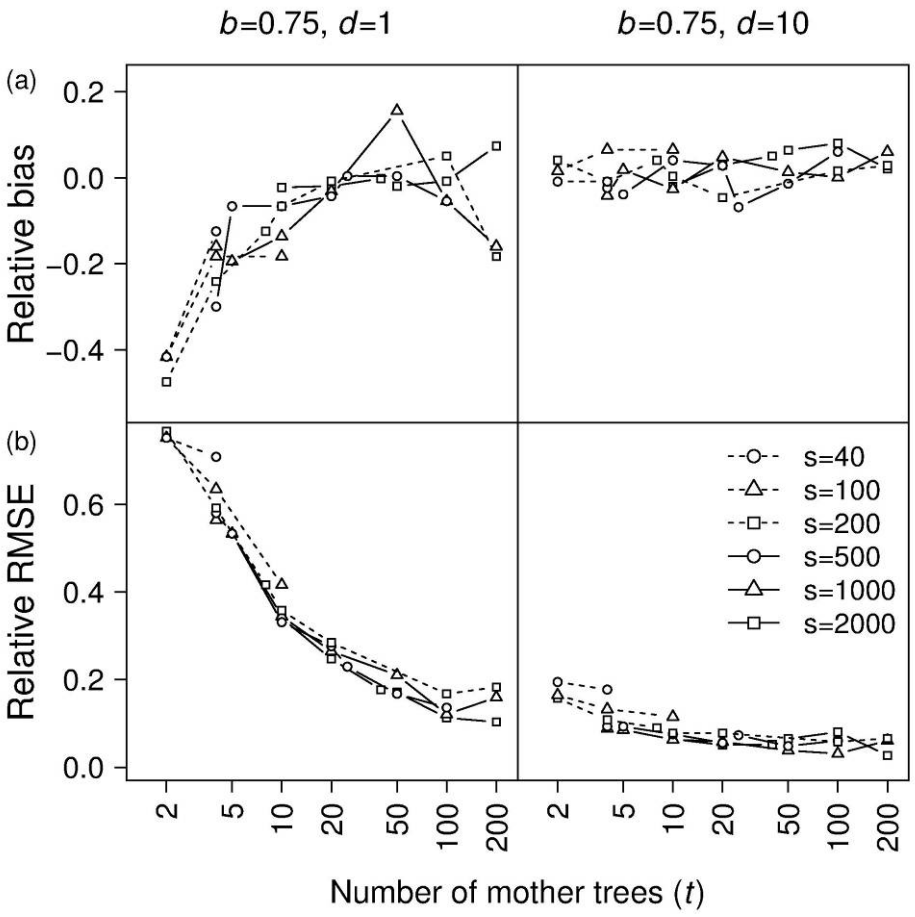

Figure 3.14: Relative bias (a) and RMSE (b) of pollen immigration estimates with increasing number of sampled mother trees (*t,* x-axis in log scale) in SSs characterized by constant total number of seeds (*s*). Each line links SSs characterized by equal *s* and different combinations of *t* and *s/t*. The first columns reports results for a Weibull dispersal kernel with $b = 0.75$ and $d = 1$, the second the ones for a Weibull dispersal kernel with $b = 0.75$ and $d = 10$.

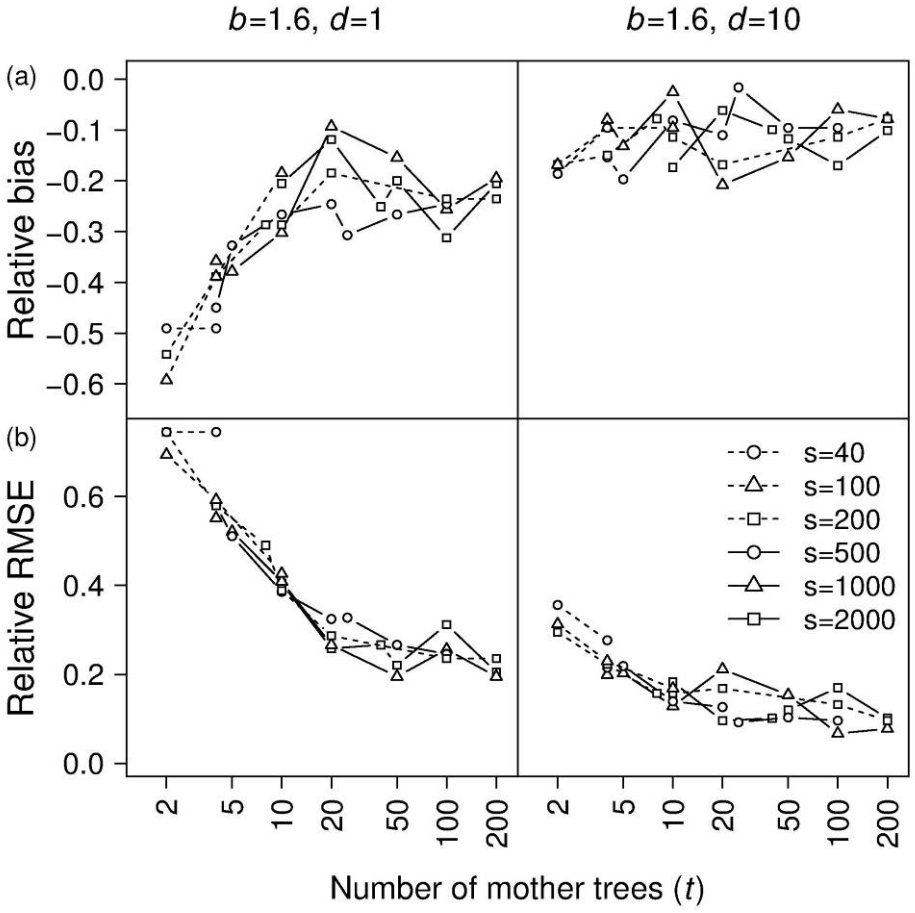

Figure 3.15: Relative bias (a) and RMSE (b) of pollen immigration estimates with increasing number of sampled mother trees (*t,* x-axis in log scale) in SSs characterized by constant total number of seeds (*s*). Each line links SSs characterized by equal *s* and different combinations of *t* and *s/t*. The first columns reports results for a 2Dt dispersal kernel with $b = 1.6$ and $d = 1$, the second the ones for a 2Dt dispersal kernel with $b = 1.6$ and $d = 10$.

Assessment of methodological reliability

The fitting procedure converged in 99.85% of the 1,920,000 simulations. Lack of convergence reached a maximum of 13% when $s = 20$ and dispersal was simulated according to an exponential power with low b (0.5) and high d (10). The percentage of convergence was positively related to the percentage of realized father-trap connections (Spearman correlation: $\rho = 0.182$, $P < 0.001$). The percentage of realized father-trap connections is computed as the ratio between the number of father-trap connections with at least one realized pollination and the total number of possible father-trap connections. Basically, it is the ratio between non-zero cells and total cells in the x_{ij} matrix of observed data used for the computation of likelihood in the CSM (Robledo-Arnuncio and Garcia, 2007). The percentage of realized father-trap connections was mainly dependent on the number of propagules sampled in each trap. In fact, *s/t* alone explained 73% of the variability in the percentage of realized father-trap

connections. When this percentage was low, successful fitting often produced outliers in $\hat{a}$ and $\hat{b}$ distributions that in turn determine non-robust estimates of d, p_{99} and pollen immigration rate. Because in most cases a clear cut rationale to discard outliers could not be defined, I decided to rely on the median over 1000 simulations, instead of the trimmed mean, as a robust statistic to represent central values.

Estimates of bias and RMSE of $\hat{b}$ and $\hat{d}$ in the 8 PDS-SS combinations shared with Robledo-Arnuncio and Garcia (2007) were highly correlated with their results (intercept = -0.008, t = -4.18, $P < 0.001$; slope = 0.92, t = 33.08, $P < 0.001$, df = 30, $R^2 = 0.97$). The comparison among pollen immigration estimates showed very similar results as well (intercept = 0.04, t = 4.54, $P < 0.01$; slope = 1.07, t = 41.34, $P < 0.001$, df = 10, $R^2 = 0.97$).

3.4 Discussion

In this work, the effect of sampling effort on the estimation of pollen dispersal parameters in paternity studies was investigated through a simulation approach. The aim was to provide the first evaluation of sampling efforts reported in published paternity studies. I found that: *i*) scale and shape parameters of the dispersal kernel were generally overestimated, causing mean dispersal distances and 99th percentiles in the tail of the curve to be underestimated; *ii*) except for rare exceptions, pollen immigration from outside the sampling plot was generally underestimated; *iii*) pollen immigration estimates were highly sensitive to the number of sampled mother trees, whereas precision and accuracy of dispersal kernel parameters only depended on the total number of sampled seeds; and *iv*) errors in pollen immigration estimates were high when simulated dispersal was restricted, whereas the opposite was found for dispersal kernel parameter estimates. Our findings show that, when sampling effort is constrained, as in most paternity experiments, it should be fine-tuned to the specific aims of the study.

Estimation of dispersal kernel parameters

Assessing uncertainty in dispersal kernel estimation has emerged as an urgent need in recent literature (Robledo-Arnuncio and Garcia, 2007; Jones and Muller-Landau, 2008; Niggemann *et al.*, 2012; Nathan *et al.*, 2012). Errors in such estimates are mainly dependent on sampling design (sampling effort and trap spatial arrangement) and, when genetic tools are adopted, on marker resolution and frequency of genotyping errors (e.g. Moran and Clark, 2011). Focusing

on the sampling effort, I quantified the accuracy and precision in estimating shape and scale parameters describing curves commonly used in paternity studies.

Diminishing accuracy and precision of parameter estimates as the total number of sampled seeds (*s*) decreases is an obvious result and confirms what was found in Robledo-Arnuncio and Garcia (2007). However, I showed that $s \sim 200$ is a turning point below which both bias and RMSE can reach warning levels, especially for long distance pollen dispersal scenarios (PDSs). When sampling effort is low, positive biases in estimated scale ($\hat{a}$) and shape ($\hat{b}$) parameters determine negative biases in mean dispersal distance ($\hat{d}$) and 99th percentile ($\hat{p}_{99}$); $\hat{d}$ can be ~70% and $\hat{p}_{99}$ ~50% of the true ones when simulated pollen dispersal was extensive with respect to the plot size ($l/d = 2$). Contrary to what previously thought, *d* can be severely underestimated unless an adequately high sampling effort is adopted, even when using the Competing Sources Model (CSM). Therefore, published paternity studies in which *d* is estimated solely from the observed distribution of pollination distances are likely to be doubly biased: once for the sampling artifact worked out by CSM (the effect of competing pollen sources) and once for inadequate sampling. Together with results for the p_{99}, this indicates that, when dispersal is high, kernel fitting on within-population data depicts dispersal as being shorter than it actually is.

Underestimation of long dispersal distances is due to several sampling limitations. First, the curve tail is truncated by plot boundaries, although CSM lowers this sampling artifact (Robledo-Arnuncio and Garcia, 2007). Second, no spatial data regarding immigrant pollen are available, even though information from censored data could be included (Jones *et al.,* 2005). Finally, the number of traps near the plot boundaries is rarely adequate when a random spatial distribution of traps is adopted, thus preventing within-plot long distances to be properly sampled. A precise characterization of long distance dispersal (LDD) provides information on relevant biological and ecological processes, such as species invasions, hybridization and contamination events, conservation of genetic diversity, population responses to habitat fragmentation and climate change (Bullock and Clarke, 2000). The quantitative description of LDD is considered the most challenging task in dispersal studies because it requires tracking of rare dispersal events occurring at high distance from the source (Nathan, 2006; Hardy, 2009). The characterization of the tail of the kernel can be improved by including minimum dispersal distance for censored data (i.e. the distance from propagule source to the nearest plot boundary). Few studies have included censored data in kernel estimation (e.g. Jones *et al.*, 2005; Goto *et al.,* 2006; Piotti *et al.*, 2009; Chybicki and Burczyk, 2010a). In a simulation study, Hirsch *et al.* (2012) showed that unbiased estimates of seed dispersal kernel can be

achieved by including censored data if the study area includes at least 40-50% of total dispersed seeds. Robledo-Arnuncio and Garcia (2007) showed that accounting for the immigration rate in the CSM did not significantly improve kernel estimation for $s \geq 200$. According to my findings, this might not hold true when dispersal is high.

The spatial arrangement of traps is relevant for estimating dispersal kernel parameters. A number of studies have focused on optimizing the arrangement of traps in space and the number of traps at different distances from the source (e.g. Klein and Laredo, 1999; Bullock and Clarke, 2000; Stoyan and Wagner, 2001; Skarpaas *et al.*, 2005; Skarpaas and Shea, 2007). These studies generally highlighted that a number of traps should be placed far from the source in order to characterize the tail of the curve, and that trap spatial arrangement should allow an even sampling of as many dispersal distances as possible. Nonetheless, these results are difficult to apply to paternity studies, because they are based on a single source of propagules (but see Stoyan and Wagner (2001) for an exception), on extremely high numbers of traps and on the possibility to decide the location of traps. Therefore, future simulation studies should focus on the combined effect of the number and criteria to choose mother trees, considering multiple sources with a non-random spatial distribution, as in the case of natural populations.

This study also showed that, under the simplistic assumptions upon which it is based, adequate estimates of both $\hat{d}$ and $\hat{p}_{99}$ can be generally achieved with manageable sampling effort ($100 < s < 200$). It should also be stressed that in the explored scenarios dispersal is mainly local, since p_{99} is at most 4 times the side of the plot (l). Recent findings have shown that pollen is still viable after traveling much longer distances, resulting in LDD being more frequent than previously thought (Williams, 2010; Kremer *et al.,* 2012). In addition, LDD is governed by more complex laws than those used to model local dispersal, and depends more on stochastic processes linked to meteorological conditions than on distance itself (Nathan, 2006; Hardy, 2009). Therefore, LDD by pollen requires new approaches for its tracing (Kremer *et al.,* 2012). The disparity between local distance-controlled *vs*. distance-independent meteorology-controlled pollination patterns should also be carefully taken into account in future practical and theoretical work on the reconstruction of the dispersal kernel.

Estimation of pollen immigration rate

Negative biases were also generally found in estimates of pollen immigration from outside. Underestimation became substantial when dispersal was short (low d) and the number of

mother trees was low ($t \leq 10$). In these cases, estimated pollen immigration rate can be as low as ~50% of the true one. It should be noted that a 50% underestimation can have different relevance depending on the actual pollen immigration. However, even a small absolute difference between the actual and estimated pollen immigration rates might affect biological processes (e.g. population differentiation) besides being relevant when addressing debatable issues such as quantifying GM pollen contamination.

The effect of *d* on pollen immigration estimates has a rather simple explanation: when *d* is low pollen immigration is detectable only near plot boundaries. Consequently, studies on species characterized by low pollen dispersal capability may suffer from sampling flaws if plot boundaries are inadequately sampled. In my simulations, this happens when *t* is low. This argument holds true also when the *l/d* ratio is high, i.e. when the sampling area is disproportionately large with respect to mean dispersal distance. Thus, enlarging the sampling area can produce a large underestimation of pollen immigration rates if *t* is not properly adjusted. On the other hand, when *d* is high (or *l/d* is low), estimation is less affected by sampling effort and allocation. Species' dispersal characteristics and study-specific sampling designs should indeed be carefully considered when comparing results from different studies.

My results indicate that different sampling strategies should be used depending on whether the aim is to estimate pollen immigration rate or kernel parameters. Sampling allocation (i.e. the combination of mother trees and seeds per mother tree) has a marginal effect on kernel parameters' estimation. On the contrary, *t* is much more important than *s/t* for both precision and accuracy of pollen immigration estimates when simulated dispersal is low. This is again related to trap coverage of plot boundaries and to the heterogeneity of immigration rates among mother trees. Published pollen immigration estimates based on low *t* (≤10) are common (64% of cases in Leonarduzzi *et al.*, 2012). In these cases there may be a significant underestimation of pollen immigration if the study species is characterized by short dispersal and the spatial arrangement of traps is inappropriate.

When studying pollen immigration, it becomes of particular interest that p_{99} and pollen immigration rate should not be considered as interchangeable measures of LDD. In fact, for constant *d*, pollen immigration rate may counter-intuitively decrease as p_{99} increases (see Table 3.1, Fig. 3.16). This is because the pollen immigration rate quantifies the proportion of propagules coming from outside the study plot, regardless of the distance they traveled. Even when pollen immigration rate is high, all immigrant pollen can originate from a tree just outside the study plot. Obviously, the association between pollen immigration rate and long distance dispersal holds when highly isolated populations are exhaustively sampled in a

paternity experiment.

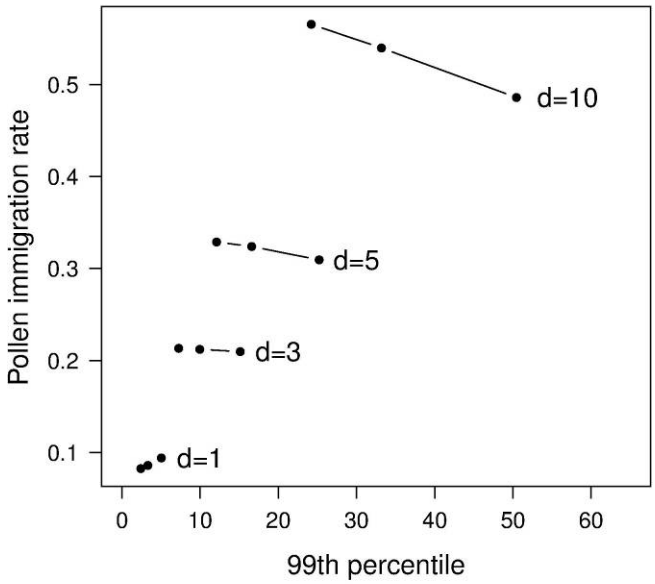

Figure 3.16: Relationship between the tail of the dispersal kernel (99th percentile) and the rate of pollen immigration in the exponential power PDSs.

Considerations of sampling effort in published paternity studies

Robledo-Arnuncio and Garcia (2007) found that a relatively small *s* (200) seems sufficient to estimate a kernel's mean accurately (bias ~ 0.01), but reducing the RMSE of the kernel's shape estimates to low levels (< 0.10) will require higher *s* (500). I found that further decreasing *s*, in particular when $t \leq 10$, may drastically reduce robustness of estimates from paternity studies aimed at exploring simultaneously different components of pollen dispersal. The positive relationship between sampling effort and precision was also experimentally determined by doubling *s* and quadruplicating *t* in one of the rare cases in which two paternity experiments were conducted in the same plot (Oddou-Muratorio *et al.*, 2005).

The median values of *s* and *t* in paternity studies reviewed by Leonarduzzi *et al.* (2012) were 240 and 8, respectively. When a sampling scenario (SS) with comparable sampling effort and allocation (*s* = 250, *t* = 10, *s/t* = 25) was simulated, I found a severe underestimation of *d* (~ 50%) if pollen dispersal was extensive (the PDS with highest p_{99}), and a substantial underestimation of pollen immigration rate (~ 30%) if pollen dispersal was limited (e.g. exponential power PDSs with *d* = 1). Higher errors are expected for lower sampling efforts. Such errors can lead to misinterpretation of results and low comparability among studies. This should be carefully taken into account in future reviews on pollen and seed dispersal results

obtained by parentage analysis. Finally, it should be noted that sampling efforts recorded in Leonarduzzi *et al.* (2012) could be downwardly biased because in multi-plot studies (where many small stands are sampled with low sampling effort) the sampling effort for each single stand was considered as a data point.

Given the increasing number of paternity studies comparing several plots (e.g. Oddou-Muratorio *et al.,* 2010; Craft and Ashley, 2010; Piotti *et al.*, 2012), my results showed that the effectiveness of such comparisons can be greatly improved through sound sampling designs. For instance, when a large number of plots is sampled to investigate the effect of habitat fragmentation on pollen dispersal patterns (e.g. Lander *et al.*, 2010; Wang *et al.,* 2010) particular attention should be paid to sampling strategy in large fragments. In fact, when *l/d* becomes high, concentrating a limited number of traps far from plot boundaries can introduce large errors in pollen immigration rate. This holds true also for single plot experiments, in particular when sampling allocation is unbalanced towards *s/t* (e.g. Lian *et al.*, 2001; Hanaoka *et al.,* 2007). On the other hand, since evidence is accumulating about pollen dispersal being extensive in forest trees at the local scale, values of *l/d* are likely to be low, and the most severe bias might occur for *d* and p_{99} (see below).

Sampling guidelines

My findings stress the relevance of carefully planning the sampling design and offer a piece of advice for future paternity studies. Considerations of optimal sampling design will inevitably have to cope with natural constraints posed by the study population and with limited resources.

If a robust estimation of dispersal kernel parameters is the goal of a paternity study, the total number of sampled seeds (*s*) is the key factor to control. Conversely, when the focus is on pollen immigration from outside, increasing the number of mother trees (*t*) to the detriment of the number of seeds per mother tree (*s/t*) is the advisable strategy. Estimation of pollen dispersal parameters can also benefit from *a priori* knowledge of the study species' dispersal capability:

1) For forest species whose dispersal is characterized by high *d*, or in cases where *l/d* is low, minimum requirements for dispersal kernel parameters depend on simulated dispersal curve families. When studying a species usually characterized by large tail fatness (as in 2Dt kernel), lowering bias to acceptable levels may require at least 1000 and 200 seeds for *d* and p_{99}, respectively, whereas even a large sampling effort (s = 2000) is not sufficient to provide precise estimates of *d* and p_{99}. On the contrary, lower sampling

efforts ($100 < s < 200$) can provide unbiased estimates for d and p_{99} when dispersal is less fat-tailed, whereas $s \geq 500$ is needed for achieving good precision in parameters' estimates. With regard to the estimation of pollen immigration rate, a limited number of seeds and mother trees ($s \sim 200$ and $t \sim 5$) can be sufficient to achieve robust estimates, regardless of the simulated dispersal curve family.

2) For forest species whose dispersal is characterized by low d, or in cases where l/d is high, a small sampling effort ($s \sim 200$) is generally sufficient to obtain adequate accuracy and precision in dispersal kernel parameters. On the other hand, a high t (at least 20-50 depending on the specific PDS and on the curve family) should be used for obtaining adequate estimates of pollen immigration rate. In these cases, the robustness of pollen immigration estimates will also benefit from carefully choosing the spatial distribution of mother trees (e.g., avoiding placing all traps in the centre of the plot). As previously mentioned, *ad hoc* mother tree arrangements are likely to markedly improve pollen immigration and kernel parameter estimates. To this end, it will be important to asses the combined effect of sampling effort and trap arrangement on pollen dispersal parameters.

When using these guidelines, it is important to note that my simulations are based on many simplifying assumptions which could impact our results. Assumptions are mainly linked to: *i*) the randomness of spatial population structure, *ii*) the complexity of dispersal processes, and *iii*) the correctness of paternity assignment. With regard to spatial population structure, I assumed that trees are randomly distributed and that tree density does not hamper pollen dispersal, whereas in natural populations trees are more likely to be clumped, the landscape is heterogeneous and high tree density can limit dispersal (Hardy, 2009). Moreover, I described a simplified dispersal process assuming that all trees disperse according to the same isotropic dispersal kernel and have equal fecundity. Variation in tree fecundity was shown to have negligible effect on kernel reconstruction when the total number of sampled seeds is ≥ 200 and trees are randomly placed in space (Robledo-Arnuncio and Garcia, 2007; Robledo-Arnuncio, 2008). However, high variation in tree fecundity combined with a scarce sampling effort can produce misrepresented dispersal distances, especially when the number of traps is low, and further investigation is needed. Also including complex spatial patterns and more realistic dispersal characteristics (e.g. by introducing flowering phenology) in future *ad hoc* simulation studies will improve the usefulness of these sampling guidelines.

Finally, paternity was reconstructed using categorical allocation and assuming no genotyping error, but genotyping errors are know to highly affect paternity assignments (Jones *et al.*, 2010). Dealing with uncertainty associated with genotyping errors has become increasingly

important in parentage analysis. In fact, recently developed methods enable simultaneous estimation of the dispersal kernel and mating patterns, explicitly taking into account possible genotyping errors (Chybicki and Burczyk, 2010b). Bayesian approaches are the most promising tools to achieve a thorough description of dispersal patterns (Klein and Oddou-Muratorio, 2011; Moran and Clark, 2011) and could benefit from a careful description of how sampling design can affect dispersal estimates. This would enhance our knowledge on the evolutionary dynamics of populations in changing environments and provide sound information for management and conservation actions.

Chapter 4:

Development of polymorphic microsatellites from transcriptome and genomic data for *Abies alba* Mill. and congeneric species

4.1 Introduction

Silver fir (*Abies alba* Mill.) is a widespread European conifer. It is a keystone species of many mountain forest ecosystems with high ecological and economic value and is also found at low density associated with other widespread European species such as beech (*Fagus sylvatica* L.) and spruce (*Picea abies* Karst.) (Wolf 2003). Despite being tolerant of a relatively broad range of environmental conditions (e.g. it is cold-hardy and shade-tolerant), silver fir is more sensitive than other conifers to changes in temperature, water availability and air pollution. In particular, showing lower water-use efficiency compared to other fir species from more xeric areas, it is expected to be severely affected by drought under a changing climate (Guehl *et al.* 1991; Aussenac 2002; Macias *et al.* 2006; Linares and Camarero 2012). In the last 200 years, its range has significantly decreased due to both environmental changes and human impact through deforestation, overexploitation, silvicultural choices in favor of faster growing conifers, improper management and air pollution (Wolf 2003). In other parts of its range, particularly at the upper tree limit, its distribution has increased over the same period due to land use changes, i.e. natural recolonization of abandoned agricultural and low productivity lands (Chauchard *et al.* 2007, 2010). Peripheral *A. alba* populations, in particular at the southern edge of the distribution, are expected to be the most affected by climate change due to their small population size, fragmented distribution and bio-geographical position (Piovani *et al.* 2010; Maiorano *et al.* 2013).

Its relevance for European forest ecosystems has made *A. alba* the object of several genetic surveys using terpenes, isozymes, mitochondrial DNA markers, chloroplast and nuclear microsatellites (SSRs) (e.g. Konnert and Bergmann 1995; Vendramin *et al.* 1999; Liepelt *et al.* 2002; Sagnard *et al.* 2002; Liepelt *et al.* 2009; Piovani *et al.* 2010; Gömöry *et al.* 2012), providing important information on the species' postglacial recolonization history and on the spatial distribution of its genetic diversity at different scales. There is now an urgent need for conservation-oriented population genetic studies with increased resolution to assess current

dynamics and future evolutionary trajectories of *A. alba* populations strongly exposed to environmental change and to evaluate current conservation strategies in Europe (Lefèvre *et al.* 2013). The set of currently available SSRs for *A. alba* is not suitable for this task due to their limited number and the presence of null alleles (Cremer *et al.* 2012; Gömöry *et al.* 2012). For this reason, I developed and characterized in natural populations new SSRs using both transcriptomic and genomic resources.

Transcriptome sequencing using next-generation sequencing (NGS) is an effective tool for generating genomic resources and identifying polymorphic molecular markers for non-model organisms, particularly for species characterized by large and repetitive genomes such as conifers (Parchman *et al.* 2010; Roschanski *et al.* 2013). Developing SSRs from transcriptome sequences (EST-SSRs hereafter) is labor- and cost-effective compared to the conventional procedure for the identification of genomic SSRs (i.e. screening of cloned libraries by Sanger sequencing, Schoebel *et al.* 2013). Moreover, EST-SSRs are easily transferable to other species due to the higher level of sequence conservation of transcribed DNA across species (Varshney *et al.* 2005; Zalapa *et al.* 2012; Fan *et al.* 2013). They are expected to be less polymorphic than genomic SSRs (gSSRs hereafter), but also less prone to null alleles, making them ideal for genetic studies in which genotyping errors should be strictly avoided, e.g. fine-scale population genetic studies and parentage studies (Kim *et al.* 2008; Oddou-Muratorio *et al.* 2009).

In this work, I introduce two sets of new multiplexed EST-SSRs for high-resolution and cost-effective genetic analyses in *A. alba* and several congeneric taxa. To do this, I took advantage of available transcriptome data (Roschanski *et al.* 2013). In addition, two new gSSRs were developed and multiplexed with previously available gSSRs displaying high amplification quality (Hansen *et al.* 2005; Cremer *et al.* 2006). I describe the procedure to identify and optimise the new EST-SSRs and to design multiplex sets, paying particular attention to quality controls (e.g. null-allele detection). I compared the performance of EST-SSRs and gSSRs and tested the transferability of EST-SSRs to 17 congeneric taxa from the Mediterranean, Asia and North-America.

4.2 Materials and methods

Plant material

Plant material was collected from four populations: Northern (Abetone, 44°8'28"N, 10°40'3"E, coded as ABE) and Southern (Sila, 39°7'57"N, 16°38'19"E, SIL) Apennines

(Italy), Bulgaria (Bansko, Pirin mountains, 41°50'35" N, 23°23'7" E, BL) and Romania (Arges, Fagaras mountains, 45°26'28"N, 24°41'40"E, RH) (Fig. 4.1). Forty-eight adult trees (pair-wise minimum distance between trees >20 m) were sampled in each population (total N=192 individuals). Fresh needles were dried in silica gel and then stored at -80°C until DNA extraction. *Abies* spp. samples used for transferability tests were collected either *in situ*, in provenance trials or in the arboretum of the botanical garden at the University of Marburg (Germany).

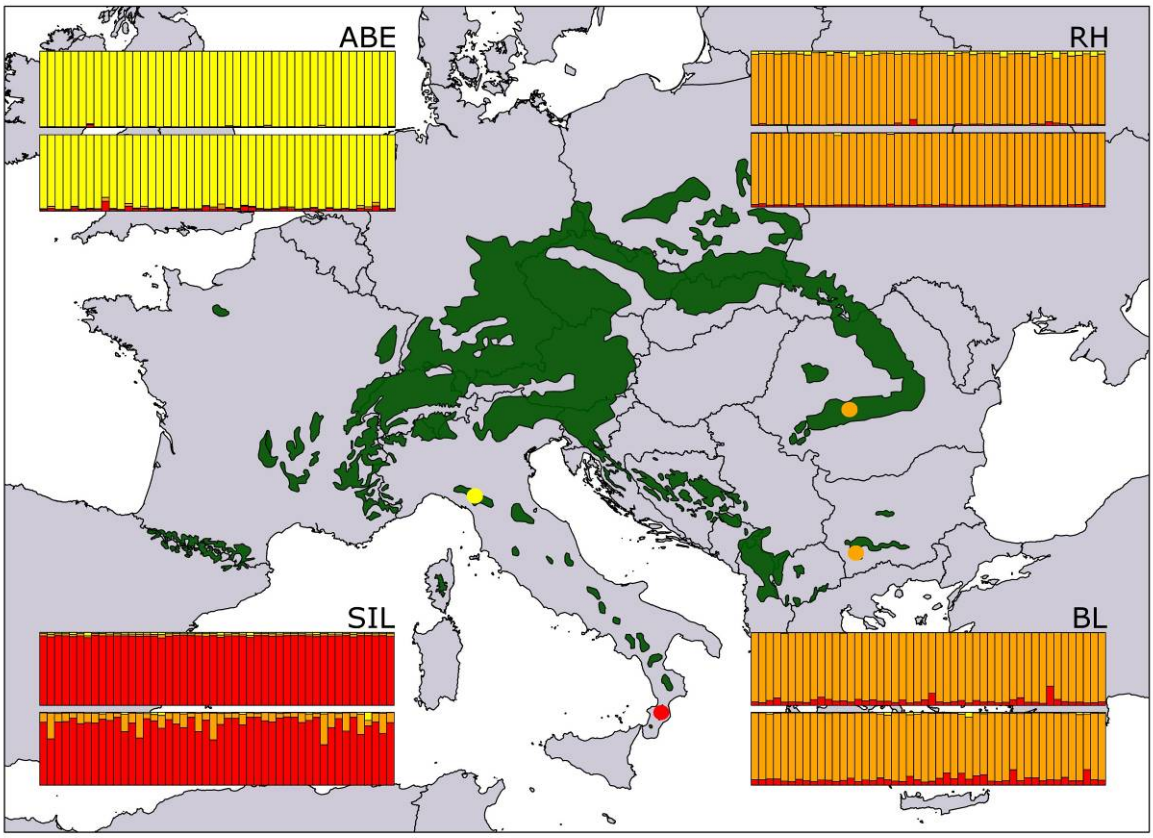

Figure 4.1: Map of the 4 populations sampled (see text for details) and results of clustering analysis using STRUCTURE (K=3). For each population, upper barplots refer to tSSRs and lower barplots to gSSRs. Population code names: ABE (Abetone, Northern Italy), SIL (Sila, Southern Italy), RH (Arges, Romania) and BL (Bansko, Bulgaria). Green colored areas indicate the distribution range of *Abies alba* (Wolf 2003).

DNA isolation

DNA extraction was performed from 50 mg of frozen needles with the DNeasy 96 Plant Kit (Qiagen) following the manufacturer's instructions. For disruption of plant material I added a 3-mm diameter tungsten bead to each well of the 96-well plates. Plates were frozen in liquid nitrogen for 30 seconds before two cycles of 1 min disruption at 25 Hz using a Mixer Mill MM300 (Retsch, Germany). DNA quality was estimated on a 1% agarose gel stained with

GelRed (Biotium, USA). DNA concentration was measured using a spectrophotometer NanoDrop ND-1000 (Thermo Scientific, Wilmington, USA).

Multiplex PCR optimization

SSRs from the transcriptome (EST-SSRs). EST-SSR discovery was carried out based on the analysis of assembled contigs from a transcriptome of *A. alba* (Accession numbers: JV134525-JV157085; Roschanski *et al.* 2013) using the SPUTNIK software (http://espressosoftware.com/sputnik/index.html). SPUTNIK finds perfect, compound and imperfect repeats using a recursive algorithm (Duran *et al.* 2009). The minimum number of repeats was set to six for di-SSRs and five for tri-, tetra- and penta-SSRs.

Sequence output from SPUTNIK was subsequently analyzed using Websat (http://www.wsmartins.net/websat/) to identify candidates with appropriate flanking regions suitable for primer design. Primers were designed using PRIMER3 (Rozen and Skaletsky 2000) applying the following parameters: product size between 100 and 500 bp, annealing temperature (T_a) between 57°C and 62ºC, optimum GC content = 50, maximum self-complementarity = 4.00, maximum 3' self-complementarity = 2.00, maximum Poly-X = 4.

A total of 67 EST-SSR primer pairs were designed and tested on a set of eight samples (four samples from ABE and four samples from BL) by PCR amplification. The PCR thermal profile was: denaturation at 94°C for 4 min, followed by 10 cycles at 94° C for 30 s, 63°C for 30 s (decreasing 1°C/cycle), and 72°C for 30 s, followed by 27 cycles at 94°C for 30 s, 53°C for 30 s, and 72°C for 40 s, with a final 10 min extension step at 72°C. PCR products were quality-checked on 2% agarose gels stained with GelRed (Biotium, USA).

The quality and polymorphism of all 67 EST-SSRs were first checked using the M13-tail labeling technique (Schuelke 2000). PCR products were analyzed on an ABI 3500 automatic sequencer (Applied Biosystems, USA) using LIZ-500 as internal size standard. To evaluate EST-SSR polymorphism, 48 samples (24 samples from ABE and 24 samples from BL), were genotyped with 24 EST-SSRs that exhibited high quality amplification and clear microsatellite peaks at the expected size. Excluding monomorphic loci, 16 EST-SSRs were finally selected and subsequently multiplexed by taking into account size ranges (Table 4.1). Mendelian segregation and the possible presence of null alleles were tested by progeny tests (Gillet and Hattemer 1989; Tarazi *et al.* 2010) on six open-pollinated progenies. For each progeny, six to 28 offsprings were genotyped using simplex PCR and the M13 tail labeling technique and their multi-locus genotype was compared to that of their seed-parent.

The Type-it Microsatellite PCR kit (Qiagen, Germany) was used to carry out multiplex

reactions. The final volume of PCR was optimized to 6 µl to reduce costs. The PCR mix for both multiplexes was: 3 µl Type-it Microsatellite Buffer, 2 µl of primers premix and 1 µl of DNA (~20 ng⁄µl). Concentrations and fluorescent dyes for each primer pair in the primer premix are presented in Table 1. Both EST-SSR multiplex sets had the same PCR thermal profile: an initial step at 95°C for 5 min, followed by 32 cycles at 95°C for 30 s, 57°C for 90 s and 72°C for 30 s, with a final 30 min extension step at 60°C. PCR products were run on an ABI 3500 automatic sequencer (Applied Biosystems, USA), with LIZ-500 as internal size standard. The match between simplex and multiplex profiles of all 192 samples was also checked to control for allele amplification competition and possible allelic drop-out. Chromatograms were analyzed using GeneMapper v4.1 (Applied Biosystems, USA).
Using the same PCR conditions as above, I tested the transferability of the newly developed EST-SSRs to ten Mediterranean, three North-American and four Asian *Abies* species and sub-species and one to seven individuals per taxon (Table 4.2).

SSRs from genomic DNA. In this panel, I included six gSSRs developed by Hansen *et al.* (2005) and Cremer *et al.* (2006), some of which had already been multiplexed (Hansen *et al.* 2008; Cremer *et al.* 2012; Gömöry *et al.* 2012), and two gSSRs developed from an *Abies alba* enriched library (Malausa *et al.* 2011). They were selected from two sets of 12 and six gSSRs, respectively, according to their quality and polymorphism tested on eight samples (four samples from ABE, and four samples from BL). The PCR thermal profile was: denaturation at 94°C for 4 min, followed by 10 cycles at 94° C for 30 s, 61°C for 40 s (decreasing 1°C/cycle), and 72°C for 40 s, followed by 29 cycles at 94°C for 30 s, 51°C for 40 s, and 72°C for 45 s, with a final 10-min extension step at 72°C. The quality-check of PCR products followed the same procedure as for EST-SSRs.
All gSSRs were first validated using M13-tail labeling technique and then multiplexed according to their size range.(Table 1). The multiplex reactions, PCR amplification and sizing of PCR products were carried out as for EST-SSRs above, except that in the PCR thermal profile the annealing temperature was 59°C for 60 s.

Genotype scoring and data analyses
All population genetic analyses were carried out on the whole dataset using both the EST-SSR and gSSR multiplexes. To estimate the error rates, the 192 samples screened with the 16 EST-SSRs and eight gSSRs were scored by two readers and Type A and B errors were estimated. Type A error refers to the case when a heterozygote is mistaken for a homozygote, or vice

versa. Type B error refers to a wrongly scored allele. In such cases, a final decision was made by joint agreement.

The software GenAlEx v6.5 (Peakall and Smouse 2012) was used to assess genetic diversity of the four *A. alba* populations. For each population, the total number of alleles (A), observed (H_O) and expected heterozygosity (H_E) and the fixation index (F_{IS}) were calculated at each locus. The program INEst (Chybicki and Burczyk 2009) was used to estimate the frequencies of null alleles in the dataset, running the individual inbreeding model (IIM) with a Gibbs sampler of 10^5 iterations. Computation of allelic richness (A_R) was carried out using the program HP Rare (Kalinowski 2005) in order to make it independent from sample size. Rarefaction was carried out with a common total sample size of 64 genes (32 diploid individuals). The software GENEPOP v4.2.1 was used to test for genotypic disequilibrium among loci using log likelihood ratio statistics (Rousset 2008) and Markov chain parameters provided by default.

Single parent and parent pair exclusion probabilities were calculated from allele frequencies according to the formula by Jamieson and Taylor (1997) using FaMoz (Gerber *et al.* 2003). Differentiation indices (Jost's D and Weir and Cockerham F_{ST}) were estimated using the diveRsity package in R (Keenan *et al.* 2013) and GenAlEx v6.5 (Peakall and Smouse 2012), respectively.

A Bayesian clustering approach was used to detect population genetic structure using the software STRUCTURE v2.3 (Pritchard *et al.* 2000; Falush *et al.* 2003). The admixture model was used, in which the fraction of ancestry from each cluster is estimated for each individual and allowed for correlated allele frequencies, as well as the "locprior" option when population identity is used as *a priori* information for clustering. Five independent runs for each K value ranging from 1 to 7 were performed after a burn-in period of 10^4 steps followed by 5×10^4 Markov Chain Monte Carlo replicates. To identify the number of cluster (K) that best explained the data, the rate of change of L(K) (ΔK) between successive K values was calculated following Evanno *et al.* (2005) using the web application "StructureHarvester" (Earl and von Holdt 2012).

Functional annotation EST-SSRs

Accurate functional annotation of silver fir EST-SSRs is a difficult task due to the limited availability of reference genome/gene sequences in public databases for conifer species. In order to maximize successful annotation, assembled contigs containing EST-SSRs were compared against four different databases (ConGenIE; GenBank (Benson *et al.* 2013);

UniProtKB/TrEMBL and UniProtKB/Swiss-Prot (The UniProt Consortium, 2012)) by using BLASTx (E-value cut-off: $< 10^{-3}$), and they were searched for protein family.

4.3 Results and Discussion

Multiplex PCR optimization

I detected 2150 putative EST-SSRs. This relatively low number is in agreement with previous studies where a negative correlation between SSR frequencies and the genome size was found, suggesting that it may be challenging to develop a large number EST-SSRs for conifers (Ueno *et al.* 2012). Based on the Websat analysis, I selected and tested 67 EST-SSRs with single non-interrupted, non-compound motifs and with the highest number of repeats, which are expected to display high polymorphism (Petit *et al.* 2005).

From the original set of 67 EST-SSRs tested in simplex reactions, 16 were retained, which amplified seven di- and nine tri-nucleotide SSRs. In this selection process, I removed 41 markers because of no or very poor PCR amplification, five because of multi-banding patterns, and five because of no polymorphism. This rather low success rate (24%) is comparable to those observed in other conifer species (e.g. Pfeiffer *et al.* 1997; Pinzauti *et al.* 2012; Sebastiani *et al.* 2012; Wagner *et al.* 2012), which are characterized by large genomes, partly due to large gene families and abundance of pseudo-genes and partly due to a very high content of repetitive DNA such as transposable elements (Kovach *et al.* 2010). The primer sequences of the 16 selected markers and their main characteristics are reported in Table 4.1. In addition to high quality allele binning, the 16 EST-SSRs were selected because progeny tests confirmed their Mendelian segregation and the very low number of mother-offspring mismatches indicated a low frequency of null alleles.

The 16 EST-SSRs were assembled in two 8-plexes (multiplexes A and B). I limited the number of loci included in multiplexes to avoid possible overlap of alleles from different loci due to their high size ranges. Electropherograms and marker size ranges for the two multiplexes are shown in Fig. 4.2.

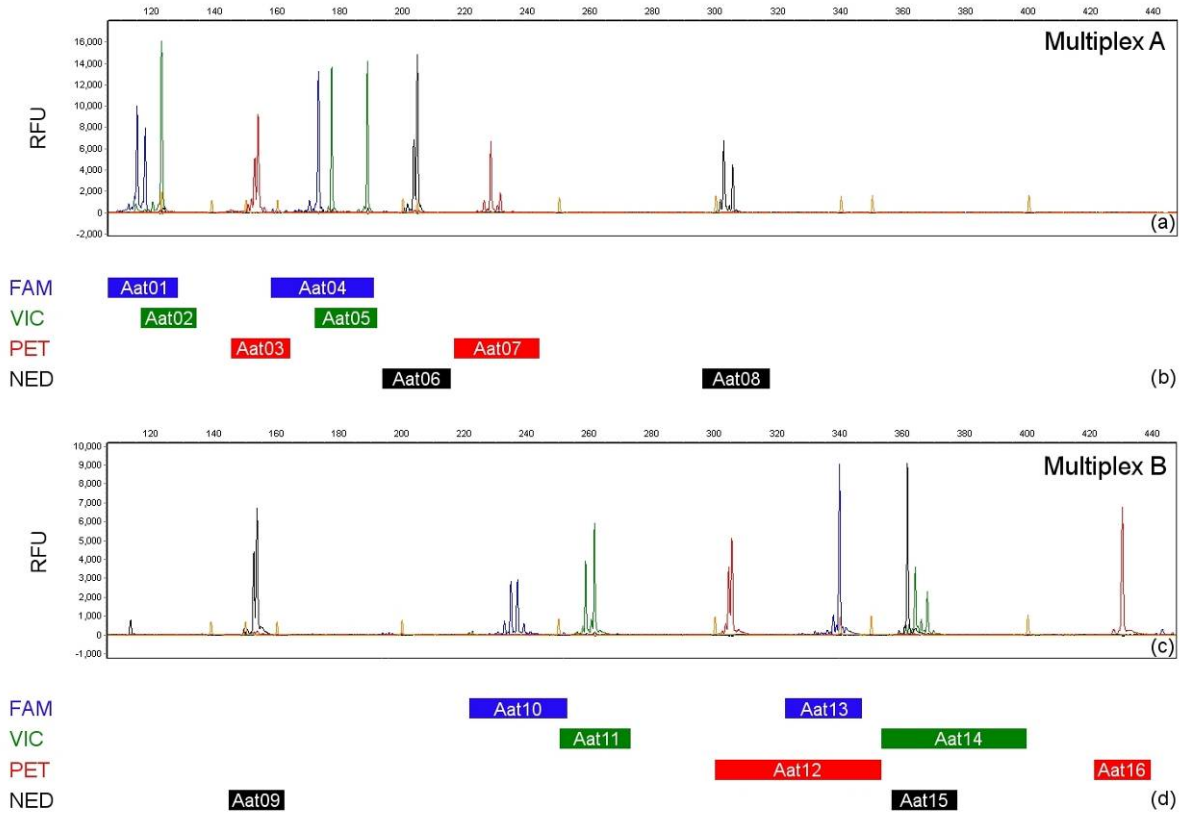

Figure 4.2: Examples of an individual electropherogram for multiplexes A and B (panels a and c, respectively) and marker ranges for multiplexes A and B (panel b and d, respectively).

Both multiplexes had high amplification quality and ample polymorphism. The comparison between single- and multi-plex amplification did not reveal allele dropout. In addition, combining already available and newly developed gSSRs, an 8-plex, showing high quality amplification and high polymorphism, was also successfully designed for gSSRs (multiplex C, Table 4.1).

Table 4.1 Characteristics of multiplexes A, B and C based on 192 *A. alba* individuals from the four populations sampled.

Locus	Reference	Primer sequences (5'→3')	Motif	Dye	Size (bp)	[C]	*A*	H_O	H_E	F_{IS}	Accession No.
EST-SSRs											
Multiplex A											
Aat01	This study	F: CCATGTCTCCGATTTCCAGT R: GGCCTAACGAAAGCAGAATC	$(GCG)_{10}$	FAM	103-127	0.20	7	0.581	0.517	-0.125	KF304594
Aat02	This study	F: AGAAGATTTCCCGGCTTTTC R: ATCCAGACAGCGAACTTTGG	$(CAG)_7$	VIC	123-129	0.06	3	0.312	0.331	0.057	KF304595
Aat03	This study	F: TCCCCATGGTTTGGTTAAAA R: CGAAGAAAATGTTGCGGAAT	$(AT)_9$	PET	149-161	0.10	6	0.476	0.515	0.075	KF304596
Aat04	This study	F: CCATGTATGGTGCTCCTCCT R: CCTTCATTGCAGAAAAGCAA	$(CAG)_{11}$	FAM	158-191	0.27	9	0.423	0.404	-0.046	KF304597
Aat05	This study	F: AGCATCCACATTCCGTAACC R: AGTTGACCGTTGGAGAGCAG	$(GCA)_7$	VIC	177-192	0.06	3	0.280	0.247	-0.135	KF304598
Aat06	This study	F: TTATGCGGAGCAGTTCTGTG R: TGTTGCTGGCGTACTGGTAG	$(GCA)_8$	NED	196-214	0.20	5	0.115	0.113	-0.019	KF304599
Aat07	This study	F: GCTAGCAGAACCCTGGAATG R: GGTGGGATATTTCCAGCAAG	$(AT)_{11}$	PET	219-241	0.10	10	0.556	0.656	0.154	KF304600
Aat08	This study	F: ACTCCATCACGGTGGTCTTC R: GCCATTCAGGCTCTCAGTTC	$(AT)_9$	NED	302-312	0.08	3	0.171	0.163	-0.048	KF304601
Multiplex B											
Aat09	This study	F: CAGATCCTCCCACATCCAAC R: TGACACCACAGGAAACCATC	$(TCA)_8$	NED	150-156	0.05	3	0.032	0.031	-0.016	KF304602
Aat10	This study	F: GAGCACGATGAAGAGGAAGC R: AAAACCCCCACGCGGTAT	$(AT)_{12}$	FAM	226-250	0.25	13	0.625	0.656	0.047	KF304603
Aat11	This study	F: AGCGTTGATTGGAAGCAGTC R: GAAGCATGGTGTCGTTGTTG	$(AAC)_9$	VIC	255-270	0.08	5	0.561	0.535	-0.048	KF304604
Aat12	This study	F: ATCCATATCTCCTGCCTTGC R: CTTTCCAGGTGATCTGATTGC	$(AG)_{12}$	PET	303-349	0.21	19	0.610	0.600	-0.016	KF304605
Aat13	This study	F: ACTCAAAGCCAAGCTGGAGA R: TGCATAAGACAGCCGAGTCA	$(AG)_8$	FAM	326-342	0.30	4	0.163	0.180	0.093	KF304606
Aat14	This study	F: GACTGGGGATCCTGCTGTTA R: AGAGGAGGCAGCCCATACAT	$(TA)_9$	VIC	358-394	0.13	16	0.734	0.749	0.020	KF304607
Aat15	This study	F: AGGAGGAGGTTCAGCATGTC R: CTTGCTCTCTGACCCAGTTG	$(AGA)_8$	NED	361-373	0.08	4	0.133	0.132	-0.006	KF304608
Aat16	This study	F: AACCACCGCTGATATTTTGG R: GGGTTCAAGAAATGGGAATG	$(GAA)_7$	PET	427-430	0.20	2	0.269	0.288	0.067	KF304609

gSSRs											
Multiplex C											
SFg6	Cremer *et al.* 2006	F: GTAACAATAAAAGGAAGCTACG R: TGTGACACATTGGACACC	$(AC)_9$	VIC	103-111	0.11	5	0.332	0.577	0.425	DQ218456
SF324	Cremer *et al.* 2006	F: TTTGAACGGAAATCAAATTCC R: AAGAACGACACCATTCTCAC	$(CCG)_8$	PET	105-120	0.24	5	0.296	0.473	0.374	DQ218461
NFF7	Hansen *et al.* 2005	F: CCCAAACTGGAAGATTGGAC R: ATCGCCATCCATCATCAGA	$(GA)_{33}$	VIC	116-174	0.13	26	0.857	0.896	0.043	AY966495
SFb5	Cremer *et al.* 2006	F: AAAAAGCATCACTTTTCTCG R: AAGAGGAGGGGAGTTACAAG	$(CT)_{15}$	FAM	138-160	0.20	10	0.373	0.713	0.477	DQ218455
SFb4	Cremer *et al.* 2006	F: GCCTTTGCAACATAATTGG R: TCACAATTGTTATGTGTGTGG	$(GT)_{16}$	NED	149-205	0.30	25	0.667	0.865	0.229	DQ218454
Aag01	This study	F: GCTTATTCTCACTGCTCGCC R: ATGACTTGAAGGTGGATGCC	$(CTT)_{15}$	PET	193-250	0.15	13	0.804	0.768	-0.046	KF304592
SF1	Cremer *et al.* 2006	F: TTGACGTGATTAACAATCCA R: AAGAACGACACCATTCTCAC	$(CCG)_9$	VIC	208-229	0.17	6	0.555	0.511	-0.086	DQ218453
Aag02	This study	F: TATTCCTCCACTTGGGTGCT R: GGTGGAGATCCGTATGCAAT	$(GA)_{13}$	FAM	208-250	0.37	19	0.363	0.855	0.575	KF304593

[C], final concentration in each primer premix [μM]; A, number of alleles; H_O and H_E, observed and expected heterozygosities; F_{IS}, inbreeding coefficient.

Genotype scoring and analysis

Binning of alleles was consistent across the whole dataset, indicating that loci display allele sizes according to the expected di- and tri-nucleotide repeat variation. Two loci, Aat07 and Aat08, showed intermediate size variants (1-bp variation) for some individuals which could be clearly distinguished from the other size classes, thus having no impact on binning precision. The possibility to correctly and easily score additional variants can increase the precision of the analysis (Guichoux *et al.* 2011) and is an indication of the existence of mutation types other than insertion/deletion of SSR motifs within the sequence or the flanking regions (Barthe *et al.* 2012). At locus Aat12, more than two amplification products in single individuals (up to 4) in the 2 Italian populations were detected. This could be due to a duplication of this locus in some populations, but further analyses are needed to confirm this hypothesis.

Type A error ranged from 0 to 0.53 % and type B error ranged from 0 to 0.26 % across the whole EST-SSR dataset. Mean type A error was 0.40 % for multiplex A and 0.15 % for multiplex B, whereas mean type B error was 0.20 % for multiplex A and 0.11 % for multiplex B. Higher type A and B error rates have been observed for the gSSR multiplex (type A error ranged between 0 and 2.6 %, type B error ranged between 0 and 1.6 %), due to the higher stuttering displayed by some loci which made the reading less straightforward. The error rates were significantly reduced by adopting well-defined reading rules.

The 16 EST-SSRs selected were all polymorphic and displayed a low to moderate level of diversity. Observed (H_O) and expected (H_E) heterozygosity and inbreeding coefficient (F_{IS}) per locus are reported in Table 1. H_O ranged between 0.115 and 0.581 for multiplex A, and 0.133 and 0.734 for multiplex B. H_E was between 0.113 and 0.656 for multiplex A and between 0.031 and 0.749 for multiplex B. The number of alleles per locus ranged from 2 to 19 and the mean number of alleles was 5.75 and 8.25 for multiplex A and B, respectively. Allelic richness varied between 1.75 (Aat08) and 5.81 (Aat07), and between 1.90 (Aat09) and 9.10 (Aat12), for multiplex A and B respectively (see Fig. 3). By contrast, gSSRs exhibited higher H_O and H_E values, with a maximum up to 0.857 and 0.896 respectively, as well as a higher number of alleles per locus (Table 4.1) and a higher allelic richness (Fig. 4.3), a result already reported in other studies (e.g. Sullivan *et al.* 2013). Lower diversity at EST-SSR than gSSR loci can be explained by the higher degree of conservation of the transcribed regions of the genome.

Out of the possible 120 combinations involving the EST-SSR loci, no significant linkage disequilibrium was detected among loci ($P<0.05$). When linkage disequilibrium analysis was

performed on the total number of SSRs (16 EST-SSRs + 8 gSSRs), out of the possible 276 combinations, only three (about 1%) were significant ($P < 0.05$).

Weir and Cockerham's F_{IS} values at EST-SSR loci were generally low or slightly negative. This suggests a low frequency of null alleles, which was confirmed by the non-significant null allele frequencies estimated at all loci in the 4 populations (Fig. 4.3). When inbreeding coefficients were estimated taking into account null allele frequencies by INEST, I found no F_{IS} significantly different from 0 at the population level. Therefore, positive and high Weir and Cockerham's F_{IS} values estimated at some gSSRs (Table 4.1) are likely to be the consequences of a high frequency (>20%) of null alleles (Fig. 4.3). Discarding the most null allele prone loci or making adjustments for the presence of null alleles will thus be necessary when using gSSRs for estimating diversity and differentiation parameters (Chapuis and Estoup 2007) or for paternity and parentage analyses (Oddou-Muratorio *et al.* 2009; Piotti *et al.* 2012). The higher mutation rate expected for the non-coding portion of the genome, affecting also the annealing sites, is possibly the main reason for the higher frequency of null alleles at gSSR than EST-SSR loci (Kovach *et al.* 2010).

The estimates of differentiation among populations were generally slightly higher for EST-SSRs (F_{ST} up to 0.243 for locus Aae08) than for gSSRs (F_{ST} up to 0.120 for SFg6; on average $F_{ST} = 0.087$ and 0.065 at EST-SSRs and gSSRs, respectively). Lower differentiation estimates for gSSRs are expected due to the lower frequency of the most frequent alleles and the higher within-population genetic diversity (Jakobsson *et al.* 2012). However, when polymorphism within locus was taken into account, population differentiation measured using Jost's D was smaller for EST-SSRs than for gSSRs (Fig. 4.3).

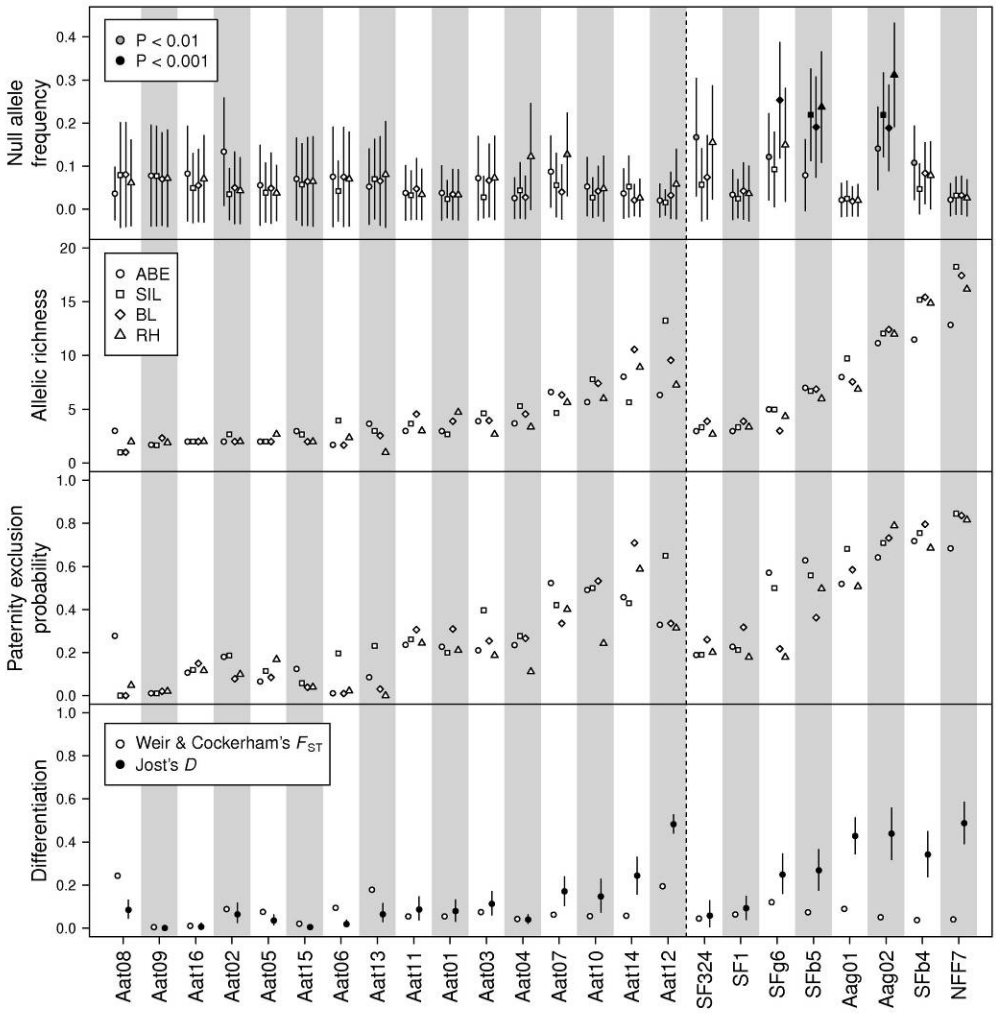

Figure 4.3: Null allele frequencies, allelic richness, and paternity exclusion probabilities for the 16 EST-SSRs and the 8 gSSRs analyzed in each population. In the bottom panel, a comparison between Weir and Cockerham F_{ST} and Jost's D in presented for each marker analyzed.

The Bayesian clustering approach revealed a clear geographic pattern, with a best grouping at K=3 (Fig. 4.1). There was a clear separation between the Italian and Balkan gene pools as well as within the Italian gene pool, reflecting different quaternary histories (see Cheddadi *et al.* 2013 for a recent synthesis). Interestingly, gSSRs showed a very similar pattern, although with a slightly higher degree of admixture (Fig. 4.1).

The exclusion probability was higher for gSSRs than for EST-SSRs (Fig. 4.3), as a result of a greater number of alleles at genomic loci. However, assignment biases related to null alleles suggest the use of a large enough set of EST-SSRs or a carefully selected set of EST-SSRs and gSSRs not affected by the presence of null alleles for population genetic studies. As an example, in the SIL population paternity exclusion probabilities > 0.999 can be achieved by

using only 6 markers (NFF7, SFb4, Aag01, Aat12, Aat10, Aat14) seemingly not affected by null alleles.

The 16 EST-SSRs selected have been also tested for possible outliers using the Bayesian test of Foll and Gaggiotti (2008), setting the parameters as in Soto-Cerda and Cloutier (2013), and using the software BayeScan v2.1 (http://cmpg.unibe.ch/software/bayescan/). BayeScan, which is based on what is recognized as the best approach to avoid detection of false positives (Pérez-Figueroa *et al.* 2010; Narum and Hess 2011), did not identify any outliers.

Functional annotation of EST-SSRs

Functional annotation of contigs containing 16 *A. alba* EST-SSRs revealed that seven contigs had homology with known proteins and five contigs had homology with putative proteins, as shown in Table 4.3.

Table 4.2 Transferability of 16 *A. alba* EST-SSRs into 17 congeneric taxa from the Mediterranean (first 10 taxa, sections *Abies* and *Piceaster*), Asia (taxa 11 to 14, sections Momi and Balsamea) and America (last 3 taxa, sections Balsamea and Grandis). (-): no amplification; (+/-): some amplification, but optimization is needed; (+): successful high quality amplification; (++): successful high quality amplification and locus is polymorphic. N: number of samples analyzed per taxon. Taxonomic reference: USDA, ARS, National Genetic Resources Program, Germplasm Resources Information Network - (GRIN), except *A. borisii-regis* (IUCN Red List of Threatened Taxa)

Species	Aat01	Aat02	Aat03	Aat04	Aat05	Aat06	Aat07	Aat08	Aat09	Aat10	Aat11	Aat12	Aat13	Aat14	Aat15	Aat16	transferability rate
A. borisii-regis Mattf.	+	++	++	++	+	+	++	+	+	++	++	+	+	+	++	+	1
A. cephalonica Loudon	++	++	++	++	+	+	++	+	+	++	+	++	++	++	+	+	1
A. nordmanniana (Steven) Spach	++	++	+	++	+	+	++	++	++	++	++	++	++	++	++	++	1
A. nordmanniana subsp. equi-trojani (Asch. and Sint. ex Boiss.) Coode and Cullen	+/-	+	++	+	+	+	+/-	+	+	++	+	++	++	++	+	+	1
A. nordmanniana subsp. bornmuelleriana (Mattf.) Coode and Cullen	++	++	++	++	+	+	++	+	++	++	++	++	++	++	+	+	1
A. nebrodensis (Lojac.) Mattei	++	++	++	++	++	++	++	++	++	++	++	++	++	++	++	++	1
A. cilicica (Antoine and Kotschy) Carrière	++	+	++	++	++	+	+/-	++	++	+/-	++	++	+	++	+	+	1
A. pinsapo Boiss.	++	++	++	+	+	+	+/-	+/-	++	++	++	+/-	++	++	+	+	1
A. pinsapo var. marocana (Trab.) Ceballos and Bolaños	++	++	+	+	+	+	+/-	+	+	+/-	+	++	++	+/-	+	+	1
A. numidica de Lannoy ex Carrière	+	++	++	++	++	+	++	+	+	++	+	++	+	++	+	+	1
A. recurvata Mast.	+	++	+	+	+	+	+	+	+	+	+	++	+	++	+	+	1
A. sibirica Ledeb.	+	+	-	++	+	++	-	+	++	++	++	-	++	+	+	+	0.81
A. veitchii Lindl.	++	+	-	++	++	++	-	+	++	+	+	+/-	+	-	+	+	0.81
A. koreana E. H. Wilson	++	+	-	+	+	+	-	+	+	+	+	-	+	-	+	+	0.75
A. lasiocarpa (Hook.) Nutt.	+	++	-	+	+	++	-	++	+	++	++	+	+	-	+	+	0.81
A. concolor (Gordon and Glend.) Lindl. ex Hildebr.	++	++	-	++	++	+	-	+	+	+	+	++	++	+	+	+	0.88
A. grandis (Douglas ex D.Don) Lindl.	+	++	-	+	+	+	-	+	+	++	+	++	-	+	+	+	0.81

Table 4.3. Functional annotation of assembled contigs containing 16 *A. alba* EST-SSRs using BLASTx. (PF) : Protein Family (-) No hit

Locus	Protein name	UniProtKB/ Swiss-Prot	E-value	NCBI Conserved Domains	E-value	UniProtKB	E-value	ConGenIE	E-value
EST-SSRs									
Multiplex A									
Aat01	Indole-3-acetic acid-induced protein ARG7	PF02519	8.0E-15	PF02519	1.20E-14	PF02519	1.0E-42	PF02519	2.7E-29
Aat02	Putative uncharacterized protein	-	-	-	-	PF00096	6.0E-38	PF00096	2.79E-34
Aat03		-	-	-	-	-	-	-	-
Aat04		-	-	-	-	-	-	-	-
Aat05	Putative uncharacterized protein	-	-	-	-	PF03479	2.0E-14	-	-
Aat06	Growth-regulating factor 3	PF08879 PF08880	5.0E-12	PF08880	1.33Ee-06	PF08879 PF08880	2.0E-14	PF08879 PF08880	7.03E-62
Aat07	Protochlorophyllide reductase, chloroplastic	PF00106	7.0E-23	PF00106	0E+00	PF00106	2.0E-25	PF00106	5.71E-23
Aat08		-	-	-	-	-	-	-	-
Multiplex B									
Aat09	Putative uncharacterized protein	-	-	-	-	PF00096	8.0E-36	PF00096	1.33E-28
Aat10	Putative uncharacterized protein	-	-	-	-	PF00892	6.0E-21	PF00892	1.67E-17
Aat11	B2 protein	PF10539	1.0E-14	PF10539	4.93E-03	PF10539	6.0E-52	PF10539	4.96E-12
Aat12		-	-	-	-	-	-	-	-
Aat13	30S ribosomal protein 3-1, chloroplastic	PF04839	9.0E-42	PF04839	5.79E-26	PF04839	7.0E-47	PF04839	1.24E-40
Aat14	Putative uncharacterized protein	-	-	-	-	PF00304	1.0E-20	PF02493	5.04E-28
Aat15	31 kDa ribonucleoprotein, chloroplastic	PF00076	6.0E-6	PF00076	4.85E-03	PF00076	7.0E-60	PF00076	1.28E-55
Aat16	17.7 kDa class II heat shock protein	PF00011	2.0E-21	PF00011	6.28E-09	PF00011	1.0E-24	PF00011	3.36E-34

EST-SSR transferability

The cross-transferability of the newly developed EST-SSRs was high and reflected the degree of relatedness among taxa (Table 4.2). In particular, the amplification rate was 100% for the eight Mediterranean *Abies* taxa belonging to section *Abies* (same as *A. alba*) but also for *A. numidica* and *A. pinsapo* (section *Piceaster*) and the Asian fir *A. recurvata* (section *Momi*). It ranged from 75% to 81% for North American firs (sections *Balsamea* and *Grandis*) and Asian firs (section *Balsamea*). Most of the markers appeared to be polymorphic across the different taxa when sample sizes were large enough for tests to be made. Some amplifications were of poor quality (+/- in Table 4.2). This could be due either to low DNA quality or a need for optimization. Although Mendelian segregation analysis and additional amplifications still need to be performed on a larger sample size, our pilot transferability and polymorphism results across the genus *Abies* are very promising and suggest the usefulness of the EST-SSRs I developed for population genetic studies in this genus.

4.4 Conclusion and perspectives

The two EST-SSR multiplexes designed for *A. alba* allow fast, cost-effective and accurate genotyping of a large number of individuals and populations. The time spent for their optimization is significantly compensated by the accuracy of allele binning which allows for rapid and efficient screening of large sample sizes. The two newly developed EST-SSR multiplexes are currently applied on a range-wide sample of 28 Italian populations (each represented by 50 individuals) to resolve in detail their past population dynamics, which is expected to be more complex than previously hypothesized (Cheddadi *et al.* 2013). The comparison of two EST-SSR multiplexes to a control set of gSSRs revealed their lower diversity and frequency of null alleles, as expected considering the lower mutation rates assumed in the coding portion of the genome (Kovach *et al.* 2010). Careful selection of markers from the three multiplexes will facilitate the identification of the best combination to get the highest possible exclusion probability in gene flow studies. In general, these newly developed SSRs will be useful for conservation genetic studies and to improve our knowledge about population dynamics of *Abies* species.

The EST-SSR markers developed here showed a high transferability rate across the genus *Abies*, even for phylogenetically distant species. Although additional optimization should be performed and polymorphism should be assessed in a larger sample, these first results look very promising for population genetic studies within the genus *Abies*.

It should also be stressed that these EST-SSRs, being less prone to homoplasy (due to their putative lower mutation rates) and highly transferable across species, could also be used for phylogenetic analysis, as already shown in the genus *Epimedium* (Zeng *et al.* 2010). Moreover, considering the BLASTx results (Table 4.3), some of the EST-SSRs might be linked to genes involved in controlling important traits, thus providing a potentially powerful tool for genetic mapping.

Chapter 5:

Biogeographical patterns of silver fir (*Abies alba* Mill.) in the Apennines

5.1 Introduction

Silver fir (*Abies alba* Mill.) is one of the most important forest species in Europe and several articles aiming at describing his Quaternary history have been published. Contrasting results from available palaeobotanical and genetic studies (e.g. Huntley and Birks 1983, Liepelt *et al.* 2009, Linares 2011, Cheddadi *et al.* 2013) showed how the biogeographical patterns of silver fir in the Apennines are highly complex and peculiar, a likely outcome of a complex evolutionary history shaped by several interacting factors. The paucity of adequate data from this geographic area (generally only few populations per study were sampled and analyzed) adds further uncertainty to the different hypotheses put forward so far. It is urgent to unravel the evolutionary history of fragmented silver fir populations in the Italian peninsula because this refugial area has pivotal evolutionary, ecological and conservation value due to its genetic and eco-physiological peculiarity (Hansen and Larsen 2004, Carrer *et al.* 2010, Cheddadi *et al.* 2013).

One major factor in determining the present distribution of forest tree genetic diversity is constituted by the Quaternary cycles of range retraction and expansion following glaciations (Petit *et al.* 2003). Paleoecological and genetic studies have indicated the presence of putative glacial refugia for silver fir in the Apennines, and that these refugia may have acted as starting point of the post-glacial expansion from the Italian Peninsula towards Central Europe (Konnert and Bergman 1995, Terhurne-Berson *et al.* 2004, Liepelt *et al.* 2009, Cheddadi *et al.* 2013). Despite many hypotheses have been formulated, the clear identification and location of refugial areas in the Apennines is still uncertain (Cheddadi *et al.* 2013). In particular, one important point that still needs to be clarified is the distinction between isolated refugia (i.e. refugia that did not expand after the Ice Ages) and refugia that acted as starting points during the recolonization process. While the latters contributed to the current genetic diversity pattern throughout Europe, isolated refugia did not, but they may now represent valuable reservoirs of genetic diversity (Petit *et al.* 2003, Liepelt 2009, Cheddadi *et al.* 2013).

Another factor that makes the reconstruction of the evolutionary history of silver fir in this area a complex task is that the Apennines have been historically affected by human activities that might have markedly altered the natural dynamics of forest tree populations. Silver fir has

been greatly affected by human exploitation. Historical data indicate that monastic orders have intensively managed conifer forests since 1000 AD in the Central Apennines (Urbinati and Romano 2012). Forest management through coppicing has also favoured beech, causing the substitution of former mixed stands with pure beech forests. In addition, forest management also led to the plantation and spread of material of unknown origin (e.g. one of the largest silver fir forests in Northern Apennines, Foreste Casentinesi, which is currently used as a seed source for reafforestation, is partly of Boemian origin, Cavagna and Cian 2003). In some areas of the Apennines (e.g. North-Western Apennines) only small relic autochthonous populations exist nowadays. These populations are highly fragmented and located in remote areas characterized by harsh environmental conditions, often surrounded by silver fir plantations (Piovani *et al.* 2010).

Apennine populations represent the rear-edge of the current silver fir distribution. These populations show genetic peculiarity (Vendramin *et al.* 1999, Parducci *et al.* 2001) and they are likely to be long-term stores of the species genetic diversity, especially for genes linked to warm- and drought-adaptation (e.g. Larsen and Mekic 1991, Grivet *et al.* 2011). Populations from Southern Apennine have been shown to have high genetic diversity and to be genetically distinct compared to populations from Central Europe (Vendramin *et al.* 1999, Parducci *et al.* 2001). Using a set of chloroplastic microsatellites, Parducci *et al.* (2001) found an F_{ST} among Calabrian and German populations equal to 0.19, confirming previous results (Vendramin *et al.* 1999). On the other hand, rear-edge populations often suffer from to the negative consequences of habitat fragmentation (Eckert 2008, Aguilar *et al.* 2008), i.e. genetic drift-related effects, increased inbreeding effect, demographic contraction, population isolation, making these populations more vulnerable (Piovani *et al.* 2010; Maiorano *et al.* 2013). In addition, peripheral populations may also suffer from the negative effect of maladaptive asymmetric gene flow from the centre of the species distribution (Davis and Shaw 2001). With the current climate change, gene flow from central populations to the southernmost ones could lead to the introduction of unfit alleles, not suitable for facing increasingly warmer or drier conditions, thus reducing their fitness and increasing the risk of extinction of the population.

The geography itself makes the Apennines a hotspot of diversity. The Apennines are a mountain chain with a 1000 km long latitudinal span. This broad latitudinal gradient includes a wide variety of ecological, environmental and climatic conditions, with many clines but also abrupt changes. Silver fir populations growing along the Apennines face different conditions (e.g. from montane to mediterranean environments, from pure to mixed stands, Carrer *et al.* 2010) and are likely to experience different selective pressure. In addition, mountain peaks

and sudden changes in soil composition (from sandstone to limestone and granite) may also have acted as post-glacial barriers to gene flow and massive migration, leading to population isolation and differentiation.

Genetic studies based on a variety of molecular markers (from allozyme to mitochondrial and chloroplast sequences) and carried out at the biogeographical scale have shown marked differences among populations along the Italian peninsula (Konnert and Bergmann 1995, Liepelt *et al.* 2002, Liepelt *et al.* 2009, Piovani *et al.* 2010). Such differentiation has also been confirmed by eco-physiological and dendrochronological studies (Larsen and Mekic 1991, Hansen and Larsen 2004, Carrer *et al.* 2010), where populations belonging to different "latitudinal" regions were shown to have different eco-physiological responses in common environments. These findings suggest the existence of putative "eco-genetic" clusters (i.e. groups of nearby populations characterized by peculiar genetic and/or eco-physiological profiles) for silver fir along the Apennines. Nonetheless, the exact boundaries of these putative distinct clusters and the causes that led to their differentiation have not been clearly identified yet. The relatively high differentiation found in previous studies between relatively close populations (e.g. Konnert and Bergmann 1995, Vicario *et al.* 1995, Liepelt *et al.* 2009) is unusual for a forest tree species with high dispersal potential, where extensive gene flow - especially by pollen - is expected to homogenize allele frequencies at the biogeographical scale (Kremer *et al.* 2012). It is not clear whether these groups originated as the results of post-glacial recolonization events, due to drift following population fragmentation and isolation, due to more recent adaptation to local conditions, or to a mix of these putative causes.

This chapter aims at identifying putative eco-genetic clusters by investigating the spatial patterns of genetic variation. I overcome the issue of the general paucity and inadequacy of data from the Apennine region by sampling populations of silver fir across the whole Apennine range (*i.e.* from North-Western Apennine to Calabria), including populations from all the putative eco-genetic clusters identifiable from previously published studies, and populations from surrounding regions (i.e. the Alps and the Balkans).

This study also aims at clarifying the post-glacial history of silver fir populations in the Italian penisisula. In particular, two main hypotheses found in previous published studies are tested: i) the existence of a single glacial refugium in Southern Italy from where the recolonization of the whole Italian peninsula started (as reported by Linares 2011); ii) the presence of a second refugium in Northern Apennine and its role as a main contributor to the recolonization towards Central Europe (Cheddadi *et al.* 2013). Also the role of the Balkanian refugium and of contact zones among different migration routes will be evaluated.

To unravel the complex biogeographical patterns of silver fir in the Apennines and identify areas of genetic discontinuity is the starting point for the future study of the underlying evolutionary and ecological processes using spatial patterns of adaptive genetic diversity and for planning conservation and management actions in the Apennines.

5.2 Materials and methods

Plant material collection

Twenty-four putatively autochthonous populations were sampled: 16 in the Apennines (5 in the Northern, 5 in the Central, 6 in the Southern Apennine), 5 in the Alps (3 in Western and 2 in Eastern Alps) and 3 in the Balkans (1 from Serbia, 1 from Bulgaria and 1 from Romania (Fig. 5.1 and Tab. 5.1). Apennine populations are located along a 1000 km latitudinal stretch and were selected in order to cover the whole Apennine distribution. I aimed at sampling at least one population (ideally 2: one on top, one on bottom) from each of the putative different eco-genetic clusters identifiable from previous studies (Larsen and Mekic 1991, Konnert and Bergmann 1995, Parducci *et al.* 1996, Liepelt 2002, Carrer *et al.* 2010, Liepelt *et al.* 2010, Piovani *et al.* 2010) in order to maximize the chances to detect discontinuity areas and population differentiation. I included populations from the Alps and from the Balkan peninsula as outgroups because, according to previous findings (Liepelt *et al.* 2002, Terhurne-Berson *et al.* 2004, Liepelt *et al.* 2009), they should have experienced different evolutionary histories with respect to the Apennine ones. From each population I collected young needles from 50 adult individuals (N = 1200). Sampled trees were at least 20 meters apart. Individual spatial coordinates were recorded using a high precision GPS device. Fresh needles were dried with silica gel and then stored at -80°C until DNA extraction.

Figure 5.1: Location of the sampled populations. Population codes refer to Tab. 5.1.

Table 5.1: Location and geographical characteristics of the sampled populations

Code	Location	Latitude (°N)	Longitude (°E)	Mean altitude (m a.s.l.)	Region
ABE	Abetone (MO)	44° 08' 28"	10° 40' 02"	1391	Northern Apennine
BTR	Bocca Trabaria (PG-PU)	43° 36' 03"	12° 13' 32"	1004	
CER	Cerreto (RE)	44° 17' 16"	10° 14' 22"	1638	
NER	Monte Nero (PC)	44° 33' 28"	9° 30' 15"	1657	
PIG	Pigelleto, M. Amiata (GR)	42° 48' 43"	11° 39' 25"	799	
ABS	Pescopennataro (IS)	41° 51' 31"	14° 17' 16"	1481	Central Apennine
COR	Cortino (TE)	42° 37' 11"	13° 29' 26"	1304	
CPL	Ceppo (TE)	42° 40' 10"	13° 26' 10"	1416	
TOS	Tossicia (TE)	42° 31' 55"	13° 36' 33"	1362	
VCL	Valle della Corte (AP)	42° 42' 19"	13° 22' 24"	1370	
CIL	Cilento (SA)	40° 28' 06"	15° 26' 39"	1126	Southern Apennine
GAM	Aspromonte (RC)	38° 08' 37"	15° 50' 42"	1541	
LAU	Laurenzana (PO)	40° 24' 20"	15° 57' 33"	1115	
SIL	Sila Grande (CS)	39° 07' 54"	16° 38' 25"	1740	
SSB	Serra San Bruno (VV)	38° 33' 42"	16° 20' 59"	1089	
TNP	Terranova nel Pollino (PO)	39 57' 34"	16° 13' 22"	1128	
NOA	Val Noana (TN)	46° 07' 41"	11° 50' 52"	1129	Eastern Alps
TAR	Tarvisio (UD)	46° 29' 21"	13° 35' 59"	897	
PES	Val Pesio (CN)	44° 12' 40"	7° 40' 12"	1242	Western Alps
SAL	Val di Susa (TO)	45° 02' 55"	6° 53' 18"	1745	
TOC	Toceno (VB)	46° 10' 05"	8° 27' 31"	1409	
BLG	Pirin mountains (Bulgaria)	41° 50' 41"	23° 23' 04"	1213	Balkans
ROM	Fagaras mountains (Romania)	45° 26' 28"	24° 41' 41"	1477	
SER	Tara mountains (Serbia)	43° 56' 23"	19° 18' 33"	1161	

DNA isolation and SSRs amplification

DNA was extracted from 40 mg of frozen needles using the DNeasy 96 Plant Kit (Qiagen) according to the manufacturer's instructions. For disrupting the material, needles were first frozen in liquid nitrogen for 30 seconds and then ground on a Mixer Mill MM300 (Retsch, Germany) for 1 minute at 25 Hz using also one 3-mm diameter tungsten bead (Qiagen). Two complete disruption cycles were performed. DNA quality was estimated using 1% agarose gels stained with GelRed (Biotium, USA). DNA concentration was measured using a spectrophotometer NanoDrop ND-1000 (Thermo Scientific, Wilmington, USA).

For each population 48-50 individuals were genotyped at 20 microsatellite loci (Tab. 5.2). These loci include all 16 EST-SSRs and 4 gSSRs described in Chapter 4 of this thesis. From the set of 8 gSSRs I selected the 4 loci that displayed the highest amplification success and

the clearest band pattern: Sf324 and Sf1 from Cremer *et al.* (2006), NFF7 from Hansen *et al.* (2005), Aag01 described in Chapter 4 (Tab. 5.2). All details regarding the multiplexing and the amplification procedure for the 16 EST-SSRs are reported in Chapter 4 (see Tab. 4.1). Some adjustments were made for the amplification of the 4 gSSRs. Three loci (Sf324, NFF7, Aag01) were amplified in a 3-plex PCR, using the same primer concentration of the original protocol described in Chapter 4. The locus Sf1 was amplified in a single PCR performed in a 10-µl reaction volume containing 20 ng template DNA, 0.2 µM of each primer, 0.2 mM of each dNTP, 1 X Buffer (Promega), 2.5 mM $MgCl_2$, 1 U of Taq (Promega). The forward primer of Sf1 was labeled with FAM dye (previously it was labeled with VIC). For this marker, I used a touch-down PCR program, performed as follows: an initial step at 94°C for 3 min, followed by 10 cycles at 94°C for 30 s, 60°C for 30 s (annealing temperature decreasing 1 °C at each cycle) and 72°C for 40 s, then 25 cycles at 94°C for 30 s, 50°C for 30 s and 72°C for 40 s with a final 7 min extension step at 72°C. The product of the single PCR was diluted 1:2 and pulled together with products of the 3-plex also diluted 1:2 to be run at the sequencer.
All PCR were performed on a GeneAmp PCR System 9700 thermal cycler (Perkin Elmer). PCR products were run on AB 3500 (Applied Biosystems, USA), with LIZ-500 as internal size standard. The resulting profiles were sized using GeneMarker (Softgenetics), jointly with some of the samples used in Chapter 4 to check consistency in peak size and band pattern.
Two EST-SSRs (Aat07 and Aat12) and one gSSR (Sf324) were not considered for further analyses because they gave unclear amplification profiles (i.e. inconsistent repetition motif, number of PCR products higher than 2, frequent allele drop-out). Individuals having more than three non-amplifying loci were excluded from the dataset.

Statistical analyses

Genetic diversity parameters

Standard genetic diversity indexes were estimated in the 24 silver fir populations by GenAlEx v6.5 (Peakall and Smouse 2012). For each population, the total number of alleles (*A*), the number of private alleles (*P*), observed (H_O) and expected heterozygosity (H_E) and the fixation index (F_{IS}) were calculated at each locus. Global F_{ST} according to Weir and Cockerham (1984) and Jost's *D* (Jost 2008) were also calculated using GenAlEx. Allelic richness (*Ar*) was calculated using Fstat v2.9.3.2 (Goudet 2001), based on a minimum sample size of 42 diploid individuals, using a rarefaction method to account for uneven sample size (El Mousadik and Petit 1996). Computation of private allelic richness (*PAr*) was carried out using the program HP Rare (Kalinowski 2005). Potential deviations from Hardy Weinberg

equilibrium (HWE) were tested by GENEPOP v4.2.1 (Rousset 2008) at each locus within each population using a modified Fisher's exact text (Guo and Thompson 1992) with 5 x 10^6 Markov chain iterations. Also linkage disequilibrium between loci across populations was tested using log likelihood ratio statistic based on 5 x 10^6 Markov chain iterations in GENEPOP.

The presence of null alleles was investigated using different approaches implemented in 3 softwares: 1) the program INEst (Chybicki and Burczyk 2009) was run using the individual inbreeding model (IIM) with a Gibbs sampler of 10^5 iterations; 2) FreeNA (Chapuis and Estoup 2007) uses the Expectation Maximization (EM) algorithm of Dempster *et al.* (1977) to estimate the frequency of null alleles. This software was also used to calculate global and pairwise F_{ST} refined by excluding null alleles using the ENA correction (Chapuis and Estoup 2007); and 3) MICRO-CHECKER v2.2.3 (van Oosterhout *et al.* 2004) was run with 1000 randomizations using the "Oosterhout" method. The presence of null alleles was predicted when there was locus-specific significant homozygosity excess (compared to HWE) that was evenly distributed across all allele sizes.

I tested for recent changes in the effective population size by using BOTTLENECK v.1.2.02 (Piry *et al.* 1999). Expected heterozygosity from Hardy-Weinberg equilibrium was compared to the heterozygosity at mutation-drift equilibrium predicted on the basis of the observed number of alleles through a Wilcoxon sign test (Piry *et al.* 1999). The program was run under a two-phase model of mutation (TPM) that generally fits microsatellite evolution better than either pure stepwise or infinite allele models (Di Rienzo *et al.* 1994). One thousand simulations were performed for each sample based on a TPM consisting of 80% single-step mutations and 20% multistep changes. Positive values of BOTTLENECK statistics reflect a gene diversity excess possibly caused by recent founder events, whereas negative values are consistent with heterozygote advantage. I also checked for population size reductions lasting several generations and, therefore, for "ancient" bottlenecks, estimating the M ratio (Garza and Williamson 2001). The M ratio was calculated in R as the ratio between the number of alleles (k) and the range of allele size divided by the repeat motif (r) for each locus in each populations. Since the M ratio is particularly sensitive to alleles at low frequency, the estimation was also carried out excluding alleles having frequency lower than 0.02 ($M_{>0.02}$).

Finally, all 17 loci were screened for neutrality in all 24 populations using BayeScan v1.0 (Foll and Gaggiotti 2008), an F_{ST} based outlier detection program. Results are not presented here because they seem to be highly affected by the biogeographical structure. The number of outlier loci markedly decreased when analyzing only populations geographically close to each other, while when using the bulk of all populations the number of outliers is high. Therefore,

further analysis that take into account the effect of population structure are required.

Genetic structure and population differentiation

Several complementary analyses were used to investigate population differentiation and genetic structure.

Pairwise-F_{ST} among all pairs of populations was calculated in GenAlex using AMOVA with 99 permutations to calculate the significance of *P*-values. Pairwise-R_{ST} were calculated in GENEPOP.

Principal Component Analysis (PCA) on square root arcsin-tranformed allelic frequencies was run using the `prcomp` function in R. The overall pattern was further visualized using an UPGMA (Unweighted Pair Group Method with Arithmetic mean) dendrogram based on Nei's distance (1978). The UPGMA dendrogram was built by using TreeFit v1.2 (Kalinowski 2009). Bootstrap support values for each branch length were calculated using 1000 bootstrapped trees.

A Bayesian clustering approach was also used to detect population genetic structure using the software STRUCTURE version 2.3 (Pritchard *et al.* 2000; Falush *et al.* 2003).I used the admixture model, where the fraction of ancestry from each cluster is estimated for each individual and allowed for correlated allele frequencies. I used no prior information on sample geographical origin. Ten independent runs for each K value ranging from 1 to 10 were performed after a burn-in period of 10^5 steps followed by 5×10^5 Markov Chain Monte Carlo replicates. To identify the number of cluster (K) that best explained the data, the rate of change of L(K) (ΔK) between successive *K* values was calculated following Evanno *et al.* (2005) using the web application StructureHarvester (Earl and von Holdt 2012).

Isolation by distance was tested comparing genetic and geographic distances between each pair of populations. F_{ST} /(1 − F_{ST}) was used as genetic distances. Several analyses were performed: including all populations, separating the populations according to the clusters obtained the Bayesian analysis in Structure. The P-values of the Mantel test correlation coefficients were obtained from 10 000 permutations using the `mantel` function in the R package `vegan`.

Finally, the analysis of molecular variance (AMOVA, Excoffier et al 1992) was used to test different hypotheses about the origin of current biogeographical structure based on clustering results and previously published studies (e.g. Linares *et al.* 2011, Cheddadi *et al.* 2013, see Tab. 5.5). AMOVA was used to rank different possible clustering options from the one that showed the highest differentiation among groups and the lowest differentiation among populations within groups. Molecular variance was partitioned:

- among and within the 24 populations (Hp 0 = no biogeographical structure);
- among geographical groups, among populations within geographical group and within population (Hp 1= biogeographical structure mainly coincident with geography);
- among biogeographical groups (as inferred by clustering analyses), among populations within biogeographical group and within population. This analysis was run at different "hierarchical" scales (Hp 3 and Hp 5);
- among groups according to what was found in previously published hypotheses on the biogeographical structure of silver fir (Hp 2 and Hp 4).

All analyses were done using GenAlex, with 999 permutations.

5.3 Results

Genetic diversity parameters

Genotypes were obtained for 1167 individuals for 17 loci (14 EST-SSRs and 3 gSSRs). Twenty-one individuals were excluded due to amplification issues.

Standard diversity parameters are presented in Table 5.2 and 5.3. Similarly to what was previously found in Chapter 4, values of genetic diversity were generally larger for gSSRs compared to EST-SSRs (Tab. 5.2). Mean allelic richness per locus was 4.44 for EST-SSRs and 11.94 for gSSRs. Also H_O and H_E were higher in gSSRs (mean = 0.70 ± 0.02 and 0.070 ± 0.02, respectively) compared to EST-SSRs (0.37 ± 0.03 and 0.37 ± 0.02).

No clear differences were found between gSSRs and EST-SSRs in terms of inbreeding coefficient, null allele frequencies, and results about the genetic structure of investigated populations. Deviations from Hardy Weinberg equilibrium were found only in 8 (out of 408) locus-population pairs but were not consistent across loci within populations. Values of F_{IS} per locus were low, except for loci Aat02 and Aat13 where it reached 0.113 (± 0.051) and 0.114 (± 0.043), respectively. Estimates of null allele frequencies were low for all loci with all 3 approaches used. The maximum frequency for null allele was 0.047 for locus Aat02 as estimated by FreeNA. Results from INEst and Microchecker confirmed this pattern (Tab. 5.2).

Linkage disequilibrium tests showed that all loci, except for the Aat01-Sf1 pair, were independent in every population. Loci Aat01 and Sf1 showed highly significant linkage disequilibrium across all populations. In the analysis presented here they are both included because their presence did not bias the overall biogeographical pattern (almost identical results were obtained by comparing results from PCA performed on the whole dataset and excluding locus Sf1, correlation>0.99). I will consider to exclude them in future analyses where independence among loci is assumed and it can determine large bias in the results, e.g. paternity and parentage experiments. A large difference among F_{ST} values was detected among loci, with values ranging from 0.009 for locus Aat09 to 0.185 for locus Aat08. Values of D were generally comparable to F_{ST}, except in the case of the more polymorphic loci, where D was always larger (Tab. 5.2).

Table 5.2: Genetic diversity parameters of the set of 14 EST-SSRs and 3 gSSRs used. *A* (mean number of alleles), *Ar* (allelic richness), H_O (observed heterozygosity), H_E (expected heterozygosity), Null (INest and Microchecker: significant presence of null allele; $p < 0.01$), Null (FreeNA: estimated frequency of null allele). *D* (Jost's D), F_{ST} (global differentiation index), F_{ST} ENA (global differentiation index corrected for the presence of null alleles). Values in parentheses are standard errors.

Locus	*A*	Ar_{42}	H_O	H_E	F_{IS}	Null (INest)	Null (Microchecker)	Null (FreeNA)	*D*	F_{ST}	F_{ST} *ENA*
Aat01	3.96 (0.21)	4.67	0.59 (0.02)	0.56 (0.01)	-0.057 (0.023)	0/24	0/24	0.006	0.088	0.063	0.063
Aat02	2.38 (0.15)	2.64	0.34 (0.03)	0.38 (0.02)	0.113 (0.051)	2/24	3/24	0.047	0.032	0.048	0.052
Aat09	1.75 (0.11)	1.96	0.03 (0.01)	0.03 (0.01)	0.048 (0.041)	0/24	0/24	0.007	0.000	0.009	0.025
Aat03	3.75 (0.24)	4.98	0.47 (0.04)	0.46 (0.04)	-0.016 (0.025)	0/24	1/24	0.011	0.104	0.106	0.106
Aat04	4.54 (0.24)	6.15	0.41 (0.03)	0.39 (0.03)	-0.042 (0.031)	0/24	1/24	0.011	0.048	0.068	0.066
Aat05	2.17 (0.10)	2.97	0.18 (0.03)	0.17 (0.02)	0.005 (0.035)	0/24	0/24	0.014	0.020	0.085	0.082
Aat06	3.04 (0.23)	4.80	0.19 (0.03)	0.19 (0.03)	0.051 (0.041)	0/24	2/24	0.018	0.028	0.102	0.098
Aat10	6.33 (0.24)	7.73	0.66 (0.02)	0.66 (0.02)	0.001 (0.013)	0/24	0/24	0.011	0.178	0.079	0.079
Aat11	3.54 (0.13)	3.94	0.51 (0.01)	0.53 (0.01)	0.044 (0.022)	0/24	3/24	0.024	0.096	0.075	0.073
Aat08	2.71 (0.19)	3.98	0.38 (0.05)	0.39 (0.05)	0.001 (0.026)	0/24	2/24	0.018	0.146	0.185	0.190
Aat13	2.96 (0.2)	3.87	0.22 (0.03)	0.26 (0.03)	0.114 (0.043)	1/24	2/24	0.037	0.048	0.119	0.120
Aat14	6.96 (0.35)	9.78	0.70 (0.02)	0.69 (0.02)	-0.006 (0.020)	0/24	1/24	0.013	0.210	0.081	0.079
Aat15	2.25 (0.11)	2.64	0.22 (0.03)	0.21 (0.03)	-0.007 (0.029)	0/24	1/24	0.011	0.055	0.159	0.157
Aat16	2.00 (0.00)	2.00	0.24 (0.02)	0.24 (0.02)	0.015 (0.028)	0/24	0/24	0.019	0.008	0.025	0.026
Average EST-SSRs	3.45 (0.18)	4.44	0.37 (0.03)	0.37 (0.02)	0.019 (0.03)			0.018	0.076	0.086	0.087
NFF7	16.33 (0.95)	20.8	0.81 (0.02)	0.84 (0.02)	0.037 (0.016)	0/24	4/24	0.023	0.439	0.07	0.07°
Aag01	8.21 (0.31)	10.2	0.70 (0.03)	0.71 (0.03)	0.014 (0.017)	0/24	2/24	0.014	0.347	0.118	0.117
SF1	3.88 (0.22)	4.77	0.59 (0.02)	0.56 (0.01)	-0.062 (0.021)	0/24	0/24	0.007	0.089	0.064	0.064
Average gSSRs	9.47 (0.50)	11.9	0.70 (0.02)	0.70 (0.02)	-0.004 (0.018)			0.015	0.292	0.084	0.084

Table 5.3: Genetic diversity parameters of the 24 populations studied. *A* (mean number of alleles), *Ar* (allelic richness), *P* (number of private alleles), *PAr* (number of private alleles after rarefaction), H_O (observed heterozygosity), H_E (expected heterozygosity), F_{IS} (fixation index), F_{IS} *null* (fixation index accounting for the presence of null alleles).

	Population	*A*	Ar_{42}	*P*	*PAr*	H_O	H_E	F_{IS}	F_{IS} *null*
Northern Apennine	ABE	4.41 (0.77)	4.300	0	0.000	0.43 (0.07)	0.43 (0.06)	0.002 (0.031)	0.009 (0.010)
	BTR	3.47 (0.54)	3.421	0	0.000	0.40 (0.06)	0.40 (0.06)	-0.020 (0.021)	0.011 (0.013)
	CER	4.29 (0.64)	4.206	1	0.057	0.46 (0.05)	0.45 (0.05)	-0.027 (0.018)	0.009 (0.010)
	NER	4.76 (0.85)	4.678	2	0.110	0.44 (0.05)	0.47 (0.06)	0.056 (0.029)	0.010 (0.011)
	PIG	4.41 (0.76)	4.307	0	0.000	0.44 (0.06)	0.46 (0.06)	0.068 (0.034)	0.009 (0.010)
Central Apennine	ABS	4.12 (0.79)	4.031	0	0.001	0.42 (0.06)	0.42 (0.06)	0.014 (0.018)	0.010 (0.011)
	COR	3.88 (0.62)	3.810	1	0.049	0.48 (0.07)	0.44 (0.06)	-0.054 (0.045)	0.006 (0.007)
	CPL	4.12 (0.76)	4.029	0	0.008	0.39 (0.06)	0.41 (0.06)	0.060 (0.027)	0.009 (0.01)
	TOS	4.35 (0.78)	4.246	0	0.000	0.41 (0.07)	0.41 (0.07)	0.003 (0.028)	0.009 (0.010)
	VCL	4.00 (0.70)	3.915	0	0.000	0.39 (0.07)	0.38 (0.07)	-0.017 (0.025)	0.008 (0.009)
Southern Apennine	CIL	5.00 (1.14)	4.890	0	0.010	0.45 (0.07)	0.45 (0.07)	0.020 (0.040)	0.012 (0.012)
	GAM	5.12 (1.19)	5.002	2	0.134	0.45 (0.07)	0.45 (0.06)	0.040 (0.040)	0.010 (0.010)
	LAU	4.94 (1.13)	4.804	0	0.019	0.47 (0.07)	0.49 (0.07)	0.040 (0.030)	0.010 (0.011)
	SIL	5.29 (1.16)	5.105	2	0.115	0.49 (0.06)	0.47 (0.06)	-0.064 (0.027)	0.016 (0.015)
	SSB	5.18 (1.14)	5.027	2	0.110	0.44 (0.06)	0.45 (0.07)	0.020 (0.033)	0.011 (0.012)
	TNP	5.71 (1.48)	5.501	2	0.121	0.44 (0.06)	0.46 (0.06)	0.032 (0.024)	0.011 (0.011)
Eastern Alps	NOA	4.18 (0.79)	4.129	0	0.000	0.40 (0.06)	0.44 (0.06)	0.092 (0.055)	0.012 (0.012)
	TAR	4.35 (0.69)	4.228	1	0.060	0.44 (0.06)	0.42 (0.06)	-0.037 (0.047)	0.006 (0.007)
Western Alps	PES	4.24 (0.68)	4.220	0	0.008	0.40 (0.06)	0.42 (0.06)	0.055 (0.055)	0.010 (0.011)
	SAL	3.71 (0.65)	3.635	0	0.000	0.38 (0.05)	0.41 (0.05)	0.089 (0.062)	0.011 (0.012)
	TOC	3.88 (0.72)	3.842	0	0.000	0.44 (0.06)	0.44 (0.05)	0.021 (0.054)	0.009 (0.009)
Balkans	BLG	5.12 (1.12)	5.018	4	0.225	0.41 (0.07)	0.41 (0.08)	-0.016 (0.016)	0.009 (0.009)
	ROM	4.76 (1.06)	4.623	1	0.070	0.36 (0.07)	0.36 (0.07)	0.004 (0.037)	0.008 (0.009)
	SER	5.06 (1.27)	4.921	3	0.190	0.38 (0.07)	0.37 (0.08)	-0.026 (0.014)	0.008 (0.009)

When comparing within-population genetic diversity parameters across populations some general trends emerged. The highest diversity values were found in Southern Apennine, Balkans and Northern Apennine, whereas Western Alps and Central Apennines exhibited the lowest diversity values (Tab. 5.3, Tab. 1 in Appendix 2). This was confirmed also by the higher number of private alleles. Interestingly, the general pattern was inverted for few loci where H_E was higher in Central Apennine populations (e.g. locus Aat08 and Aat15, data not shown). In all populations inbreeding coefficients were low.

Table 5.4: Summary of population size reduction tests. M-statistic (Garza and Williamson, 2001) averaged among loci, values < 0.70 are highlighted in bold. H_E excess: fractions of loci showing heterozygosity deficiency (Cornuet and Luikart, 1996), *: one ore more monomorphic loci. T2 statistic of Cornuet and Luikart (1996) Wilcoxon test p-values for heterozygosity tests: Bottleneck test and Expansion test using TPM model.

Population	M	$M_{>0.02}$	H_E excess	T2		TPM (20%) Bottleneck test	TPM (20%) Expansion test
ABE	0.761 (0.228)	0.780 (0.239)	7/17	-1.581	°	ns	°
BTR	0.803 (0.234)	0.784 (0.258)	9/16*	-0.147	ns	ns	ns
CER	0.652 (0.207)	0.685 (0.259)	8/17	-1.315	°	ns	ns
NER	0.703 (0.239)	0.699 (0.239)	7/17	-1.633	°	ns	ns
PIG	0.751 (0.238)	0.748 (0.235)	9/16*	-0.93	ns	ns	ns
ABS	0.802 (0.220)	0.777 (0.241)	7/15**	-1.233	ns	ns	ns
COR	0.719 (0.275)	0.752 (0.269)	10/17	0.233	ns	ns	ns
CPL	0.713 (0.255)	0.774 (0.275)	8/17	-1.18	ns	ns	ns
TOS	0.763 (0.227)	0.742 (0.250)	7/17	-1.321	°	ns	ns
VCL	0.790 (0.225)	0.781 (0.260)	8/16*	-1.659	*	ns	ns
CIL	0.836 (0.218)	0.812 (0.253)	7/16*	-1.016	ns	ns	ns
GAM	0.779 (0.200)	0.795 (0.231)	7/17	-1.066	ns	ns	ns
LAU	0.790 (0.213)	0.792 (0.251)	9/15**	0.584	ns	ns	ns
SIL	0.795 (0.214)	0.814 (0.214)	10/17	-1.122	ns	ns	ns
SSB	0.800 (0.193)	0.823 (0.200)	6/17	-1.381	°	ns	°
TNP	0.784 (0.201)	0.775 (0.221)	5/17	-1.808	*	ns	*
NOA	0.737 (0.224)	0.730 (0.223)	7/16*	-0.504	ns	ns	ns
TAR	0.722 (0.196)	0.724 (0.268)	7/17	-2.554	**	ns	ns
PES	0.751 (0.208)	0.778 (0.230)	5/16*	-1.980	*	ns	*
SAL	0.751 (0.238)	0.778 (0.260)	7/16*	-1.185	ns	ns	ns
TOC	0.758 (0.227)	0.725 (0.256)	9/16*	-0.211	ns	ns	ns
BLG	0.828 (0.217)	0.900 (0.166)	6/17	-1.601	°	ns	°
ROM	0.736 (0.195)	0.855 (0.222)	5/16*	-3.503	***	ns	*
SER	0.785 (0.224)	0.880 (0.160)	5/16*	-1.802	*	ns	°

Only 2 populations (i.e. CER and NER in Northern Apennine) were characterized by M < 0.70, such low M values suggest an "ancient" reduction in the effective population size. The populations COR and CPL (Central Apennine), and NOA and TAR (Eastern Alps) showed M values close to 0.70. Recent population reductions were not by the Bottleneck test, but, a significant heterozygosity excess ($P < 0.05$, Wilcoxon test) was detected in populations TNP, PES, ROM (and, with a weaker signal, in ABE, SSB, BLG and SER) indicating a possible recent population expansion in these stands.

Biogeographical structure and population differentiation

By using the Bayesian clustering approach implemented in Structure, the highest ΔK occurred at $k = 2$ (Fig. 5.2, Tab. 2 in Appendix 2) with probability of membership exceeding 0.90 for 1018 individuals out of 1167 (87%). The two inferred clusters are concordant with large scale geography. One of the inferred clusters included populations from the Balkans and Southern Apennine, whereas the second comprised the Northern and Central Apennine populations and the Alpine ones (Fig. 5.3a, 5.3c). The only exception was population ABS, the southernmost population of Central Apennine, that clustered with Southern Apennine and Balkans (Fig. 5.3a). This major separation in 2 genetic clusters was also found in PCA along the 1st principal component (explaining 36% of the overall variance, Fig. 5.4a and Fig. 1 in Appendix 2) and in the UPGMA tree with strong bootstrap support (Fig.5.5, Tab. 3 in Appendix 2).

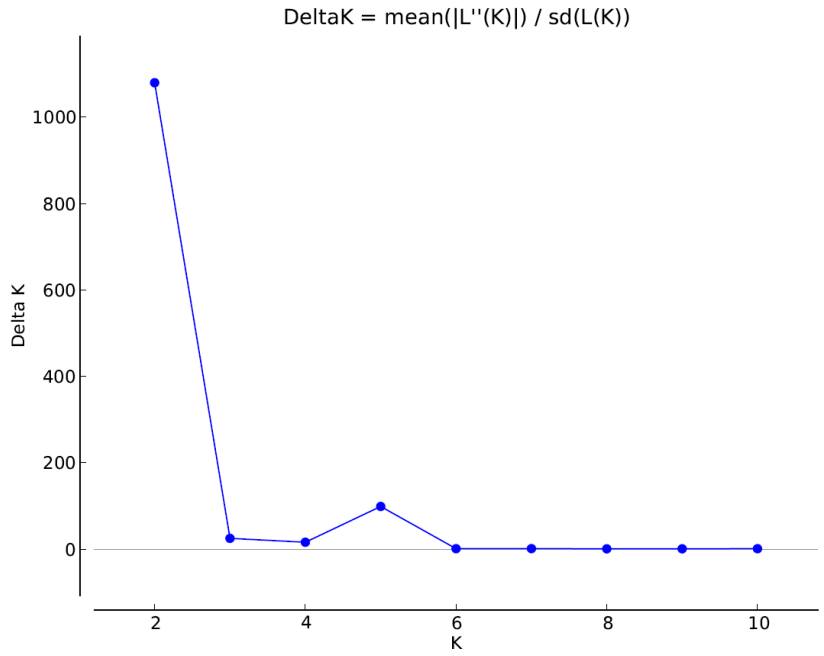

Figure 5.2: Values of ΔK from Structure analysis.

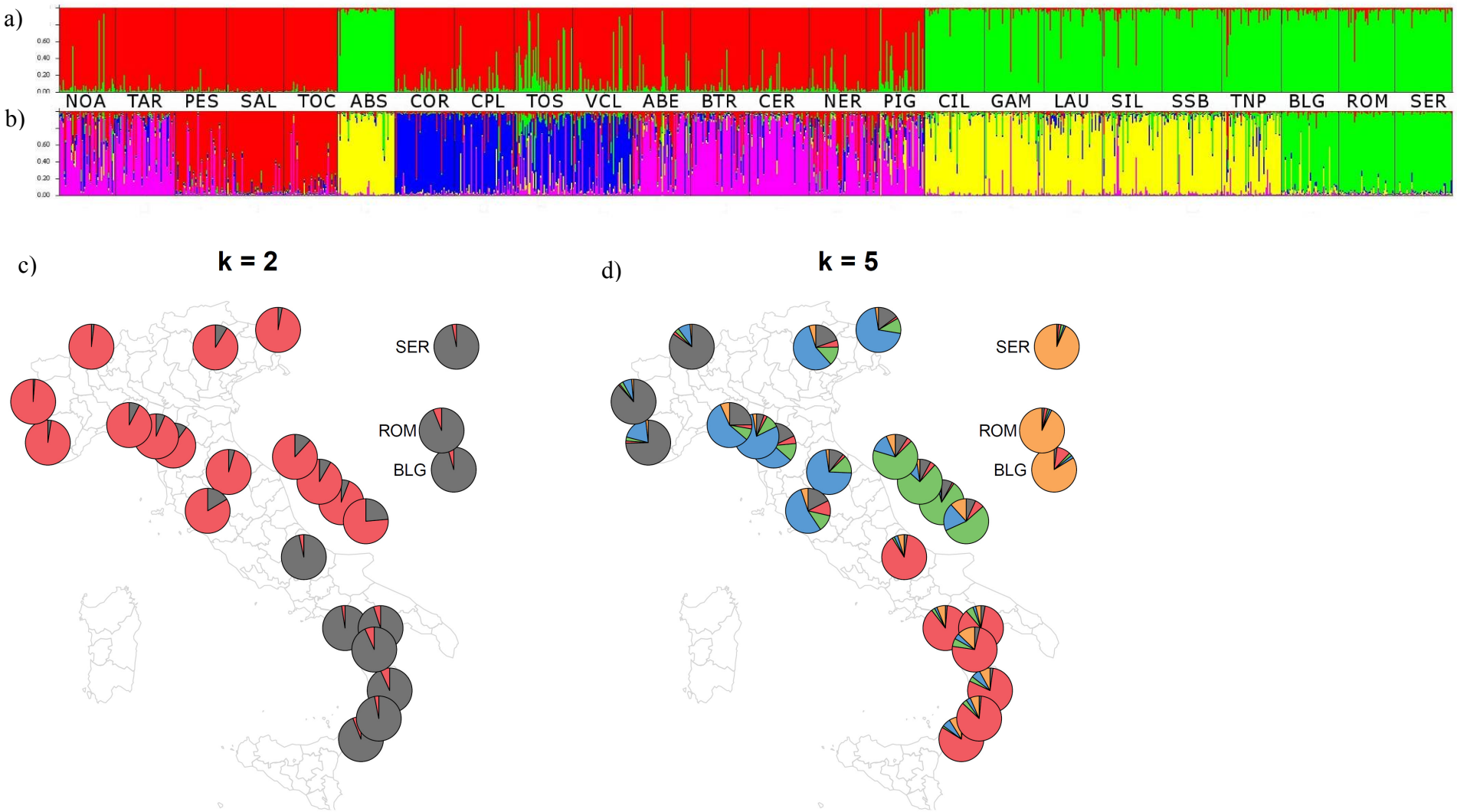

Figure 5.3: Above, bar graphs of individual and population membership to each of the k genetic clusters: a) $k = 2$, b) $k = 5$. Below, Map of mean population membership to each genetic cluster: c) $k = 2$, d) $k = 5$.

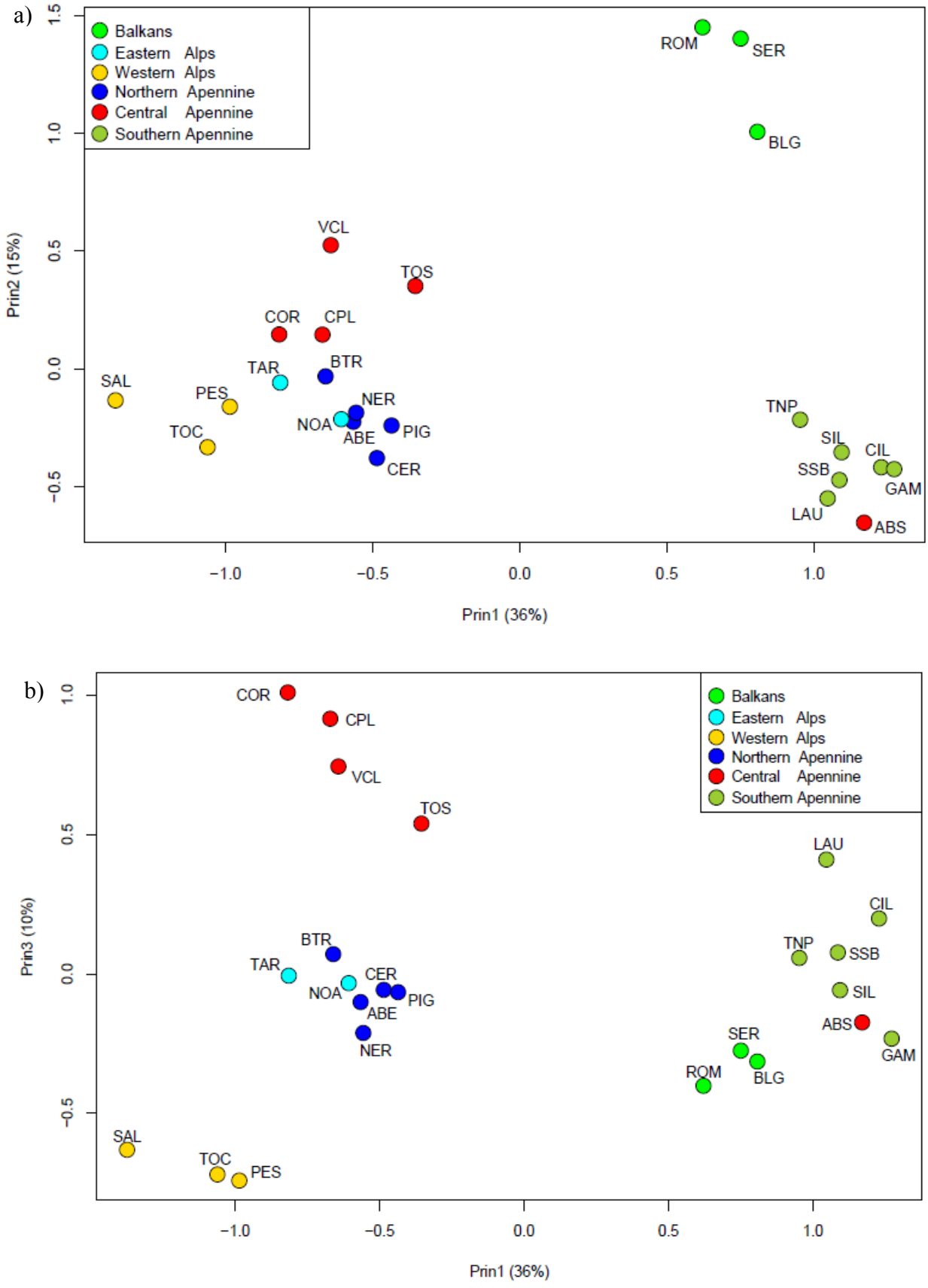

Figure 5.4: Principal Component Analysis ordination diagrams a) Principal Component 1 x Principal Component 2, b) Principal Component 1 x Principal Component 3.

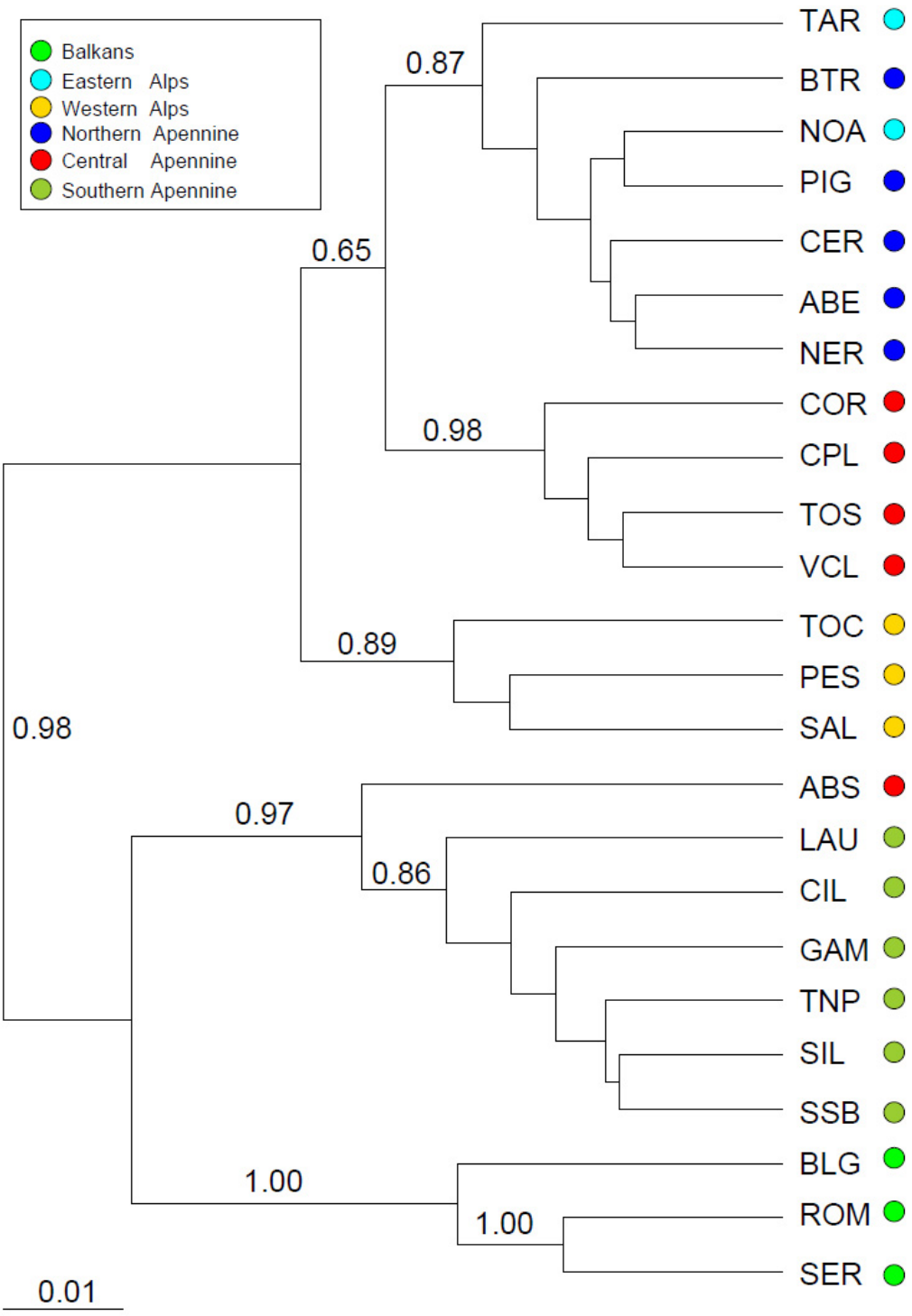

Figure 5.5: UPGMA tree based on Nei's genetic distance among populations and bootstrap support values for the branches.

All clustering approaches highlighted the existence of smaller clusters at a lower hierarchical level, generally reflecting the geography at a finer scale. In Structure, a second smaller but evident ΔK peak was found for k = 5 (Fig. 5.2, Tab. 2 in Appendix 2) with probability of membership higher than 0.90 for 596 individuals (51%). The two core clusters are further subdivided into 5 geographically concordant groups: 1) Balkans, 2) Southern Apennine (again together with ABS from Central Apennine), 3) Central Apennine, 4) Northern Apennine together with Eastern Alps, and 5) Western Alps (Fig. 5.3b and 5.3d). The same subtle biogeographical structure was clearly detected also by UPGMA (Fig. 5.5, Tab. 3 in Appendix 2) and by PCA (Fig. 5.4 a,b, Fig. 1 in Appendix 2). The second principal component (explaining 15% of variance) divided the Balkan populations from the Southern Apennine ones, whereas the third one (10% of explained variance) separated Central Apennine populations, Northern Apennine-Eastern Alps populations, and Western Alps populations in 3 distinct groups.

The support for the finer biogeographical structure, despite being lower than the one for the major division in 2 cluster, was sound (see q values, group separation in the PCA ordination, bootstrap support for the branches). The biogeographical structure emerged from the clustering approaches were confirmed by pair-wise F_{ST} and R_{ST} values (Tab. 4 in Appendix 2). The degree of pair-wise population differentiation generally increased with distance among populations (Mantel test considering all 24 populations: r = 0.60, P < 0.001, see Fig. 5.6 and Fig. 2 in Appendix 2). I found some cases where populations geographically relatively close to each other were highly genetically differentiated (e.g. pairwise F_{ST} between ABS and COR was 0.176, pairwise F_{ST} between NOA and TOC was 0.105, pairwise F_{ST} between PES and CER was 0.085) or, on the opposite, where populations relatively far from each other showed low differentiation (e.g. pairwise F_{ST} values between NOA and TAR from PIG were 0.012 and 0.028, respectively). The population ABS was clearly differentiated from all Northern and Central Apennine populations (pairwise F_{ST} values always higher than 0.095), but genetically close to the Southern Apennine ones (pairwise F_{ST} values ranging from 0.028 and 0.044).

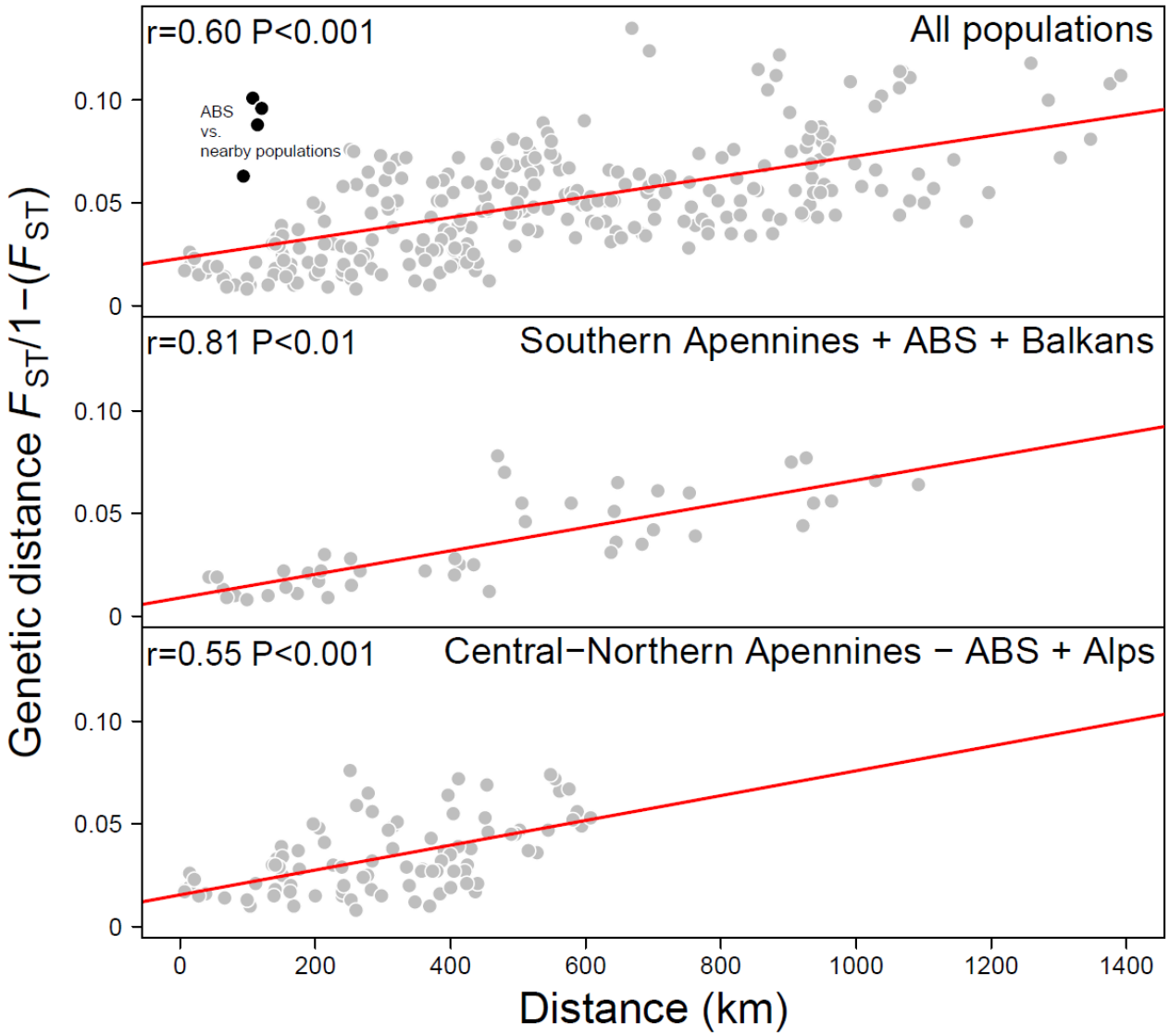

Figure 5.6: Relationship between genetic and geographical distance, tested using Mantel test.

Results from AMOVA analyses showed that most of the total genetic variation is attributable to within-population differentiation (~90%) (Tab. 5.5). Remaining genetic variation (~10%) was differently partitioned among and within groups depending on the grouping hypotheses (Tab. 5.5). The highest percentage of variation among groups (7.43%) was obtained when populations are grouped within the 5 clusters identified by clustering analyses (Hp 5). In this case, Φ_{RT} is almost double then Φ_{SR} (0.074 *vs*. 0.029).

Table 5.5 Hypotheses of population differentiation and biogeographical structure based on clustering results and on previously published studies. Hypotheses were tested using AMOVA partition of molecular variance among groups, among populations within group and within populations.

k = number of groups.

				% Variation			Φ-statistic		
Hypothesis	Source	k	Group composition	Among groups	Among populations	Within populations	Φ_{RT}	Φ_{SR}	Φ_{ST}
(0) No biogeographical structure		0	All 24 populations separated	-	8.77	91.23	-	-	0.088**
(1) Geographical groups	Geography	6	(1) Balkans (2) Southern Apennine (3) Central Apennine (4) Northern Apennine (5) Western Alps (6) Eastern Alps	5.94	3.67	90.38	0.059**	0.039**	0.096**
(2) 2 groups from 2 different glacial refugia	Linares *et al.* 2011	2	(1) Balkans (2) all Italian populations	4.46	7.74	88.07	0.045**	0.078**	0.119**
(3) 2 core genetic clusters	Structure, UPGMA, PCA	2	(1) Balkans + Southern Apennine (+ABS) (2) Central (-ABS) + Northern Apennines + all Alps	6.16	5.38	88.46	0.062**	0.057**	0.115**

(4) 3 groups from 3 different glacial refugia	Cheddadi *et al.* 2013,PCA	3	(1) Balkans (2) Southern Apennine (+ABS) (3) Central (-ABS) + Northern Apennines + all Alps	7.18	4.31	88.50	0.072**	0.046**	0.115**
(5) 5 genetic clusters	Structure, UPGMA, PCA (deeper hierarchical structure)	5	(1) Balkans (2) Southern Apennine (+ABS) (3) Central Apennines (-ABS) (4) Northern Apennines + Eastern Alps (5) Western Alps	7.43	2.69	89.88	0.074**	0.029**	0.101**

5.4 Discussion

In this chapter, I investigated the biogeographical pattern of genetic variation in silver fir populations from the Apennines using a newly developed set of 17 microsatellites (see Chapter 4). To my knowledge, this is the first study where an ad-hoc sampling scheme was applied to unravel past biogeographical dynamics of silver fir in the Apennines. To this aim, a high number of carefully selected populations were sampled across the whole Apennine range, together with populations from the Alps and the Balkans as outgroups. Previous published studies on silver fir biogeography were generally based on surveys carried out at a broader scale with low resolution, with populations from Southern Europe being underrepresented (to this regard, indicative are Fig.1 in Liepelt *et al.* 2002 and Liepelt *et al.* 2010 for genetic data, and Fig.1a in Cheddadi *et al.* 2013 for pollen fossil data). Some information were also available from small experiments on a geographically limited area (e.g. Vicario *et al.* 1995, Piovani *et al.* 2010). However, the heterogeneity in spatial scales investigated and the variety in the molecular markers used have made it difficult to tell a reliable and compelling story about *Abies alba* past dynamics in the Apennines so far.

All clustering approaches used to investigate the genetic structure and the amount of differentiation among sampled populations (i.e. Bayesian analysis in Structure, PCA and UPGMA tree based on Nei's genetic distance) led to the same clear biogeographical pattern highlighting the presence of genetically distinct clusters highly concordant with geography, even though spatial data were not included in the analyses. Differentiation occurred at different hierarchical levels (at least 2), with major and minor separations that indentify geographically coherent clusters at a broader and finer scale, respectively (the only exception in the correspondence between geographic and genetic clusters is the population ABS. This is the southernmost population of Central Apennines but it clustered with Southern Apennine populations and was actually highly differentiated from the Central Apennine ones, therefore hereafter it will be not be considered as being part of the Central Southern Apennine cluster). Surprisingly, a major separation divided populations from Southern Apennines and Balkans from all the other populations (from Central and Northern Apennine and from the Alps) (Fig. 5.7). Within these two major clusters deeper subdivisions separated: *i*) the Southern Apennines from the Balkans, and *ii*) populations from Northern Apennines and Eastern Alps from Western Alps and from Central Apennines (with Western Alps and Central Apennines also clearly separated) (Fig. 5.7). This finer division in 5 groups accounted for a high percentage of molecular variation among groups as compared to what was found when analyzing using only major separations. Interestingly, I found that the separations among the

inferred genetic clusters were generally characterized by abrupt genetic discontinuities, even between geographically close populations (the strongest separation between ABS and TOS). To investigate the causes of such steep genetic discontinuities deserves further attention.

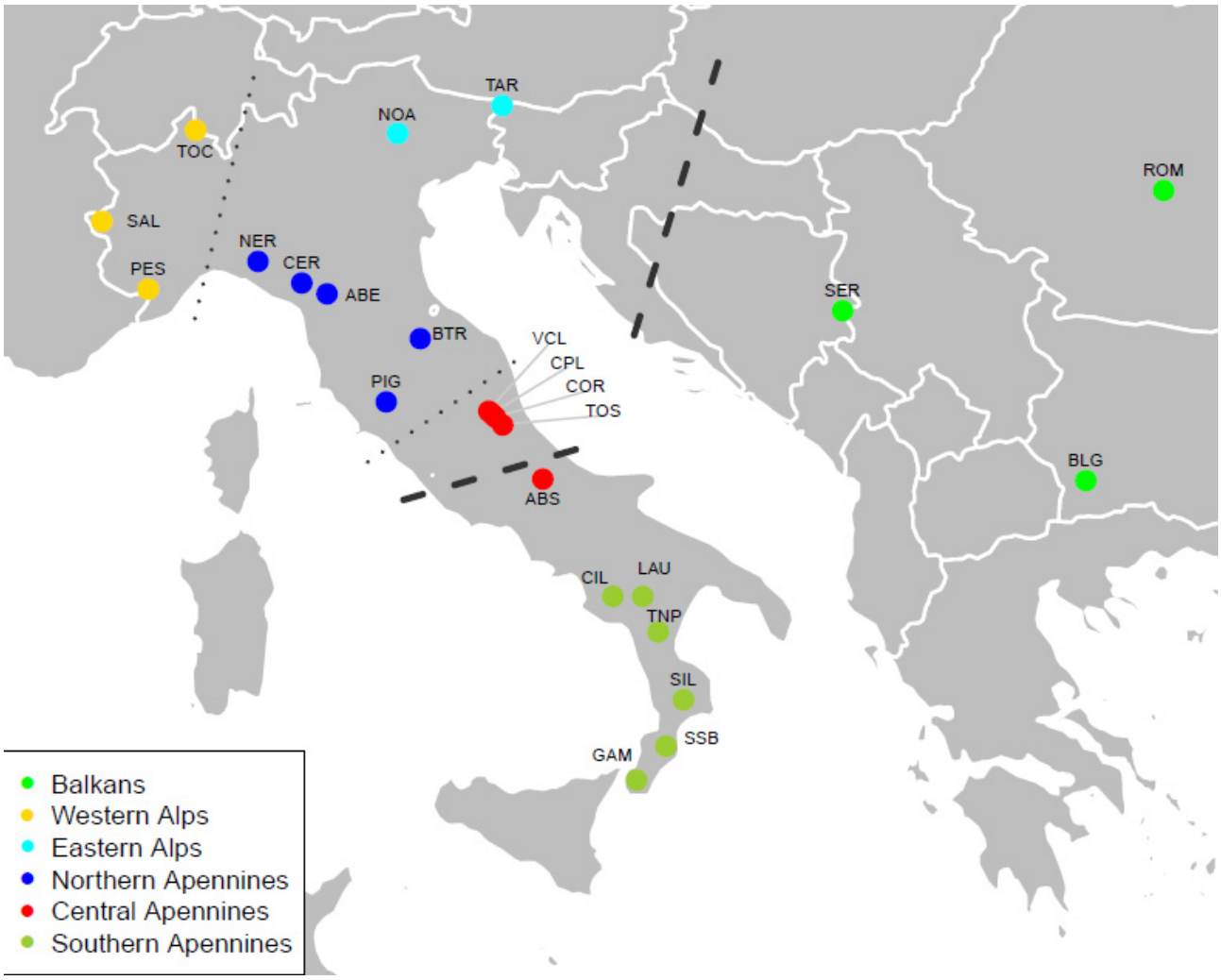

Figure 5.7: Map of the sampled populations and separations among genetic clusters. Major separations are indicated in bold, minor separations with the dotted line.

Post-glacial migration routes and dynamics

Given that in many forest trees current biogeographical patterns have been moulded by glacial range retreat and post-glacial recolonization (Taberlet and Cheddadi 2002, Petit *et al.* 2003), my results on the biogeographical structure of the investigated populations allows drawing some conclusions on their past dynamics. The high genetic diversity, number of private alleles and genetic distinctiveness that characterize all Southern Apennine populations are concordant with the hypothesis of an isolated refugia in Southern Italy (Larsen 1986, Konnert and Bergmann 1995, Vendramin *et al.* 1999, Cheddadi *et al.* 2013), but do not support the hypothesis of a post-glacial migration route starting from Southern Italy and recolonizing the

entire Italian peninsula (as postulated by Parducci *et al.* 1996 and Linares 2011). Previous genetic studies have mainly focused on Calabrian populations, highlighting their high genetic diversity and peculiarity with respect to all the other Italian and Apennine populations (e.g. Parducci *et al.* 2001). In this study I found that also the other Southern Apennine populations (CIL, TNP, LAU), despite being characterized by very different geological and environmental conditions, have levels of genetic diversity and allelic richness equivalent to the ones of Calabrian populations (i.e. GAM, SIL, SSB). Therefore, according to my data, the entire Southern Apennine group appears as a hotspot of genetic diversity with local dynamics and high within-group gene flow, but isolated from the rest of Apennines.

High genetic diversity, allelic richness and number of private alleles were also found in Northern Apennine, especially in populations NER, PIG and ABE. This is surprising because these populations are currently highly fragmented and have small population size (Piovani *et al.* 2010), therefore a strong effect of genetic drift is expected. These populations resulted genetically very close to the ones from Eastern Alps (pairwise F_{ST} among populations belonging to the 2 geographic groups ranged from 0.012 to 0.055), but have generally higher genetic diversity and allelic richness. Since my sampling scheme allowed to compare Northern Apennine populations with all surrounding silver fir forests at a biogeographical scale, these findings represent the first thorough genetic evidence supporting the existence of a glacial refugium in this area. This refugium may have been the starting point of a recolonization route through the Po plain towards the Alps and Central Europe (Cheddadi *et al.* 2013). This hypothesis is also supported by studies that found the presence of fossil pollen in North-Western Apennine and in the Po plain during the Late-glacial (13,000 - 10,000 BP) (Ravazzi *et al.* 2006, Kaltenrieder *et al.* 2009, Vescovi *et al.* 2010), highlighting that low altitude populations in the Po plain may have acted as a fast connection between the Apennines and the Alps in the Late-glacial (Kaltenrieder *et al.* 2009, Piovani *et al.* 2010, Cheddadi *et al.* 2013). According to our data, this migration route is not likely to have involved the Western Alps since they are clearly differentiated from both Northern Apennine and Eastern Alps (pairwise F_{ST} with populations belonging to Northern Apennine and Eastern Alps ranged from 0.038 and 0.134 and from 0.069 and 0.123, respectively).

With regard to Central Apennine populations, no clear evidence about their post-glacial origin was found. Interestingly, these populations showed lower genetic diversity when compared with the surrounding ones. In additions, despite being geographically very close (6-30 km), they were clearly separated among them in the PCA ordination, suggesting that the effect of drift may be particularly strong in these populations as a consequence of small population size, probably due to anthropic activities. The most likely hypothesis regarding their history is

that they were colonized from Northern Apennine populations with a southward migration, but their survival during the last glaciations in this area can not be excluded.

This work focused on the Apennines, but also some information about the Alps emerged. The low within-population genetic diversity - despite large population sizes - found in Western Alp populations does not support the presence of a refugium in this area (as postulated by Terhurne-Berson *et al.* 1994). A more likely hypothesis would be that these populations originated from a refugium in the Central Massif (Konnert and Bergmann 1995), in South-Eastern France (Terhurne-Berson *et al.* 1994) or from a secondary westward expansion from Eastern Alps (Cheddadi *et al.* 2013), but our data are not adequate to draw firm conclusions on this issue. With regard to Eastern Alps, our results are concordant with studies supposing their origin from Northern Apennine (Konnert and Bergmann 1995, Piovani *et al.*2010, Cheddadi *et al.* 2013), but not with the ones depicting them as a contact zone between the recolonization routes from the Apennines and from the Balkans, or as originating solely from the Balkan route (Parducci *et al.* 1996, Liepelt *et al.* 2002, Liepelt *et al.* 2009, Liepelt *et al.* 2010, Cheddadi *et al.* 2013).

Eco-genetic clusters in the Apennines

Results from this study highlighted the presence of 3 clearly distinct genetic clusters in the Apennines (Northern Apennine, Central Apennine and Southern Apennine populations) separated by 2 discontinuity areas (Fig. 5.7). The first sharp discontinuity area occurred at the boundary between Central and Southern Apennine, with an abrupt change in genetic composition (the differentiation between Central and Southern Apennine is actually higher than the one between Southern Apennine and Balkans). As previously mentioned, this is concordant with results from previous genetic studies (e.g. Vendramin *et al.* 1999, Parducci *et al.* 2001, Liepelt 2010), bus also with eco-physiological studies (e.g. Larsen and Mekic 1991, Hansen and Larsen 2004). Larsen and Mekic (1991) in a provenance test on 15 different provenances from Italy and Central Europe found that populations from Southern Italy behaved very differently from all the other provenances in all traits measured. Compared to populations from Central Europe, these populations exhibited higher rates of photosynthesis, transpiration, growth and higher water use efficiency, suggesting a better adaptation to drought. Hansen and Larsen (2004) confirmed these findings but also found earlier bud burst and poor winter-frost resistance in these populations. The division among Central and Southern Apennines was instead not supported by the study of the relationship between climate and growth patterns carried out by Carrer *et al.* (2010). By using tree-ring chronology, they found similar growth patterns in Central and Southern Apennines.

The second discontinuity zone was found between Northern and Central Apennine at a finer differentiation scale. Despite studies on this part of the Apennine range are generally very scarce, our data confirm the results found by Liepelt *et al.* (2010) and by the dendro-ecological study of Carrer *et al.* (2010), where they showed that tree-ring growth patterns were highly different among Northern Apennine and Central-Southern Apennine.

A third discontinuity area was found at the northern edge of the Apennines between Northern Apennine and Western Alps (described above).

The differentiation patterns detected are uncommon for a forest tree, generally thought to be characterized by genetic uniformity over large scale due to extensive gene flow, especially by pollen (Liepelt *et al.* 2002), and implies the presence of some barriers to gene flow. These barriers can have different origin: geographical (e.g. mountain peaks), due to anthropic activities (e.g. deforestation of wide areas), or ecological (populations growing in different environments can be locally adapted. Whatever the original cause, the differentiation currently observed among clusters seems both eco-physiological and genetic. Results from this study call for further investigation to disentangle if differentiation is mainly due to demographical and historical causes or if it is driven also by selective processes. To understand how the distinct clusters arose and what processes currently maintain them, genetic discontinuities are the most interesting areas for intensifying the sampling effort establishing small-scale transects. In future experiments following this thesis work, populations from discontinuity areas will be sampled at a finer scale to precisely detect geographically circumscribed areas where to set up gene flow experiments at the border of eco-genetic clusters. Populations to be sampled have already been identified for one of the two main discontinuity areas along the Apennines (area number 2 in Fig. 5.8), between the Northern Apennine cluster and the Central Apennine one. Nine small stand at low altitude (500-9000 m a.s.l.) have already been identified in the northern part of the transect (Maciaroni 2012) and five more in the southern part (C. Urbinati, personal communication).

Figure 5.8: Map of the occurrence of silver fir along the Apennine from Rovelli (1995). On this map I indicated (red areas) populations sampled in this study. Orange zones are areas of particular interest for the future experiments. In particular, orange areas indicated as 2 and 3 correspond to the genetic discontinuity areas.

The plan is to study effective gene flow using paternity/parentage analysis, and using both the set of markers used for this study and a set of putatively selective markers (SNPs), already available for silver fir. The aim would be to compare the rate and direction of neutral and potentially adaptive gene flow in order to identify the currently ongoing evolutionary processes and predict the future dynamics of these populations.

Chapter 6:
Conclusions

My thesis work aimed at providing a conceptual and methodological framework for the study of adaptive gene flow. In particular, I focused on the development of all elements needed for performing an experiment to estimate adaptive gene flow in populations of silver fir (*Abies alba* Mill.), a European conifer species with high ecological and economical value.

There is a critical need to predict how plants will cope with the currently ongoing climate change and whether they will be able to adapt or to migrate fast enough not to perish (Corlett and Wescott 2013). In forest trees, the relationship between adaptation and gene flow has rarely been studied with *ad-hoc* experiments. The strength of local adaptation has been traditionally studied using common garden or transplant experiments (see Savolainen *et al.* 2007 for some examples), where gene flow was mainly seen as secondary/disturbing factor.
On the other hand, gene flow *per se* (i.e. estimating gamete immigration and dispersal kernel parameters) has been widely studied within natural populations of forest trees using neutral molecular markers (i.e. allozymes, microsatellites or, less often, AFLPs). Therefore, a variety of robust and sophisticated analytical methods for the study of neutral gene flow using paternity and parentage analysis is currently available (Jones et al 2010, Klein and Oddou-Muratorio 2011, Robledo-Arnuncio 2012). These methods, together with recent advances in the development of molecular and genomic tools, may allow using potentially adaptive markers to directly track adaptive gene flow.
Given that molecular and analytical tools are available, what is still missing to study adaptive gene flow is to assemble all elements needed in a cohesive workflow. My thesis gives a major contribution to this point using as case study the scattered silver fir populations along the 1000 km long Apennine chain.
The main elements needed to perform an experiment for estimating adaptive gene flow are: having an adequate set of molecular markers, selecting a suitable combination of study species x biogeographical zone upon which designing the experiment, and choosing an optimal sampling strategy both for identifying relevant areas where it is most likely to detect adaptive gene flow, and for correctly describing gene flow patterns. In my thesis, I focused on these issues endeavoring to make a positive and significant contribution to each of them.

First of all, to deepen my knowledge of gene flow literature and to scrutinize previously neglected methodological aspects, I investigated the effect of sampling strategy on the estimation of gene flow rates. I focused on paternity analysis because is it among the most used methods to directly estimate gene flow via-pollen (Ashley 2010), which can be more pervasive and faster than gene flow via seed in sustaining adaptive gene flow over long distances (Robledo-Arnuncio 2011). Therefore, I decided to evaluate if sampling efforts commonly found in paternity published studies were adequate to estimate pollen dispersal parameters, in particular, pollen immigration rate. For doing this, the first step has been to conduct a thorough review of published paternity studies. This revealed that sampling effort was generally low (even when only pollen immigration was estimated) and that the importance of sampling strategy was widely overlooked, resulting in large and unjustified differences in sampling effort among studies.

Subsequently, I evaluated the adequacy of sampling efforts commonly adopted in paternity experiments by carrying out a simulation study. This study aimed at assessing the accuracy and precision in the reconstruction of pollen dispersal patterns (i.e. pollen dispersal kernel and immigration rate). I tested the effect of a wide range of sampling efforts and of dispersal scenarios ranging from very restricted to extensive. I found that low sampling efforts can result in highly biased and imprecise estimates of the parameters characterizing pollen dispersal patterns. Therefore, according to my results, an optimal sampling effort for obtaining adequate estimates in a paternity study should comprise at least 200 overall seeds sampled from at least 10 different mother trees.

The second issue I addressed was linked to the availability of adequate neutral and potentially adaptive molecular markers for the study species, silver fir (*Abies alba* Mill.). Since previously available microsatellite had poor amplification success and were null allele prone, a new set of 16 polymorphic microsatellites was developed from transcriptome sequencing. These markers were assembled in 2 8-plexes with high amplification success and clear band pattern. A third multiplex of genomic microsatellites was assembled choosing the best among the available markers and including 2 newly developed genomic microsatellites. A set of 763 SNPs was developed in a parallel project (Roschanski et al (2013), D. Postolache personal communication) using transcriptome data and candidate genes linked to metabolism, growth and responses to stress. Among them 406 SNPs were successfully amplified and 273 were polymorphic. Polymorphic loci have recently been screened using individuals from two populations I sampled in Central Apennine.

The last issue I focused on was the selection of an adequate zone to study adaptive gene flow. A thorough analysis of biogeographical structure is a key prerequisite for the study of local adaptation and adaptive gene flow in natural populations. The investigation of the spatial patterns of genetic variation and its main drivers is important to understand if the biogeographical structure is influenced solely by the demographic history of the species or also by environmental factors constraining the populations. The biogeographical information can be used to: *i*) identify genetically homogeneous clusters where it is possible to investigate clinal genetic variation along environmental gradients, that can provide evidence of natural selection (Gram and Sork 2001, Grivet et al 2011), and *ii*) areas representing the borders between different genetic clusters characterized by abrupt genetic discontinuities. These areas maximize the probability of observing adaptive gene flow, if any, and thus they are the ideal candidate sites for performing an adaptive gene flow experiment.

In my thesis I investigated the spatial pattern of genetic variation in natural populations of silver fir along the Apennine range. These populations have high evolutionary and conservation value because they are at the rear-edge of the species distribution and they were shown to include putative glacial refugia, thus representing hotspots of genetic diversity. Furthermore, they grow in a wide variety of environmental conditions, because the Apennines are a 1000 km mountain chain with large environmental heterogeneity. Individuals from 16 populations sampled across the whole Apennine range were genotyped using a set of 17 microsatellites (15 chosen from the ones developed in this thesis work). By using multiple clustering approaches, I found clearly separated genetic clusters concordant with geography. These results allowed me to identify: *i*) groups of populations ideal for testing gene-environment correlation, and *ii*) geographically circumscribed areas with abrupt genetic discontinuities. In addition, this study shed light on the post-glacial evolutionary history of silver fir populations in a refugial area, showing surprising phylogeographic relationships and internal dynamics for southern populations and a strong support for a northern refugium close to ice margins.

My thesis work ended with the analysis of the abovementioned biogeographical patterns, but the project is still going on. Gene-environment correlation will be studied within genetic clusters by using SNP markers. Furthermore, the most interesting genetic discontinuity areas will be intensively sampled in order to identify differentiation at finer scale. When fine-scale differentiation is found, gene flow among the closest divergent populations will be studied using paternity and parentage analysis. For these analyses, the sampling guidelines and the molecular markers developed in my thesis will be used. Once immigrant seeds and seedlings

are identified, their allele frequencies at SNP loci will be compared with the ones of locally sired/produced seeds and seedlings. If adaptive gene flow is present allele frequencies in local and immigrant offspring are expected to be significantly different, providing important information on the rate and direction of adaptive gene flow. The use of both paternity and parentage analysis will also allow to consider differences in the survival and establishment rate between local and immigrant individuals in subsequent life-stages.

References

Adams WT, Birkes DS (1991). Estimating mating patterns in forest tree populations. In: *Biochemical Markers in the Population Genetics of Forest Trees* (eds Fineschi S, Malvolti ME, Cannata F, Hattemer HH), pp. 157-172. SPB Academic Publishing, ThevHague, the Netherlands.

Adams WT (1992). Gene dispersal within forest tree populations. New Forests, 6: 217-240.

Aguilar R, Quesada M, Ashworth L, Herrerias-Diego Y, Lobo J (2008). Genetic consequences of habitat fragmentation in plant populations: susceptible signals in plant traits and methodological approaches. Molecular Ecology: 17, 5177–5188.

Aitken SN, YeamanS, Holliday JA, Wang T, Curtis-McLane S (2008). Adaptation, migration or extirpation: climate change outcomes for tree populations. Evolutionary Applications, 1: 95–111.

Aldrich PR, Hamrick JL (1998). Reproductive dominance of pasture trees in a fragmented tropical forest mosaic. Science, 281: 103–105.

Alleaume-Benharira M, Pen IR, Ronce O (2006). Geographical patterns of adaptation within a species' range interactions between drift and gene flow. Journal of Evolutionary Biology, 19:203–215

Apsit VJ, Hamrick JL Nason JD (2001). Breeding population size of a fragmented population of a Costa Rican dry forest tree species. Journal of Heredity, 92: 415-420.

Ashley MV (2010). Plant parentage, pollination, and dispersal: how DNA microsatellites have altered the landscape. Critical Reviews in Plant Sciences, 29: 148-161.

Aussenac G (2002) Ecology and ecophysiology of circum-Mediterranean firs in the context of climate change. Annals of Forest Science, 59, 823–832.

Austerlitz F, Dick CW, Dutech C, Klein EK, Oddou-Muratorio S, Smouse PE et al. (2004). Using genetic markers to estimate the pollen dispersal curve. Molecular Ecology, 13: 937-954.

Bacles CFE, Burczyk J, Lowe AJ, Ennos RA (2005). Historical and contemporary mating patterns in remnant populations of the forest tree *Fraxinus excelsior* L. Evolution, 59: 979-990.

Bacles CFE, Lowe AJ, Ennos RA (2006) Effective seed dispersal across a fragmented landscape. Science, 311:628.

Bacles CFE, Ennos RA (2008). Paternity analysis of pollen-mediated gene flow for *Fraxinus excelsior* L. in a chronically fragmented landscape. Heredity, 101: 368-380.

Bacles CFE, Jump AS (2011). Taking a tree's perspective on forest fragmentation genetics. Trends in Plant Science, 16: 13-18.

Bai W-N, Zeng YF, Zhang DY (2007). Mating patterns and pollen dispersal in a heterodichogamous tree, *Junglans mandshurica* (Junglandaceae). New Phytologist, 176: 699-707.

Barthe S, Gugerli F, Barkley NA, Maggia L, Cardi C, Scotti I (2012). Always look on both sides: Phylogenetic information conveyed by simple sequence repeat allele sequences. PLoS ONE, 7, e40699.

Bittencourt JVM, Sebbenn AM (2007). Patterns of pollen and seed dispersal in a small, fragmented population of the wind-pollinated tree *Araucaria angustifolia* in southern Brazil. Heredity, 99: 580-591.

Begon, M., C. R. Townsend, and J. L. Harper. 2006. Ecology. Fourth edition. Blackwell, Oxford, UK.

Bittencourt JVM, Sebbenn AM (2008). Pollen movement within a continuous forest of wind-pollinated *Araucaria angustifolia*, inferred from paternity and TwoGener analysis. Conservation Genetics, 9: 855-868.

Boshier D, Chase MR, Bawa KS (1995). Population genetics of *Cordia alliodora* (Boraginaceae), a Neotropical tree. 3. Gene flow, neighborhood, and population substructure. American Journal of Botany, 82: 484-490.

Braga AC, Collevatti RG (2011). Temporal variation in pollen dispersal and breeding structure in a bee-pollinated Neotropical tree. Heredity, 106: 911-919.

Brown B, Mitchell R (2001). Competition for pollination: effects of pollen of an invasive plant on seed set of a native congener. Oecologia, 129: 43-49.

Buiteveld J, Bakker EG, Bovenschen J, de Vries SMG (2001). Paternity analysis in a seed orchard of

Quercus robur L. and estimation of the amount of background pollination using microsatellite markers. Forest Genetics, 8: 331-337.

Bullock JM, Clarke RT (2000). Long distance seed dispersal by wind: measuring and modelling the tail of the curve. Oecologia, 124: 506-521.

Bullock JM, Shea K, Skarpaas O.(2006). Measuring plant dispersal: an introduction to field methods and experimental design. Plant Ecol. 186: 217–234.

Burczyk J, Adams WT, Shimizu JY (1996). Mating patterns and pollen dispersal in a natural knobcone pine (Pinus attenuata Lemmon.) stand. Heredity, 77: 251-260.

Burczyk J, Adams WT, Moran GF, Griffin AR (2002). Complex patterns of mating revealed in a *Eucalyptus regnans* seed orchard using allozyme markers and the neighbourhood model. Molecular Ecology, 11: 2379-2391.

Burczyk J (2004). Local pollen dispersal and distant gene flow in Norway spruce (*Picea abies* Karst.). Forest Ecology and Management, 197: 39-48.

Burczyk J, DiFazio SP, Adams WT (2004). Gene flow in forest trees: how far do genes really travel? Forest Genetics, 11: 179-192.

Burczyk J, AdamsWT, Birkes DS, Chybicki IJ (2006). Using genetic markers to directly estimate gene flow and reproductive success parameters in plants on the basis of naturally regenerated seedlings. Genetics, 173: 363-372.

Buschbom J, Yanbaev Y, Degen B (2011). Efficient long-distance gene flow into an isolated relict oak stand. Journal of Heredity, 102: 464-472.

Byrne M, Elliott CP, Yates C, Coates DJ (2007). Extensive pollen dispersal in a bird-pollinated shrub, *Calothamnus quadrifidus*, in a fragmented landscape. Molecular Ecology, 16: 1303-1314.

Byrne M, Elliott CP, Yates C, Coates DJ (2008). Maintenance of high pollen dispersal in *Eucalyptus wandoo*, a dominant tree of the fragmented agricultural region in Western Australia. Conservation Genetics, 9: 97–105.

Carneiro FS, Degen B, Kanashiro M, de Lacerda AEB, Sebbenn AM (2009). High levels of pollen dispersal detected through paternity analysis from a continuous *Symphonia globulifera* population in the Brazilian Amazon. Forest Ecology and Management, 258: 1260-1266.

Carrer M, Nola P, Motta R, Urbinati C (2010). Contrasting tree-ring growth to climate responses of *Abies alba* toward the southern limit of its distribution area. Oikos, 119: 1515-1525.

Chauchard S, Caicaillet C, Guibal F (2007) Pattern of land-use abandonment control tree-recruitment and forest dynamics in Mediterranean mountains. Ecosystems, 10: 936-948

Chauchard S, Beilhe F, Denis N, Carcaillet C (2011) An increase in the upper tree-limit of silver fir (*Abies alba* Mill.) in the Alps since the mid-20th century: a land-use change phenomenon. For Ecol Manage, 259:1406-1415

Cavagna S, Cian S (2003) The National Park of the Casentine Forests, Giunti Editore.

Chaix G, Gerber S, Razafimaharo V, Vigneron P, Verhaegen D, Hamon S (2003). Gene flow estimation with microsatellites in a Malagasy seed orchard of *Eucalyptus grandis*. Theoretical and Applied Genetics, 107: 705-712.

Chapuis MP, Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. Molecular Biology and Evolution, 24: 621–631.

Chase MR, Moller C, Kesseli R, Bawa KS (1996). Distant gene flow in tropical trees. Nature, 383: 398-399.

Cheddadi R, Birks HJB, Tarroso P et al. (2013) Revisiting tree-migration rates*: Abies alba* (Mill.), a case study. Vegetation History and Archaeobotany, in press.

Chybicki IJ, Burczyk J (2009) Simultaneous estimation of null alleles and inbreeding coefficients. Journal of Heredity, 100: 106–113.

Chybicki IJ, Burczyk J (2010a). NM+: software implementing parentage-based models for estimating gene dispersal and mating patterns in plants. Molecular Ecology Resources, 10: 1071-1075.

Chybicki IJ, Burczyk J (2010b). Realized gene flow within mixed stands of *Quercus robur* L. and *Q. petraea* (Matt.) L. revealed at the stage of naturally established seedlings. Molecular Ecology, 19: 2137-2151.

Clark JS (1998). Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord. American Naturalist, 152: 204–24

Clark JS, Silman M, Kern R, Macklin E, HilleRisLambers J (1999). Seed dispersal near and far: patterns across temperate and tropical forests. Ecology, 80:1475–94

Clobert, J, Danchin E, Dhont AA Nichols JD (2001). Dispersal. Oxford University Press: Oxford, UK. Oxford University Press, Oxford.

Clobert J, Baguette M, Benton TG, Bullock JM, Ducatez S (eds) (2012) Dispersal ecology and evolution. Oxford University Press, Oxford.

Cloutier D, Hardy OJ, Caron H, Ciampi AY, Degen B, Kanashiro M, Schoen DJ (2007). Low inbreeding and high pollen dispersal distances in populations of two Amazonian Forest tree species. Biotropica, 39: 406–415.

Corlett RT, Westcott DA. 2013. Will plant movements keep up with climate change? Trends in Ecology & Evolution 28: 482–488.

Cottrell JE, Vaughan SP, Connolly T, Sing L, Moodley DJ, Russell K (2009). Contemporary pollen flow, characterization of the maternal ecological neighbourhood and mating patterns in wild cherry (*Prunus avium* L.). Heredity, 103: 118-28.

Cousens RD, Dytham C, Law R (2008). Dispersal in Plants: A Population Perspective. Oxford University Press: Oxford, UK. Oxford University Press, Oxford.

Craft KJ, Ashley MV (2010). Pollen-mediated gene flow in isolated and continuous stands of bur oak, *Quercus macrocarpa* (Fagaceae). American Journal of Botany, 97: 1999-2006.

Cremer E, Liepelt S, Sebastiani F et al. (2006) Identification and characterization of nuclear microsatellite loci in *Abies alba* Mill. Molecular Ecology Notes, 6, 374–376.

Cremer E, Ziegenhagen B, Schulerowitz K et al. (2012) Local seed dispersal in European silver fir (*Abies alba* Mill.): lessons learned from a seed trap experiment. Trees, 26, 987–966.

Curtu AL, Gailing O, Finkeldey R (2009). Patterns of contemporary hybridization inferred from paternity analysis in a four-oak-species forest. BMC Evolutionary Biology, 9: 284.

Dakin EE & Avise JC (2004). Microsatellite null alleles in parentage analysis. Heredity, 93, 504–509.

Davis M, Shaw R (2001) Range shifts and adaptive responses to Quaternary climate change. Science, 292: 673–679.

de Lacerda AEB, Kanashiro M, Sebbenn AM (2008). Long-pollen movement and deviation of random mating in a low-density continuous population of a tropical tree *Hymenaea courbaril* in the Brazilian Amazon. Biotropica, 40: 462-470.

de Moraes MLT, Sebbenn AM (2011). Pollen dispersal between isolated trees in the Brazilian savannah: a case study of the neotropical tree *Hymenaea stigonocarpa*. Biotropica, 43: 192-199.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM Algorithm. J R Stat Soc B 39: 1–38.

Dick CW (2001). Genetic rescue of remnant tropical trees by an alien pollinator. Proceedings of the Royal Society of London Series Biological Sciences, 268: 2391-2396.

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB (1994). Mutational processes of simple-sequence repeat loci in human populations. PNAS 91: 3166–3170.

Dow BD, Ashley MV (1996) Microsatellite analysis of seed dispersal and parentage of saplings in bur oak, *Quercus macrocarpa*. Molecular Ecology, 5: 615–627.

Dunphy BK, Hamrick JL (2005). Gene flow among established Puerto Rican populations of the exotic tree species, *Albizia lebbeck*. Heredity, 94: 418-425.

Dunphy BK, Hamrick JL (2007). Estimation of gene flow into fragmented populations of *Bursera simaruba* (Burseraceae) in the dry-forest life zone of Puerto Rico. American Journal of Botany, 94: 1786-1794.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources, 4, 359–361.

Eckert CG, Samis KE, Lougheed SC (2008) Genetic variation across species' geographical ranges: the central-marginal hypothesis and beyond. Molecular Ecology, 17:1170-1188

Eckert AJ, van Heerwaarden J, Wegrzyn JL et al. (2010) Patterns of population structure and environmental associations to aridity across the range of Loblolly Pine (Pinus taeda L., Pinaceae). Genetics, 185, 969–

982.

Ellstrand NC, Elam DR (1993). Population genetic consequences of small population size□: Implications for Plant Conservation. Annual Review of Ecology and Systematics, 24, 217–242.

El Mousadik A, Petit RJ (1996) High level of genetic differentiation for allelic richness among populations of the argan tree (*Argania spinosa* (L.) Skeels) endemic to Morocco. Theoretical and Applied Genetics, 92, 832–839.

Endler JA (1977). Geographic Variation, Speciation and Clines. Princeton, NJ: Princeton Univ. Press.

Ennos RA (1994). Estimating relative rates of pollen and seed migration among plant populations. Heredity, 72: 250–259.

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

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

Fahrig L (2003) Effects of habitat fragmentation on biodiversity. Annual Review of Ecology, Evolution and Systematics 34, 487–515.

Falush D, Stephens M & Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics, 164, 1567–1587.

Fan L, Zhang MY, Liu QZ, Li LT, Song Y, Wang LF, Zhang SL, Wu J (2013) Transferability of newly developed pear SSR markers to other Rosaceae species. Plant Mol Biol Rep doi: 10.1007/s11105-013-0586-z

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics, 180, 977–993.

Fuchs EJ, Hamrick JL (2011). Mating system and pollen flow between remnant populations of the endangered tropical tree, *Guaiacum sanctum* (Zygophyllaceae). Conservation Genetics, 12: 175-185.

Fukue Y, Kado T, Lee SL, Ng KKS, Muhammad N, Tsumura Y (2007). Effects of flowering tree density on the mating system and gene flow in *Shorea leprosula* (Dipterocarpaceae) in Peninsular Malaysia. Journal of Plant Research, 120: 413-420.

Funda T, Chen CC, Liewlaksaneeyanawin C, Kenawy AMA, El-Kassaby YA (2008). Pedigree and mating system analyses in a western larch (*Larix occidentalis* Nutt.) experimental population. Annals of Forest Science, 65: 705.

Gaino APSC, Silva AM, Moraes MA, Alves, PF, Moraes MLT, Freitas MLM, Sebbenn AM (2010). Understanding the effects of isolation on seed and pollen flow, spatial genetic structure and effective population size of the dioecious tropical tree species *Myracrodruon urundeuva*. Conservation Genetics, 11: 1631-1643.

Garant D, Forde SE, Hendry AP (2007). The multifarious effects of dispersal and gene flow on contemporary adaptation. Functional Ecology, 21, 434–443.

García C, Arroyo JM, Godoy JA, Jordano P (2005). Mating patterns, pollen dispersal, and the ecological maternal neighbourhood in a *Prunus mahaleb* L. population. Molecular Ecology, 14: 1821-1830.

Garza JC, Williamson EG (2001). Detection of reduction in population size using data from microsatellite loci. Mol Ecol 10: 305–318.

Gauzere J, Klein EK, Oddou-Muratorio S (2013) Ecological determinants of mating system within and between three *Fagus sylvatica* populations along an elevational gradient. Mol. Ecol. http://dx.doi.org/10.1111/mec.12435

Geng Q, Lian C, Goto S, Tao J, Kimura M, Islam MS, Hogetsu T (2008). Mating system, pollen and propagule dispersal, and spatial genetic structure in a high-density population of the mangrove tree *Kandelia candel*. Molecular Ecology, 17: 4724-4739.

Gerber S, Chabrier P, Kremer A (2003). FAMOZ: a software for parentage analysis using dominant, codominant and uniparentally inherited markers. Molecular Ecology Notes, 3: 479-481.

Gillet E, Hattemer HH (1989) Genetic analysis of isoenzyme phenotypes using single tree progenies. Heredity, 63, 135–141.

Ghosh P (1951). Random distance within a rectangle and between two rectangles. Bulletin of Calcutta

Mathematical Society, 43: 17-24

Gömöry D, Paule L, Krajmerová D, Romšáková I, Longauer R (2012) Admixture of genetic lineages of different glacial origin: a case study of *Abies alba* Mill. in the Carpathians. Plant Systematics and Evolution, 298, 703–712.

Gonzalez-Martinez SC, Krutovsky KV, Neale DB (2006) Forest-tree population genomics and adaptive evolution. New Phytologist, 170: 227-238

Goto S, Shimatani K, Yoshimaru H, Takahashi Y (2006). Fat-tailed gene flow in the dioecious canopy tree species *Fraxinus mandshurica* var. *japonica* revealed by microsatellites. Molecular Ecology, 15: 2985-2996.

Goudet J (2001) Fstat, a Program to Estimate and Test Gene Diversities and Fixation Indices (Version 2.9.3). Institut

d'Ecologie, Universite¢ de Lausanne, Dorigny, Switzerland.

GramWK, Sork VL (2001) Association between environmental and genetic heterogeneity in forest tree populations. Ecology 82:2012–2021.

Gregorius H-R, Kownatzki D, Höltken AM (2011). Spatial patterns of mating relations in wild cherry (Prunus avium L.). Perspectives in Plant Ecology, Evolution and Systematics, 13: 37-45.

Grivet D, Sebastiani F, Alia R et al. (2010) Molecular footprints of local adaptation in two Mediterranean conifers. Molecular Biology and Evolution, 28: 101-107.

Grosser C, Potts B, Vaillantcourt RE (2010). Microsatellite based paternity analysis in a clonal *Eucalyptus nitens* seed orchard. Silvae Genetica, 59: 57-62.

Guehl JM, Aussenac G, Bouachrine J et al. (1991) Sensitivity of leaf gas-exchange to atmospheric drought, soil drought, and water-use efficiency in some Mediterranean Abies species. Canadian Journal Of Forest Research, 10, 1507–1515.

Guichoux E, Lagache L, Wagner S et al. (2011) Current trends in microsatellite genotyping. Molecular Ecology Resources, 11, 591–611.

Guo SW, Thompson EA (1992) Performing the exact test of Hardy–Weinberg proportions for multiple alleles. Biometrics 48: 361–372.

Hadfield JD, Richardson DS, Burke T (2006). Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. Molecular Ecology, 15: 3715-3730.

Hampe A, Petit RJ (2005). Conserving biodiversity under climate change: the rear edge matters. Ecology Letters, 8: 461–467.

Hamrick, J. L. 2004. Response of forest trees to global environmental changes. Forest Ecology and Management, 197: 323–335.

Hanaoka S, Yuzurihara J, Asuka Y, Tomaru N, Tsumura Y, Kakubari Y, Mukai Y (2007). Pollen-mediated gene flow in a small, fragmented natural population of *Fagus crenata*. Canadian Journal of Botany, 85: 404-413.

Hansen JK, Larsen JB (2004) European silver fir (Abies alba Mill.) provenances from Calabria, southern Italy: 15 year results from Danish provenance field trials. Eur. J. Forest Res., 123: 127–138.

Hansen OK, Vendramin GG, Sebastiani F, Edwards KJ (2005). Development of microsatellite markers in *Abies nordmanniana* (Stev.) Spach and cross-species amplification in the Abies genus. Molecular Ecology Notes, 5, 784–787.

Hansen OK, Kjær ED (2006). Paternity analysis with microsatellites in a Danish *Abies nordmanniana* clonal seed orchard reveals dysfunctions. Canadian Journal of Forest Research, 36: 1054-1058.

Hansen OK, Nielsen UB, Kongevej H (2008) Crossing success in Abies nordmanniana following artificial pollination with a pollen mixture of *A. nordmanniana* and *A. alba*. Silvae Genetica, 57, 70–75.

Hanson TR, Brunsfeld SJ, Bryan F, Waits LP (2008). Pollen dispersal and genetic structure of the tropical tree *Dipteryx panamensis* in a fragmented Costa Rican landscape. Molecular Ecology, 17: 2060-2073.

Hardy OJ (2009). How fat is the tail? Heredity, 103: 437-438.

Hasegawa Y, Suyama Y, Seiwa K (2009). Pollen donor composition during the early phases of reproduction revealed by DNA genotyping of pollen grains and seeds of *Castanea crenata*. New Phytologist: 994-1002.

Hintze C, Heydel F, Hoppe C, Cunzea S, König A & Tackenberg O. (2013) D3: the Dispersal and

Diaspore Database – baseline data and statistics on seed dispersal. Perspectives in Plant Ecology, Evolution and Systematics, 15, 180□192.

Hirsch BT, Visser MD, Kays R, Jansen PA (2012). Quantifying seed dispersal kernels from truncated seed-tracking data. Methods in Ecology and Evolution, 3: 595-602.

Hoebee SE, Arnold U, Düggelin C, Gugerli F, Brodbeck S, Rotach P, Holderegger R (2007). Mating patterns and contemporary gene flow by pollen in a large continuous and a small isolated population of the scattered forest tree *Sorbus torminalis*. Heredity, 99: 47–55.

Hoffman AA & Sgro CM (2011). Climate change and evolutionary adaptation. Nature, 470, 479–485.

Hufford KM, Hamrick JL, Rathbun SL (2009). Male reproductive success at three early life stages in the tropical tree *Platypodium elegans*. International Journal of Plant Sciences, 170: 724-734.

Huntley B, Birks HJB (1983) An atlas of past and present pollen maps of Europe: 0–13,000 years ago. Cambridge University Press,

Cambridge.

Irwin AJ, Hamrick JL, Godt MJW, Smouse PE (2003). A multiyear estimate of the effective pollen donor pool for *Albizia julibrissin*. Heredity, 90: 187-194.

Isagi Y, Kanazashi T, Suzuki W, Tanaka H, Abe T (2004). Highly variable pollination patterns in *Magnolia obovata* revealed by microsatellite paternity analysis. International Journal of Plant Sciences, 165: 1047–1053.

Jakobsson M, Edge MD, Rosenberg NA (2013) The relationship between F_{ST} and the frequency of the most frequent allele. Genetics 193:515-528

Jamieson A, Taylor SCS (1997) Comparisons of three probability formulae for parentage exclusion. Animal Genetics, 28, 397–400.

Jones AG, Ardren WR (2003) Methods of parentage analysis in natural populations. Molecular Ecology, 12: 2511-2523

Jones FA, Chen J, Weng G, Hubbell SP (2005). A genetic evaluation of seed dispersal in the neotropical tree *Jacaranda copaia* (Bignoniaceae). The American Naturalist, 166: 543-555.

Jones FA, Muller-Landau HC (2008). Measuring long-distance seed dispersal in complex natural environments: an evaluation and integration of classical and genetic methods. Journal of Ecology, 96: 642-652.

Jones ME, Shepherd M, Henry R, Delves A (2008). Pollen flow in *Eucalyptus grandis* determined by paternity analysis using microsatellite markers. Tree Genetics & Genomes, 4: 37-47.

Jones AG, Small CM, Paczolt KA, Ratterman NL (2010). A practical guide to methods of parentage analysis. Molecular Ecology Resources, 10: 6-30.

Jordano P (2007). Frugivores, seeds and genes: analysing the key elements of seed shadows. In: Dennis A, Green R, Schupp EW and Wescott D (eds) Frugivory and seed dispersal: theory and applications in a changing world, Commonwealth Agricultural Bureau International, Wallingford, UK. pp 229-251.

Jost L (2008) G_{ST} and its relatives do not measure differentiation. Molecular Ecology, 17, 4015–4026.

Jump AS, Pefiuelas J. 2005. Running to stand still: adaptation and the response of plants to rapid climate change. Ecol. Lett. 8:1010-1020.

Jump AS, Penuelas J (2006) Genetic effects of chronic habitat fragmentation in a wind-pollinated tree. Proc Natl Acad Sci, 103: 8096-8100.

Kalinowski ST (2005) HP-Rare 1.0: a computer program for performing rarefaction on measures of allelic richness. Molecular Ecology Notes, 5. 187–189.

Kalinowski ST, Taper ML, Marshall TC (2007). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Molecular Ecology, 16: 1099-1006.

Kalinowski ST (2009) How well do evolutionary trees describe genetic relationships between populations? Heredity 102:506-513.

Kameyama Y, Isagi Y, Naito K, Nakagoshi N (2000). Microsatellite analysis of pollen flow in *Rhododendron metternichii* var. *hondoense*. Ecological Research, 15: 263-269.

Kamm U, Rotach P, Gugerli F, Siroky M, Edwards P, Holderegger R (2009). Frequent long-distance gene flow in a rare temperate forest tree (*Sorbus domestica*) at the landscape scale. Heredity, 103: 476-482.

Katul GG, Porporato A, Nathan R, Siqueira M, Soons MB, Poggi D, Horn HS, Levin SA (2005)

Mechanistic analytical models for long-distance seed dispersal by wind. American Naturalist, 166: 368–381.

Kaufman SR, Smouse PE, Alvarez-Buylla ER (1998). Pollen-mediated gene flow and differential male reproductive success in a tropical pioneer tree, *Cecropia obtusifolia* Bertol. (Moraceae): a paternity analysis. Heredity, 81: 164-173.

Keenan K, McGinnity P, Cross TF, Crozier WW, Prodohl PA (2013) diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. Methods in Ecology and Evolution, in press.

Kenta T, Isagi Y, Nakagawa M, Yamashita M, Nakashizuka T (2004). Variation in pollen dispersal between years with different pollination conditions in a tropical emergent tree. Molecular Ecology, 13: 3575-3584.

Kaltenrieder P, Belis C, Hofstetter S, Ammann B, Ravazzi C, Tinner W (2009) Environmental and climatic conditions at a potential

glacial refugial site of tree species near the southern Alpine glaciers. New insights from multiproxy sedimentary studies at Lago della Costa (Euganean Hills, northeastern Italy). Q Sci Rev, 28: 2647–2662

Kim K, Ratcliffe S, French B, Liu L, Sappington T (2008) Utility of EST-derived SSRs as population genetics markers in a beetle. Journal of Heredity, 99: 112–124.

Klein EK, Laredo C (1999). Optimal sampling designs for studies of gene flow: a comment on Assunção and Jacobi. Evolution, 53: 2002-2005.

Klein EK, Lavigne C, Picault H, Renard M, Gouyon PH (2006). Pollen dispersal of oilseed rape: estimation of the dispersal function and effects of field dimension. J Appl Ecol, 43: 141-151.

Klein EK, Oddou-Muratorio S (2011). Pollen and seed dispersal inferred from seedling genotypes: the Bayesian revolution has passed here too. Mol Ecol, 20: 1077-1079.

Knapp EE, Goedde MA, Rice KJ (2001) Pollen-limited reproduction in blue oak: implications for wind pollination in fragmented populations. Oecologia, 128: 48-55.

Konnert M, Bergmann F (1995) The geographical distribution of genetic variation of silver fir (*Abies alba*, Pinaceae) in relation to its migration history. Plant Systematics and Evolution, 196.

Konuma A, Tsumura Y, Lee CT, Lee SL, Okuda T (2000). Estimation of gene flow in the *tropical-rainforest tree Neobalanocarpus heimii* (Dipterocarpaceae), inferred from paternity analysis. Molecular Ecology, 9: 1843-1852.

Kovach A, Wegrzyn J, Parra G et al. (2010) The Pinus taeda genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics, 11: 420.

Kramer AT et al. (2008). The paradox of forest fragmentation genetics. Conservation Biology, 22: 878-885

Krutovsky KV, Burczyk J, Chybicki I, Finkeldey R, Pyhäjärvi T,Robledo-Arnuncio JJ (2012) Gene flow, spatial structure, local adaptation and assisted migration in trees in Genomics of Tree Crops, pp 71-116, Springer New York.

Kremer A, Ronce O, Robledo-Arnuncio JJ, Guillaume F, Bohrer G, Nathan R et al. (2012). Long-distance gene flow and adaptation of forest trees to rapid climate change. Ecology Letters, 15: 378-392.

Kuparinen A, Katul G, Nathan R & Schurr FM (2009). Increases in air temperature can promote wind-driven dispersal and spread of plants. Proc. R. Soc. Lond., B, Biol. Sci., 276: 3081–3087.

Lander TA, Boshier DH, Harris SA (2010). Fragmented but not isolated: contribution of single trees, small patches and long-distance pollen flow to genetic connectivity for *Gomortega keule*, an endangered Chilean tree. Biological Conservation, 143: 2583-2590.

Larsen JB (1986) Die geografische Variation der Weißtanne (Abies alba Mill.)—Wachstumsentwicklung und Frostresistenz. Forstw Centralbl, 105: 396–406

Larsen JB, Mekic F (1991) The geographic variation in in European silver-fir (Abies alba Mill.). Silv Genet, 40:188–198

Larsen AS, Kjær ED (2009). Pollen mediated gene flow in a native population of *Malus sylvestris* and its implications for contemporary gene conservation management. Conservation Genetics, 10: 1637-1646.

Latouche-Hallé C, Ramboer A, Bandou E, Caron H,

Kremer A (2004). Long-distance pollen flow and tolerance to selfing in a neotropical tree species. Molecular Ecology, 13: 1055-1064.

Lee SL, Ng KKS, Saw LG, Lee CT, Muhammad N, Tani N, Tsumura Y, Koskela J (2006). Linking the gaps between conservation research and conservation management of rare dipterocarps: a case study of *Shorea lumutensis*. Biological Conservation, 131: 72-92.

Lefevre F, Koskela J, Hubert J et al. (2013) Dynamic conservation of forest genetic resources in 33 European countries. Conservation Biology, 27: 373–384.

Leinemann L, Hattemer H (2006). Genetic variation and mating pattern in a stand of yew (*Taxus baccata* L.). Allgemeine Forst und Jagdzeitung, 177: 217-224.

Lenormand T (2002) Gene flow and the limits to natural selection. TREE, 17:183-189

Leonardi S, Piovani P, Scalfi M et al. (2012) Effect of habitat fragmentation on the genetic diversity and structure of peripheral populations of beech in Central Italy. Journal of Heredity, 103: 408–417.

Leonarduzzi,C, Leonardi S, Menozzi P, Piotti A (2012). Towards an optimal sampling effort for paternity analysis in forest trees: what do the raw numbers tell us? Iforest, 5: 18-25.

Levin, SA, Muller-Landau HC, Nathan R, & Chave J (2003). The ecology and evolution of seed dispersal: A Theoretical Perspective. Annual Review of Ecology, Evolution, and Systematics, 34: 575–604.

Lian C, Miwa M, Hogetsu T (2001). Outcrossing and paternity analysis of *Pinus densiflora* (Japanese red pine) by microsatellite polymorphism. Heredity, 87: 88-98.

Liepelt S, Bialozyt R, Ziegenhagen B (2002). Wind-dispersed pollen mediates gene flow among refugia. Proceedings of the National Academy of Sciences, 99: 14590–14594.

Liepelt S, Cheddadi R, De Beaulieu JL, Fady B, Goemoery D, Hussendoerfer E, Konnert M, Litt T, Longauer R, Terhuerne-Berson R & Ziegenhagen B (2009). Postglacial range expansion and its genetic imprints *in Abies alba* (Mill.) - a synthesis from palaeobotanic and genetic data. Review of Palaeobotany and palynology, 153: 139–149 doi:10.1016/j.revpalbo.2008.07.007

Liepelt S, Mayland-Quellhorst E, Lahme M, Ziegenhagen B (2010) Contrasting geographical patterns of ancient and modern genetic

lineages in Mediterranean Abies species. Plant Syst Evol, 284:141–151.

Linares JC (2011) Biogeography and evolution of Abies (Pinaceae) in the Mediterranean Basin: the roles of long-term climatic change and glacial refugia. JBiogeogr, 38: 619–630.

Linares J, Camarero JJ (2012) Growth patterns and sensitivity to climate predict silver fir decline in the Spanish Pyrenees. European Journal of Forest Research, 131: 1001–1012.

Lopez S, Rousset F, Shaw FH, Shaw RG, Ronce O. (2008). Migration load in plants: role of pollen and seed dispersal in heterogeneous landscapes. J. Evol. Biol., 21: 294–309.

Maciaroni M, Urbinati C, Gallucci V. (2012). Assetto strutturale e gestione di cenosi residuali di abete bianco *Abies alba* (Mill.) nell'ex Massa Trabaria. Bachelor Thesis. Università Politecnica delle Marche.

Macias M, Andreu L, Bosch O, Camarero JJ, Gutiérrez E (2006) Increasing aridity is enhancing silver fir *Abies alba* (Mill.) water stress in its south-western distribution limit. Climatic Change, 79: 289–313.

Maiorano L, Cheddadi R, Zimmermann NE et al. (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. Global Ecology and Biogeography, 22: 302–317.

Malausa T, Gilles A, Meglecz E et al. (2011). High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries. Molecular Ecology Resources, 11: 638–644.

Moran EV, Clark JS (2011). Estimating seed and pollen movement in a monoecious plant: a hierarchical Bayesian approach integrating genetic and ecological data. Mol Ecol, 20: 1248-1262.

Nakanishi A, Tomaru N, Yoshimaru H, Kawahara T, Manabe T, Yamamoto S (2004). Patterns of pollen flow and genetic differentiation among pollen pools in *Quercus salicina* in a warm temperate old–growth evergreen broad-leaved forest. Silvae Genetica, 53: 258-264.

Narum SR, Hess JE (2011) Comparison of FST outlier tests for SNP loci under selection. Mol Ecol Resour, 11: 184-194

Nathan R, Muller-Landau H (2000). Spatial patterns of seed dispersal, their determinants and consequences for recruitment. Trends Ecol Evol, 15: 278-285.

Nathan R, Safriel UN, Noy-Meir I (2001) Field validation and sensitivity analysis of a mechanistic model for tree seed dispersal by wind. Ecology, 82: 374–388.

Nathan R, Katul GG, Horn HS, Thomas SM, Oren R, Avissar R, Pacala SW, Levin SA (2002) Mechanisms of long-distance dispersal of seeds by wind. Nature, 418: 409–413.

Nathan R, Sapir N, Trakhtenbrot A, Katul GG, Bohrer G, Otte M, Avissar R, Soons MB, Horn HS, Wikelski M, Levin SA (2005) Long-distance biological transport processes through the air: Can nature's complexity be unfolded in-silico? Divers Distrib, 11:131–137.

Nathan R (2006). Long-distance dispersal of plants. Science, 313: 786-788.

Nathan R et al.(2008) A movement ecology paradigm for unifying organismal movement research. Proc Natl Acad Sci USA 105:19052–19059.

Nathan R, Klein EK, Robledo-Arnuncio JJ, Revilla E (2012). Dispersal kernels: review. In: Clobert J, Baguette M, Benton T, Bullock J and Ducatez S (eds) Dispersal ecology and evolution, Oxford University Press. pp 187-210.

Neale DB, Ingvarsson PK (2008) Population, quantitative and comparative genomics of adaptation in forest trees. Curr Opin Plant Biol, 11:149-155

Niggemann M, Wiegand T, Robledo-Arnuncio JJ, Bialozyt R (2012). Marked point pattern analysis on genetic paternity data for uncertainty assessment of pollen dispersal kernels. J Ecol, 100: 264-276.

O'Connell LM, Mosseler A, Rajora OP (2006) Impacts of forest fragmentation on the mating system and genetic diversity of white spruce (*Picea glauca*) at the landscape level. Heredity, 97: 418-426.

Oddou-Muratorio S, Houot M-L, Demesure-Musch B, Austerlitz F (2003). Pollen flow in the wildservice tree, *Sorbus torminalis* (L.) Crantz. I. Evaluating the paternity analysis procedure in continuous populations. Molecular Ecology, 12: 3427-3439.

Oddou-Muratorio S, Klein EK, Austerlitz F (2005). Pollen flow in the wildservice tree, *Sorbus torminalis* (L.) Crantz. II. Pollen dispersal and heterogeneity in mating success inferred from parent-offspring analysis. Molecular Ecology, 14: 4441–4452.

Oddou-Muratorio S, Vendramin GG, Buiteveld J, Fady B (2009) Population estimators or progeny tests: what is the best method to assess null allele frequencies at SSR loci? Conservation Genetics, 10: 1343–1347.

Oddou-Muratorio S, Bontemps A, Klein EK, Chybicki IJ, Vendramin GG, Suyama Y (2010). Comparison of direct and indirect genetic methods for estimating seed and pollen dispersal in *Fagus sylvatica* and *Fagus crenata*. For Ecol Manage, 259: 2151-2159.

Ortego J, Riordan EC, Gugger PF, Sork VL (2012) Influence of environmental heterogeneity on genetic diversity and structure in an endemic southern Californian oak. Molecular Ecology, 21: 3210–3223.

Ouborg, N.J. et al. (1999) Population genetics, molecular markers and the study of dispersal in plants. J. Ecol., 87: 551–568

Pairon M, Jonard M, Jacquemart AL (2006). Modeling seed dispersal of black cherry, an invasive forest tree: how microsatellites may help? Canadian Journal of Forest Research, 36: 1385–1394.

Pakkad G, Al Mazrooei S, Blakesley D, James C, Elliott S, Luoma-aho T, Koskela J (2008a). Genetic variation and gene flow among *Prunus cerasoides* D. Don populations in northern Thailand: analysis of a rehabilitated site and adjacent intact forest. New Forests, 35: 33-43.

Pakkad G, Ueno S, Yoshimaru H (2008b). Gene flow pattern and mating system in a small population of *Quercus semiserrata* Roxb. (Fagaceae). Forest Ecology and Management, 255: 3819-3826.

Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics, 11: 180.

Parducci L, Szmidt AE, Villani F, Wang XR, Cherubini M (1996) Genetic variation of *Abies alba* in Italy. Hereditas, 125: 11–18.

Parducci L, Szmidt AE, Madaghiele A, Anzidei M, Vendramin GG (2001) Genetic variation at chloroplast microsatellites (cpSSRs) in *Abies*

nebrodensis (Lojac.) Mattei and three neighbouring Abies species. Theor Appl Genet, 102: 733–740

Peakall R, Smouse P (2012) GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research – an update. Bioinformatics, 19: 2537–2539.

Pérez-Figueroa A, Garcia-Pereira MJ, Saura M, Rolan-Alvarez E, Caballero A (2010) Comparing three different methods to detect selective loci using dominant markers. J Evol Biol, 23: 2267-2276

Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C, Brewer S, et al. (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. Science 300: 1563–1565.

Petit RJ, Deguilloux MF, Chat J, Grivet D, Garnier-Géré P, Vendramin GG (2005) Removing the bias due to different numbers of repeats when comparing measures of diversity at microsatellite loci. Mol Ecol, 14: 885-890

Pfeiffer A, Olivieri AM, Morgante M (1997) Identification and characterization of microsatellites in Norway spruce (*Picea abies* K.). Genome, 40: 411–419.

Pielaat A, Lewis M, Lele S, Decaminobeck T (2006). Sequential sampling designs for catching the tail of dispersal kernels. Ecological Modelling, 190: 205-222.

Pinzauti F, Sebastiani F, Budde KB, Fady B, Gonzalez-Martinez SC, Vendramin GG (2012) Nuclear microsatellite for *Pinus pinea* (Pinaceae), a genetically depauperate tree and their transferability to *P. halepensis*. American Journal of Botany, 99: 362–365.

Piotti A, Leonardi S, Piovani P, Scalfi M, Menozzi P (2009). Spruce colonization at treeline: where do those seeds come from? Heredity, 103: 136-145.

Piotti A, Leonardi S, Buiteveld J et al. (2012) Comparison of pollen gene flow among four European beech (*Fagus sylvatica* L.) populations characterized by different management regimes. Heredity, 108: 322–331.

Piovani P, Leonardi S, Piotti A, Menozzi P (2010) Conservation genetics of small relic populations of Silver fir (*Abies alba* Mill.) in northern Apennines. Plant Biosystems, 144: 683–691.

Piry S, Luikart G & Cornuet JM (1999). Bottleneck: a computer program for detecting recent reductions in the effective population size using allele frequency data. J. Hered., 90: 502–503.

Pluess AR, Sork VL, Dolan B, Davis FW, Grivet D, Merg K, Papp J, Smouse PE (2009). Short distance pollen movement in a wind-pollinated tree, *Quercus lobata* (Fagaceae). Forest Ecology and Management, 258: 735-744.

Pollegioni P, Woeste K, Mugnozza GS, Malvolti ME (2009). Retrospective identification of hybridogenic walnut plants by SSR fingerprinting and parentage analysis. Molecular Breeding, 24: 321-335.

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

R Development Core Team (2012). R: A language and environment for statistical computing. (eds. R Foundation for Statistical Computing, Vienna, Austria).

Rathmacher G, Niggemann M, Köhnen M, Ziegenhagen B, Bialozyt R (2009). Short-distance gene flow in *Populus nigra* L. accounts for small-scale spatial genetic structures: implications for in situ conservation measures. Conservation Genetics, 11: 1327-1338.

Ravazzi C, Donegana M, Vescovi E, Arpenti E, Caccianiga M, Kaltenrieder P, Londeix L, Marabini S, Mariani S, Pini R, Vai GB, Wick L (2006) A new late-glacial site with *Picea abies* in the northern Apennine foothills: an exception to the model of glacial refugia of trees. Veget Hist Archaeobot 15: 357–371.

Ribbens E, Silander JA, Pacala SW (1994). Seedling recruitment in forests: calibrating models to predict patterns of tree seedling dispersion. Ecology 75: 1794-1806.

Ritland K (2002). Extensions of models for the estimation of mating systems using n indipendent loci. Heredity, 88: 221-228.

Robledo-Arnuncio JJ, Gil L (2005). Patterns of pollen dispersal in a small population of *Pinus sylvestris* L. revealed by total-exclusion paternity analysis. Heredity, 94: 13-22.

Robledo-Arnuncio JJ, Garcia C (2007). Estimation of the seed dispersal kernel from exact identification of source plants. Molecular Ecology 16: 5098–5109.

Robledo-Arnuncio JJ, Austerlitz F, Smouse PE (2007). POLDISP: a software package for indirect estimation of contemporary pollen dispersal. Molecular Ecology Notes, 7: 763-766.

Robledo-Arnuncio JJ (2011). Wind pollination over mesoscale distances: an investigation with Scots pine. New Phytologist, 190: 222-233.

Robledo-Arnuncio JJ (2012). Joint estimation of contemporary seed and pollen dispersal rates among plant populations. Mol Ecol Res, 12: 299-311.

Roschanski AM, Fady B, Ziegenhagen B, Liepelt S (2013) Annotation and re-sequencing of genes from de novo transcriptome assembly of *Abies alba* (Pinaceae). Application in Plant Sciences, 1: 1200179.

Rousset F (2008) GENEPOP'007: A complete re-implementation of the GENEPOP software for Windows and Linux. Molecular Ecology Resources, 8, 103–106.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods in Molecular Biology, 132: 365–386.

Sagnard F, Barberot C, Fady B (2002) Structure of genetic diversity in *Abies alba* Mill. from southwestern Alps: multivariate analysis of adaptive and non-adaptive traits for conservation in France. Forest Ecology and Management, 157: 175–189.

Salvini D, Bruschi P, Fineschi S, Grossoni P, Kjær ED, Vendramin GG (2009). Natural hybridisation between *Quercus petraea* (Matt.) Liebl. and *Quercus pubescens* Willd. within an Italian stand as revealed by microsatellite fingerprinting. Plant Biology, 11: 758-765.

Savolainen O, Pyhajarvi T, Knurr T (2007). Gene flow and local adaptation in trees. Annual Review of Ecology, Evolution, and Systematics, 38: 595-619.

Schiffers, K., Bourne, E.C., Lavergne, S., Thuiller, W. & Travis, J.M. (2013). Limited evolutionary rescue of locally adapted populations facing climate change. Philos. T. R. Soc. Lon. B. , 368.

Schnabel A, Hamrick JL (1995). Understanding the population genetic structure of *Gleditsia triacanthos* L.: the scale and pattern of pollen gene flow. Evolution, 49: 921-931.

Schoebel C, Brodbeck S, Buehler D et al. (2013) Lessons learned from microsatellite development for nonmodel organisms using 454 pyrosequencing. Journal of Evolutionary Biology, 26: 600–611.

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. Nature Biotechnology, 18: 233–234.

Schuster WSF, Mitton JB (2000). Paternity and gene dispersal in limber pine (*Pinus flexilis* James). Heredity, 84: 348-361.

Sebastiani F, Pinzauti F, Kujala ST, Gonzalez-Martinez SC, Vendramin GG (2012) Novel polymorphic nuclear microsatellite markers for Pinus sylvestris L. Conservation Genetic Resources, 4: 231–234.

Sexton, J. P., S. Y. Strauss, and K. J. Rice. 2011. Gene flow increases fitnessat the warm edge of a species' range. Proc. Natl. Acad. Sci. USA 108:11704–11709.

Setsuko S, Ishida K, Ueno S, Tsumura Y, Tomaru N (2007). Population differentiation and gene flow within a metapopulation of a threatened tree, *Magnolia stellata* (Magnoliaceae). American Journal of Botany, 94: 128-136.

Sharma CM, Khanduri VP (2007). Pollen-mediated gene flow in Himalayan long needle pine (*Pinus roxburghii* Sargent). Aerobiologia, 23: 153-158.

Silva MB, Kanashiro M, Ciampi AY, Thompson I, Sebbenn AM (2008). Genetic effects of selective logging and pollen gene flow in a low-density population of the dioecious tropical tree *Bagassa guianensis* in the Brazilian Amazon. Forest Ecology and Management, 255: 1548-1558.

Silva CRS, Albuquerque PSB, Ervedosa FR, Mota JWS, Figueira A, Sebbenn AM (2011). Understanding the genetic diversity, spatial genetic structure and mating system at the hierarchical levels of fruits and individuals of a continuous *Theobroma cacao* population from the Brazilian Amazon. Heredity, 106: 973-985.

Silvertown J (1991). Dorothy's dilemma and the unification of plant population biology. Trends in Ecology and Evolution 6: 346–348.

Skarpaas O, Shea K, Bullock JM (2005). Optimizing dispersal study design by Monte Carlo simulation. Journal of Applied Ecology, 42: 731-739.

Skarpaas O, Shea K (2007). Dispersal patterns, dispersal mechanisms, and invasion wave speeds for invasive thistles. Am Nat, 170: 421-430.

Skarpaas, O., Shea, K. & Jongejans, E. (2011) Watch your time step: trap-ping and tracking dispersal in autocorrelated environments. Methods in Ecology and Evolution, 2: 407–415.

Slatkin M (1987). Gene flow and the geographic structure of natural populations. Science, 236: 787–792.

Slavov GT, Howe GT, Adams WT (2005). Pollen contamination and mating patterns in a Douglas-fir seed orchard as measured by simple sequence repeat markers. Canadian Journal of Forest Research, 35: 1592-1603.

Slavov GT, Howe GT, Gyaourova AV, Birkes DS, Adams WT (2005) Estimating pollen flow using SSR markers and paternity exclusion: accounting for mistyping. Molecular Ecology, 14: 3109–3121.

Slavov GT, Leonardi S, Burczyk J, Adams WT, Strauss SH, DiFazio SP (2009). Extensive pollen flow in two ecologically contrasting populations of *Populus trichocarpa*. Molecular Ecology, 18: 357–373.

Smouse PE, Sork VL (2004) Measuring pollen flow in forest trees: a comparison of alternative approaches. For. Ecol. Manage., 197: 21–38

Sork VL, Smouse PE (2006) Genetic analysis of landscape connectivity in tree populations. Landscape Ecology, 21: 821–836.

Sork VL, Davis FW, Westfall R et al. (2010) Gene movement and genetic association with regional climate gradients in California valley oak (Quercus lobata N_ee) in the face of climate change. Molecular Ecology: 19, 3806–3823.

Soto-Cerda BJ, Cloutier S (2013) Outlier loci and selection signatures of simple sequence repeats (SSRs) in flax (*Linum usitatissimum L.*). Plant Mol Biol Rep doi:10.1007/s11105-013-0568-1

Spielman D, Brook BW, Frankham R. (2004). Most species are not driven to extinction before genetic factors impact them. Proceedings of the National Academy of Sciences, 42: 15261–15264

Stacy EA, Hamrick JL, Nason JD, Hubbell SP, Foster RB, Condit R (1996). Pollen dispersal in low-density populations of three Neotropical tree species. American Naturalist, 148: 275–298.

Stoyan D, Wagner S (2001). Estimating the fruit dispersion of anemochorous forest trees. Ecological Modelling, 145: 35-47.

Streiff R, Ducousso A, Lexer C, Steinkellner H, Gloessl J, Kremer A (1999). Pollen dispersal inferred from paternity analysis in a mixed oak stand of *Quercus robur* L . and *Q. petraea* (Matt .) Liebl . Molecular Ecology, 8: 831-841.

Tabbener H (2003). The use of PCR based DNA markers to study the paternity of poplar seedlings. Forest Ecology and Management, 179: 363-376.

Tani N, Tsumura Y, Kado T, Taguchi Y, Lee SL, Muhammad N, Ng KKS, Numata S, Nishimura S, Konuma A, Okuda T (2009). Paternity analysis-based inference of pollen dispersal patterns, male fecundity variation, and influence of flowering tree density and general flowering magnitude in two dipterocarp species. Annals of Botany, 104: 1421-1434.

Tarazi R, Sebben AM, Mollinari M, Vencovsky R (2010) Mendelian inheritance, linkage and linkage disequilibrium in microsatellite loci of *Copaifera langsdorffii* Desf. Conservation Genetics Resources, 2: 201–204.

Tarazi R, Sebbenn AM., Kageyama PY, & Vencovsky R (2013). Edge effects enhance selfing and seed harvesting efforts in the insect-pollinated Neotropical tree *Copaifera langsdorffii* (Fabaceae). Heredity, 110: 578–85.

Temunović M, Franjić J, Satovic Z, Grgurev M, Frascaria-Lacoste N, Fernández-Manjarrés JF (2012) Environmental heterogeneity explains the genetic structure of Continental and Mediterranean populations of *Fraxinus angustifolia* Vahl. PLoS ONE, 7: e42764.

Terhurne-Berson R, Litt T, Cheddadi R (2004). The spread of Abies throughout Europe since the last glacial period: combined macrofossil and pollen data. Vegetation History and Archaeobotany, 13: 257-268.

Ueno S, Moriguchi Y, Uchiyama K, Ujino-Ihara T, Futamura N, Sakurai T, Shinohara K, Tsumura Y (2012) A second generation framework for the analysis of microsatellites in expressed sequence tags and the development of EST-SSR markers for a conifer, *Cryptomeria japonica*. BMC Genomics, 13:136.

Urbinati C. e Romano R. (a cura), Foresta e Monaci di Camaldoli: un rapporto millenario tra gestione e conservazione, Codice Forestale Camaldolese: III

volume, INEA, Roma, (2012).

Vanden Broeck A, Cottrell J, Quataert P, Breyne P, Storme V, Boerjan W, Van Slycken J (2006). Paternity analysis of *Populus nigra* L. offspring in a Belgian plantation of native and exotic poplars. Annals of Forest Science, 63: 783-790.

van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting

genotyping errors in microsatellite data. Molecular Ecology Notes: 4, 535–538.

Van Rossum F, Stiers I, Van Geert A, Triest L, Hardy OJ (2011). Fluorescent dye particles as pollen analogues for measuring pollen dispersal in an insect-pollinated forest herb. Oecologia, 165: 663-674.

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: Features and applications. Trends in Biotechnology: 23, 48–55

Vendramin GG, Degen B, Petit RJ, Anzidei M, Madaghiele A, Ziegenhagen B (1999) High level of variation at *Abies alba* chloroplast microsatellite loci in Europe. Molecular Ecology, 8, 1117–1126.

Vescovi E, Ammann B, Ravazzi, C, and Tinner W (2010): A new Late-glacial and Holocene record of vegetation and fire history from Lago del Greppo, northern Apennines, Italy, Veg. Hist. Archaeobot., 19: 219–233.

Vicario F, Vendramin GG, Rossi P, Lio P, Giannini R (1995) Allozyme, chloroplast DNA and RAPD markers for determining genetic relationships between *Abies alba* and the relict population of *Abies nebrodensis*. Theor Appl Genet, 90: 1012–1018.

Wagner S, Gerber S, Petit RJ (2012) Two highly informative dinucleotide SSR multiplexes for the conifer Larix decidua (European larch). Molecular Ecology Resources, 12: 717–725.

Wang KS (2004). Gene flow in European beech (*Fagus sylvatica* L.). Genetica ,122: 105-113.

Wang J, Ye Q, Kang M, Huang H (2008). Novel polymorphic microsatellite loci and patterns of pollen-mediated gene flow in an ex situ population of *Eurycorymbus cavaleriei* (Sapindaceae) as revealed by categorical paternity analysis. Conservation Genetics, 9: 559-567.

Wang J, Kang M, Gao P, Huang H (2010a). Contemporary pollen flow and mating patterns of a subtropical canopy tree *Eurycorymbus cavaleriei* in a fragmented agricultural landscape. Forest Ecology and Management, 260: 2180-2188.

Wang H, Sork VL, Wu J, Ge J (2010b). Effect of patch size and isolation on mating patterns and seed production in an urban population of Chinese pine (*Pinus tabulaeformis* Carr.). Forest Ecology and Management, 260: 965-974.

White GM, Boshier DH, Powell W (2002). Increased pollen flow counteracts fragmentation in a tropical dry forest, an example from *Swietenia humilis* Zuccarini. Proceedings of the National Academy of Sciences, 99: 2038–2042.

Williams CG (2005). Framing the issues on transgenic forests. Nature Biotechnology, 23: 530-532.

Williams CG (2010). Long-distance pine pollen still germinates after meso-scale dispersal. American Journal of Botany, 97: 846-855.

Williams CG (2013). Forest tree pollen dispersal via the water cycle. American Journal of Botany 100: 1184-1190.

Willson MF (1993). Dispersal mode, seed shadows, and colonization patterns. Vegetatio, 107/108: 261-280.

Wolf H (2003) EUFORGEN Technical Guidelines for genetic conservation and use for silver fir (*Abies alba*). International Plant Genetic Resources Institute, Rome, Italy.

Xie CY, Knowles P (1994). Mating system and effective pollen immigration in a Norway spruce (*Picea abies* (L.) Karst) plantation. Silvae Genetica, 43: 48-52.

Yeaman, S., and A. Jarvis. 2006. Regional heterogeneity and gene flow maintain variance in a quantitative trait within populations of lodgepole pine. Proceedings of the Royal Society B-Biological Sciences, 273: 1587–1593.

Yehili JL, N'Guetta Assanvo S-P, Gnagne M, Blanc G, Rodier-Goud M, Clément-Demange A, Sequin M, Fanjavola M (2007). Flux de gènes dans un verger à graines d'hévéas sauvages (*Hevea brasiliensis* Müll. Arg.). Cahiers Agricultures, 16: 177-184.

Silvertown, J. (1991) Dorothy's dilemma and the unification of plant population biology. Trends in Ecology and Evolution, 6: 346-348.

Xu F, Feng S, Wu R, Du F (2013) Two highly validated SSR multiplexes (8-plex) for Euphrates' poplar,

Populus euphratica (Salicaceae). Molecular Ecology Resources, 13: 144–53.

Young A, Boyle T, Brown T (1996) The population genetic consequences of habitat fragmentation for plants. TREE, 11:413-418.

Zalapa J, Cuevas H, Zhu H, et al. (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. American Journal of Botany, 99: 193–208.

Zeng S, Xiao G, Guo J et al. (2010) Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. BMC Genomics, 11, 94.

Zhang JJ, Ye QG, Yao XH, Huang HW (2010). Spontaneous interspecific hybridization and patterns of pollen dispersal in ex situ populations of a tree species (*Sinojackia xylocarpa*) that is extinct in the wild. Conservation Biology, 24: 246-255.

Ziegenhagen B, Fady B, Kuhlenkamp V, Liepelt S (2005) Differentiating groups of Abies species with a simple molecular marker. Silvae Genetica, 54: 123–126.

Appendix 1

List of 187 data points obtained from the 92 paternity analysis papers collected. For each data point are reported: the botanical group and family, the breeding system and the primary pollination vector of the studied species; the number of male and female individuals in the stand (in monoecious species the number of male individuals is equal to the number of female individuals), the number of pollen traps (referred to as 'mother trees' in the text), the total number of sampled seeds, the stand area and stand density. In addition, for each paper the method and the molecular markers used for paternity assignment are reported. NA: no information available.

Species	Group	Family	Breeding system[1]	Pollination vector	N males	N females	N traps	N seeds	Area	Density[2]	Method[3]	Markers[4]	Reference
Pseudotsuga menziesii	Gymnosperms	Pinaceae	monoecious	wind	84	84	6	547	2.4	35	exclusion, NM	allozymes	Adams (1992)
					36	36	6	574	2.4	15			
Picea abies	Gymnosperms	Pinaceae	monoecious	wind	33	33	26	1920	1	27	NM	allozymes	Xie & Knowles (1994)
Cordia alliodora	Angiosperms	Boraginaceae	monoecious	insect	123	123	19	380	5.9	20.9	exclusion	allozymes	Boshier et al. (1995)
Gleditsia triacanthos	Angiosperms	Fabaceae	dioecious	insect	61	42	10	295	3.2	32.19	exclusion	allozymes	Schnabel & Hamrick (1995)[a]
					61	42	10	669	3.2	32.19			
					124	70	10	1356	4.2	46.19			
					124	70	12	208	4.2	46.19			
Pinus attenuata	Gymnosperms	Pinaceae	monoecious	wind	65	65	4	880	0.04	1128	NM	allozymes	Burczyk et al. (1996)
Pithecellobium elegans	Angiosperms	Fabaceae	monoecious	insect	28	28	6	167	16	1	exclusion	SSRs	Chase et al. (1996)
Calophyllum longifolium	Angiosperms	Clusiaceae	mixed	insect	31	31	11	352	84	0.33	exclusion	allozymes	Stacy et al. (1996)[a]
					29	29	11	616	84	0.33			
Spondias mombin	Angiosperms	Anacardiaceae	monoecious	insect	19	19	10	576	84	0.17			
					19	19	11	430	84	0.17			
Turpinia occidentalis	Angiosperms	Staphyleaceae	monoecious	insect	30	30	6	172	50	0.45			
Quercus macrocarpa	Angiosperms	Fagaceae	monoecious	wind	62	62	3	282	5	12.4	exclusion	SSRs	Dow & Ashley (1998)
Cecropia obtusifolia	Angiosperms	Cecropiaceae	dioecious	wind	47	41	41	1230	8.64	5.4	exclusion	allozymes	Kaufman et al. (1998)
Quercus petraea	Angiosperms	Fagaceae	monoecious	wind	124	124	7	537	5.76	21.5	exclusion	SSRs	Streiff et al. (1999)
Quercus robur	Angiosperms	Fagaceae	monoecious	wind	167	167	6	447	5.76	29.3			
Rhododendron metternichiii	Angiosperms	Ericaceae	monoecious	insect	173	173	4	216	1.05	17.1	Cervus	SSRs	Kameyama et al. (2000)
Neobalanocarpus heimii	Angiosperms	Dipterocarpaceae	monoecious	insect	30	30	5	348	42	0.75	Cervus	SSRs	Konuma et al. (2000)

Pinus flebili	Gymnosperms	Pinaceae	monoecious	wind	397	397	71	518	15	34.5	exclusion	allozymes	Schuster & Mitton (2000)
Enterolobium cyclocarpum	Angiosperms	Fabaceae	monoecious	animal, insect	11	11	5	783	8.9	1.23	exclusion	allozymes	Apsit et al. (2001)[a]
					11	11	5	721	8.9	1.23			
					11	11	2	193	8.9	1.23			
Quercus robur	Angiosperms	Fagaceae	monoecious	wind	57	57	3	180	4.5	57.8	exclusion	SSRs	Buiteveld et al. (2001)
Dinizia excelsa	Angiosperms	Fabaceae	monoecious	insect	36	36	11	333	NA	NA	Cervus	SSRs	Dick (2001)
Pinus densiflora	Gymnosperms	Pinaceae	monoecious	wind	154	154	1	874	9.1	16.9	exclusion	SSRs	Lian et al. (2001)
Eucalyptus regnans	Angiosperms	Myrtaceae	monoecious	animal, insect	285	285	30	1761	0.5	570	NM	allozymes	Burczyk et al. (2002)
Swietenia humilis	Angiosperms	Meliaceae	dioecious	insect	97	NA	5	150	NA	NA	exclusion	SSRs	White et al. (2002)
					44	NA	12	360	NA	NA			
					22	NA	17	510	NA	NA			
					97	NA	5	150	68	NA			
					74	NA	12	360	NA	NA			
Eucalyptus grandis	Angiosperms	Myrtaceae	monoecious	insect	349	349	30	724	0.6	580	FaMoz	SSRs	Chaix et al. (2003)
Populus spp.	Angiosperms	Salicaceae	dioecious	wind	12	9	3	103	0.2	105	exclusion	SSRs	Tabbener et al. (2003)
Picea abies	Gymnosperms	Pinaceae	monoecious	wind	557	557	10	2000	0.89	625	NM	allozymes	Burczyk et al. (2004)
Magnolia obovata	Angiosperms	Magnoliaceae	monoecious	insect	83	83	3	322	69	1.2	NA	SSRs	Isagi et al. (2004)
Dipterocarpus tempehes	Angiosperms	Dipterocarpaceae	monoecious	insect	277	277	3	147	70	3.95	Cervus	SSRs	Kenta et al. (2004)[a]
					277	277	3	188	70	3.95			
Dicorynia guianensis	Angiosperms	Fabaceae	monoecious	insect	157	157	22	246	40	3.9	FaMoz	SSRs	Latouche-Hallé et al. (2004)
Quercus salicina	Angiosperms	Fagaceae	monoecious	wind	111	111	8	276	11.56	13.5	Cervus	SSRs	Nakanishi et al. (2004)
Fagus sylvatica	Angiosperms	Fagaceae	monoecious	wind	24	24	24	511	0.35	68.6	Cervus	allozymes	Wang et al. (2004)
					70	70	70	844	0.78	90			
					99	99	99	1954	1.92	51.50			
Fraxinus excelsior	Angiosperms	Oleaeceae	mixed	wind	146	146	19	422	900	0.16	NM	SSRs	Bacles et al. (2005, 2008)
Albizia lebbeck	Angiosperms	Fabaceae	monoecious	insect	8	8	2	195	NA	NA	GFLOW	allozymes	Dunphy et al.(2005)
					3	3	1	81	NA	NA			
					11	11	4	210	NA	NA			

					2	2	1	142	NA	NA			
Prunus mahaleb	Angiosperms	Rosaeceae	monoecious	insect	196	196	20	200	26	7.5	Cervus	SSRs	Garcia et al. (2005)
Cryptomeria japonica	Gymnosperms	Cupressaceae	monoecious	insect	62	62	12	360	0.31	200	exclusion	SSRs	Moriguchi et al. (2005)
					35	35	12	360	0.95	36.8			
					26	26	12	360	1.09	23.8			
					54	54	12	360	0.1	540			
					24	24	12	360	0.11	218			
Quercus salicina	Angiosperms	Fagaceae	monoecious	wind	111	111	6	796	11.56	9.6	exclusion, Cervus	SSRs	Nakanishi et al. (2005)
Sorbus torminalis	Angiosperms	Rosaceae	monoecious	insect	185	185	14	653	472	0.36	Cervus, NM, Patri	SSRs	Oddou-Muratorio et al. (2003, 2005)[a]
					185	185	60	1016	472	0.36			
Pinus sylvestris	Gymnosperms	Pinaceae	monoecious	wind	35	35	34	813	20	1.8	exclusion	SSRs	Robledo-Arnuncio & Gil (2005)
Pseudotsuga menziesii	Gymnosperms	Pinaceae	monoecious	wind	342	342	24	240	2.1	162.8	exclusion	SSRs	Slavov et al. (2005)[a]
					342	342	24	336	2.1	162.8			
Fraxinus mandshurica	Angiosperms	Fagaceae	dioecious	wind	76	74	4	200	10.5	7.28	NM	SSRs	Goto et al. (2006)
Abies nordmanniana	Gymnosperms	Pinaceae	monoecious	wind	353	353	24	232	NA	NA	Cervus	SSRs	Hansen et al. (2006)
Shorea lumutensis	Angiosperms	Dipterocarpaceae	monoecious	insect	47	47	4	182	8	4.4	Cervus	SSRs	Lee et al. (2006)
Taxus baccata	Gymnosperms	Taxodiaceae	dioecious	wind	10	15	7	279	2.1	12	exclusion	allozymes	Leinemann & Hattemer (2006)
Populus nigra	Angiosperms	Salicaceae	dioecious	wind	14	42	4	155	0.4	140	Cervus	SSRs	Vanden Broeck et al. (2006)
Juglans mandshurica	Angiosperms	Juglandaceae	monoecious	wind	73	73	6	221	0.96	76	Cervus	SSRs	Bai et al. (2007)[b]
					73	73	5	238	0.96	76			
Araucaria angustifolia	Gymnosperms	Araucariaceae	dioecious	wind	124	104	10	210	5.4	42.2	Cervus	SSRs	Bittencourt & Sebben (2007)
					9	2	1	20	NA	NA			
Calothamnus quadrifidus	Angiosperms	Myrtaceae	monoecious	bird	22	22	9	177	NA	0.002	Cervus	SSRs	Byrne et al. (2007)
					23	23	16	318	NA	0.018			
Bursera simaruba	Angiosperms	Burseraceae	mixed	insect	9	9	5	247	NA	NA	GFLOW	allozymes	Dunphy & Hamrick (2007)
					7	7	3	124	NA	NA			
					6	6	2	222	NA	NA			
					6	6	1	19	NA	NA			
					3	3	1	17	NA	NA			

Shorea leprosula	Angiosperms	Dipterocarpaceae	monoecious	insect	55	55	19	647	100	0.55	Cervus	SSRs	Fukue et al. (2007)
Fagus crenata	Angiosperms	Fagaceae	monoecious	wind	32	32	2	162	1.5	21.3	Cervus	SSRs	Hanaoka et al. (2007)
Sorbus torminalis	Angiosperms	Rosaceae	monoecious	insect	123	123	20	824	20	6.1	Cervus	SSRs	Hoebee et al. (2007)
Entandrophragma cylindricum	Angiosperms	Meliaceae	monoecious	insect	152	152	16	269	100	1.52	FaMoz	SSRs	Lourmas et al. (2007)
					113	113	20	358	100	1.13			
					123	123	15	334	440	0.28			
Cryptomeria japonica	Gymnosperms	Cupressaceae	monoecious	insect	1144	1144	9	900	0.1	11.55	NA	SSRs	Moriguchi et al. (2007)
Picea glauca	Gymnosperms	Pinaceae	monoecious	wind	32	32	32	2967	NA	4	NM	allozymes	O'Connel et al. (2007)
Magnolia stellata	Angiosperms	Magnoliaceae	monoecious	insect	84	84	9	483	NA	NA	Cervus	SSRs	Setsuko et al. (2007)
Hevea brasiliensis	Angiosperms	Euphorbiaceae	monoecious	insect	287	287	25	388	0.89	322	Cervus, FaMoz	SSRs	Yehili et al. (2007)[a]
					287	287	9	346	0.89	322			
Araucaria angustifolia	Gymnosperms	Araucariaceae	dioecious	wind	52	56	10	190	14	7.71	Cervus	SSRs	Bittencourt & Sebben (2008)
Eucalyptus wandoo	Angiosperms	Myrtaceae	monoecious	insect	46	46	12	240	NA	0.05	Cervus	SSRs	Byrne et al. (2008)
					40	40	11	220	NA	0.009			
Hymenaea courbaril	Angiosperms	Fabaceae	monoecious	bat	130	130	20	367	546	0.238	Cervus	SSRs	de Lacerda et al. (2008)
Larix occidentalis	Gymnosperms	Pinaceae	monoecious	wind	41	41	14	551	NA	NA	Cervus	SSRs	Funda et al. (2008)
Kandelia candel	Angiosperms	Rhizophoraceae	monoecious	insect	2062	2062	11	378	0.55	3749	Cervus	SSRs	Geng et al. (2008)
Dipteryx panamensis	Angiosperms	Fabaceae	monoecious	insect	104	104	11	50	65	0.8	FaMoz	SSRs	Hanson et al. (2008)
					50	50	22	107	52	0.21			
					52	52	25	124	40	0.58			
					12	12	9	44	NA	0.19			
Eucalyptus grandis	Angiosperms	Myrtaceae	monoecious	animal, insect	192	192	6	282	4	45.5	Cervus	SSRs	Jones et al. (2008)[a]
					192	192	2	94	4	45.5			
Acacia saligna	Angiosperms	Fabaceae	monoecious	insect	107	107	10	186	0.55	194.5	Cervus	allozymes	Millar et al. (2008)
Shorea acuminata	Angiosperms	Dipterocarpaceae	monoecious	insect	58	58	11	688	40	1.45	Cervus	SSRs	Naito et al. (2008)
Prunus cerasoides	Angiosperms	Rosaeceae	monoecious	insect	16	16	5	100	NA	NA	Cervus	SSRs	Pakkad et al. (2008a)
					23	23	3	53	NA	NA			
					45	45	8	136	NA	NA			

Quercus semiserrata	Angiosperms	Fagaceae	monoecious	wind	26	26	8	174	10.8	2.4	Cervus	SSRs	Pakkad et al. (2008b)
					26	26	8	261	10.8	2.4			
Bagassa guianensis	Angiosperms	Moraceae	dioecious	insect	38	33	18	490	500	0.14	Cervus	SSRs	Silva et al. (2008)
Eurycorymbus cavaleriei	Angiosperms	Sapindaceae	dioecious	insect	19	14	8	240	33	0.57	Cervus	SSRs	Wang et al. (2008)
Symphonia globulifera	Angiosperms	Clusiaceae	monoecious	animal	161	161	56	748	500	0.33	Cervus	SSRs	Carneiro et al. (2009)
Prunus avium	Angiosperms	Rosaeceae	monoecious	insect	978	978	10	419	34	28.8	Cervus, FaMoz	SSRs	Cottrell et al (2009)
Quercus spp	Angiosperms	Fagaceae	monoecious	wind	296	296	8	320	NA	NA	FaMoz	SSRs	Curtu et al. (2009)
Castanea crenata	Angiosperms	Fagaceae	monoecious	insect	278	278	3	304	6	46.3	Cervus	SSRs	Hasegawa et al. (2009)
Platypodium elegans	Angiosperms	Fabaceae	monoecious	insect	68	68	5	500	50	0.2	exclusion, Cervus	SSRs	Hufford et al. (2009)
Sorbus domestica	Angiosperms	Rosaceae	monoecious	insect	189	189	49	1183	10	19	Cervus	SSRs	Kamm et al. (2009)
Malus sylvestris	Angiosperms	Rosaceae	monoecious	insect	50	50	12	180	12.69	20	Cervus	SSRs	Larsen & Kjaer (2009)
Quercus lobata	Angiosperms	Fagaceae	monoecious	wind	92	92	5	840	58.9	1.7	Cervus	SSRs	Pluess et al. (2009)
Juglans spp	Angiosperms	Juglandaceae	monoecious	wind	139	139	8	461	NA	NA	Cervus	SSRs	Pollegioni et al. (2009)
Quercus spp	Angiosperms	Fagaceae	monoecious	wind	295	295	30	855	6	49.17	Cervus	SSRs	Salvini et al. (2009)
Populus trichocarpa	Angiosperms	Salicaceae	dioecious	wind	419	NA	7	240	19.6	NA	Cervus	SSRs	Slavov et al. (2009)
					223	172	32	681	31400	NA			
Shorea leprosula	Angiosperms	Dipterocarpaceae	monoecious	insect	61	61	5	129	40	1.5	Cervus, NM	SSRs	Tani et al. (2009)[a]
					61	61	8	444	40	1.5			
					70	70	4	106	40	1.75			
					70	70	5	216	40	1.75			
Quercus macrocarpa	Angiosperms	Fagaceae	monoecious	wind	26	26	10	225	1	26	Cervus	SSRs	Craft & Ashley (2010)
					62	62	9	215	1	62			
					115	115	13	347	1	115			
Myracrodruon urendeuva	Angiosperms	Anacardiaceae	dioecious	wind, insect	467	467	29	414	436	1.07	Cervus	SSRs	Gaino et al. (2010)
Eucalyptus nitens	Angiosperms	Myrtaceae	monoecious	insect	50	50	10	473	0.33	151.5	Cervus	SSRs	Grosser et al. (2010)
Gomortega keule	Angiosperms	Gomortegaceae	monoecious	insect	176	176	31	196	NA	NA	Cervus	SSRs	Lander et al. (2010)
					92	92	38	218	NA	NA			
					32	32	16	96	NA	NA			
					30	30	1	1	NA	NA			

					22	22	7	30	NA	NA			
					22	22	21	189	NA	NA			
					20	20	7	69	NA	NA			
					17	17	8	65	NA	NA			
					16	16	5	39	NA	NA			
					11	11	1	1	NA	NA			
					10	10	1	10	NA	NA			
					7	7	2	10	NA	NA			
					4	4	1	2	NA	NA			
Populus nigra	Angiosperms	Salicaceae	dioecious	wind	267	244	7	625	1	511	Cervus	SSRs	Rathmacher et al. (2010)[a]
					267	244	6	2264	1	511			
Eurycorymbus cavaleriei	Angiosperms	Sapindaceae	dioecious	insect	50	52	13	239	6.85	14.9	Cervus	SSRs	Wang et al. (2010a)
					96	98	17	277	16.7	11.6			
Pinus tabulaeformis	Gymnosperms	Pinaceae	monoecious	wind	1	1	1	24	NA	NA	exclusion	Cp SSRs	Wang et al. (2010b)
					1	1	1	28	NA	NA			
					1	1	1	28	NA	NA			
					1	1	1	27	NA	NA			
					1	1	1	30	NA	NA			
					2	2	1	24	NA	NA			
					2	2	1	25	NA	NA			
					2	2	1	29	NA	NA			
					2	2	1	20	NA	NA			
					3	3	1	27	NA	NA			
					3	3	1	24	NA	NA			
					3	3	1	22	NA	NA			
					3	3	1	29	NA	NA			
					3	3	1	26	NA	NA			
					3	3	1	35	NA	NA			
					4	4	1	26	NA	NA			
					4	4	1	29	NA	NA			
					4	4	1	26	NA	NA			
					4	4	1	30	NA	NA			
					6	6	1	30	NA	NA			

					8	8	1	27	NA	NA			
					9	9	1	24	NA	NA			
					14	14	1	29	NA	NA			
					20	20	1	23	NA	NA			
					39	39	2	59	NA	NA			
Sinojackia spp	Angiosperms	Styracaceae	monoecious	insect	64	64	8	249	NA	NA	Cervus	SSRs	Zhang et al. (2010)
Tabebuia aurea	Angiosperms	Bignoniaceae	monoecious	insect	260	260	21	309	40	6.5	Cervus	SSRs	Braga et al. (2011)[a]
					260	260	21	328	40	6.5			
Quercus robur	Angiosperms	Fagaceae	monoecious	wind	27	27	2	39	NA	NA	exclusion, Cervus	SSRs	Buschbom et al. (2011)
Hymenaea stigonocarpa	Angiosperms	Fabaceae	monoecious	bat	6	6	2	34	3.62	0.0094	Cervus	SSRs	de Moraes & Sebben (2011)
					28	28	12	137	3.62	0.0094			
Guaiacum sanctum	Angiosperms	Zygophyllaceae	monoecious	insect	35	35	6	108	50	0.7	FaMoz	allozymes	Fuchs & Hamrick (2011)
					21	21	21	378	0.8	26.25			
Prunus avium	Angiosperms	Rosaeceae	monoecious	insect	39	39	10	400	6	6.5	self-made maximum likelihood	SSRs	Gregorius et al. (2011)
Fagus sylvatica	Angiosperms	Fagaceae	monoecious	wind	192	192	9	287	3.36	57	FaMoz, NM	SSRs	Piotti et al. (2011)
					235	235	9	275	1.44	163			
					286	286	4	187	1.91	150			
					90	90	5	249	1.32	68			
Theobroma cacao	Angiosperms	Malvaceae	monoecious	insect	156	156	9	450	0.56	278	Cervus	SSRs	Silva et al. (2011)

[1]mixed: polygamodioecious species (basically dioecious, but also having some bisexual flowers present in some or all plants); [2]When density was not available in the text it was estimated dividing the number of individuals within the study population for the stand area; [3]NM: neighbourhood model; [4]SSRs: short sequence repeats (aka microsatellites);
[a]Multiyear studies; [b]Protandrous and protogynous mother trees were analyzed separately.

Appendix 2

Table 1: Genetic diversity parameters of the 24 populations averaged on the 6 geographical groups. *A* (mean number of alleles), *Ar* (allelic richness), H_O (observed heterozygosity), H_E (expected heterozygosity), F_{IS} (fixation index), *P* (overall number of private alleles for geographic group).

Geographic group	*A*	Ar_{42}	H_O	H_E	F_{IS}	*P*
Northern Apennine	4.24 (0.71)	4.18	0.43 (0.06)	0.44 (0.06)	0.014 (0.028)	3
Central Apennine	4.09 (0.72)	4.01	0.42 (0.07)	0.42 (0.06)	0.007 (0.033)	1
Southern Apennine	5.21 (1.20)	5.05	0.46 (0.06)	0.46 (0.06)	0.015 (0.032)	8
Eastern Alps	4.26 (0.74)	4.18	0.42 (0.06)	0.43 (0.06)	0.028 (0.051)	1
Western Alps	4.10 (0.71)	3.89	0.41(0.06)	0.42 (0.06)	0.041 (0.054)	0
Balkans	5.17 (1.21)	4.85	0.41(0.07)	0.42 (0.07)	0.003 (0.026)	8

Table 2: Likelihood and ΔK values for each of the tested k, averaged over 10 runs for each k value.

k	Reps	Mean LnP(K)	sd LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
1	10	-40136.1900	0.0876	NA	NA	NA
2	10	-37854.5600	1.4849	2281.630000	1603.280000	1079.732119
3	10	-37176.2100	3.4987	678.350000	87.730000	25.074927
4	10	-36585.5900	6.6929	590.620000	107.660000	16.085806
5	10	-36102.6300	3.7208	482.960000	368.080000	98.924237
6	10	-35987.7500	65.5401	114.880000	77.040000	1.175463
7	10	-35795.8300	97.3369	191.920000	131.060000	1.346458
8	10	-35734.9700	43.8738	60.860000	37.800000	0.861563
9	10	-35636.3100	68.6198	98.660000	65.210000	0.950308
10	10	-35602.8600	72.7006	33.450000	84.290000	1.159413
11	10	-35485.1200	38.2673	117.740000	NA	NA

Table 3: Bootstrap support for interior branches of UPMGA tree based on 1000 bootstrapped trees in TreeFit (Kalinowski 2009).

Length	B.S.	Populations separated by the branch
0.002	0.6	(abe, ner) <---> (noa, tar, pes, sal, toc, abs, cor, cpl, tos, vcl, btr, cer, pig, cil, gam, lau, sil, ssb, tdp, blg, rom, tara)
0.0012	0.48	(sil, ssb) <---> (noa, tar, pes, sal, toc, abs, cor, cpl, tos, vcl, abe, btr, cer, ner, pig, cil, gam, lau, tdp, blg, rom, tara)
0.0029	0.67	(noa, pig) <---> (tar, pes, sal, toc, abs, cor, cpl, tos, vcl, abe, btr, cer, ner, cil, gam, lau, sil, ssb, tdp, blg, rom, tara)
0.0018	0.55	(abe, cer, ner) <---> (noa, tar, pes, sal, toc, abs, cor, cpl, tos, vcl, btr, pig, cil, gam, lau, sil, ssb, tdp, blg, rom, tara)
0.0029	0.5	(tos, vcl) <---> (noa, tar, pes, sal, toc, abs, cor, cpl, abe, btr, cer, ner, pig, cil, gam, lau, sil, ssb, tdp, blg, rom, tara)
0.0041	0.7	(sil, ssb, tdp) <---> (noa, tar, pes, sal, toc, abs, cor, cpl, tos, vcl, abe, btr, cer, ner, pig, cil, gam, lau, blg, rom, tara)
0.0037	0.68	(cpl, tos, vcl) <---> (noa, tar, pes, sal, toc, abs, cor, abe, btr, cer, ner, pig, cil, gam, lau, sil, ssb, tdp, blg, rom, tara)
0.0045	0.82	(noa, abe, cer, ner, pig) <---> (tar, pes, sal, toc, abs, cor, cpl, tos, vcl, btr, cil, gam, lau, sil, ssb, tdp, blg, rom, tara)
0.0038	0.78	(gam, sil, ssb, tdp) <---> (noa, tar, pes, sal, toc, abs, cor, cpl, tos, vcl, abe, btr, cer, ner, pig, cil, lau, blg, rom, tara)
0.0046	0.78	(noa, abe, btr, cer, ner, pig) <---> (tar, pes, sal, toc, abs, cor, cpl, tos, vcl, cil, gam, lau, sil, ssb, tdp, blg, rom, tara)
0.0089	1	(rom, tara) <---> (noa, tar, pes, sal, toc, abs, cor, cpl, tos, vcl, abe, btr, cer, ner, pig, cil, gam, lau, sil, ssb, tdp, blg)
0.0047	0.69	(pes, sal) <---> (noa, tar, toc, abs, cor, cpl, tos, vcl, abe, btr, cer, ner, pig, cil, gam, lau, sil, ssb, tdp, blg, rom, tara)
0.0054	0.85	(cil, gam, sil, ssb, tdp) <---> (noa, tar, pes, sal, toc, abs, cor, cpl, tos, vcl, abe, btr, cer, ner, pig, lau, blg, rom, tara)
0.0134	0.98	(cor, cpl, tos, vcl) <---> (noa, tar, pes, sal, toc, abs, abe, btr, cer, ner, pig, cil, gam, lau, sil, ssb, tdp, blg, rom, tara)
0.0082	0.87	(noa, tar, abe, btr, cer, ner, pig) <---> (pes, sal, toc, abs, cor, cpl, tos, vcl, cil, gam, lau, sil, ssb, tdp, blg, rom, tara)
0.0071	0.86	(cil, gam, lau, sil, ssb, tdp) <---> (noa, tar, pes, sal, toc, abs, cor, cpl, tos, vcl, abe, btr, cer, ner, pig, blg, rom, tara)
0.0129	0.89	(pes, sal, toc) <---> (noa, tar, abs, cor, cpl, tos, vcl, abe, btr, cer, ner, pig, cil, gam, lau, sil, ssb, tdp, blg, rom, tara)
0.0071	0.65	(noa, tar, cor, cpl, tos, vcl, abe, btr, cer, ner, pig) <---> (pes, sal, toc, abs, cil, gam, lau, sil, ssb, tdp, blg, rom, tara)
0.0274	1	(blg, rom, tara) <---> (noa, tar, pes, sal, toc, abs, cor, cpl, tos, vcl, abe, btr, cer, ner, pig, cil, gam, lau, sil, ssb, tdp)
0.0194	0.97	(abs, cil, gam, lau, sil, ssb, tdp) <---> (noa, tar, pes, sal, toc, cor, cpl, tos, vcl, abe, btr, cer, ner, pig, blg, rom, tara)
0.0357	0.98	(abs, cil, gam, lau, sil, ssb, tdp, blg, rom, tara) <---> (noa, tar, pes, sal, toc, cor, cpl, tos, vcl, abe, btr, cer, ner, pig)

Table 4: Pairwise F_{ST} (below the diagonal) and R_{ST} (above the diagonal) values

		R_{ST}																							
		NOA	TAR	PES	SAL	TOC	ABS	COR	CPL	TOS	VCL	ABE	BTR	CER	NER	PIG	CIL	GAM	LAU	SIL	SSB	TNP	BLG	ROM	SER
F_{ST}	NOA	0.000	0.023	0.132	0.106	0.023	0.105	0.046	-0.001	0.019	0.040	0.022	0.004	0.105	0.067	-0.005	0.076	0.061	0.025	0.035	0.077	0.060	0.097	0.081	0.050
	TAR	0.035	0.000	0.187	0.118	0.042	0.088	0.079	0.017	0.050	0.059	0.069	0.056	0.163	0.127	0.026	0.075	0.082	0.071	0.098	0.103	0.073	0.073	0.057	0.036
	PES	0.069	0.074	0.000	0.043	0.084	0.334	0.074	0.100	0.066	0.062	0.046	0.098	0.026	0.013	0.105	0.285	0.232	0.134	0.134	0.270	0.262	0.309	0.281	0.184
	SAL	0.105	0.091	0.029	0.000	0.041	0.268	0.048	0.075	0.050	0.027	0.051	0.102	0.047	0.047	0.075	0.229	0.183	0.103	0.128	0.229	0.218	0.256	0.224	0.128
	TOC	0.105	0.123	0.050	0.059	0.000	0.176	0.030	0.016	0.026	0.024	0.022	0.034	0.089	0.056	0.016	0.145	0.117	0.062	0.080	0.148	0.126	0.155	0.137	0.081
	ABS	0.116	0.141	0.155	0.221	0.199	0.000	0.188	0.084	0.124	0.189	0.194	0.153	0.275	0.252	0.106	-0.004	0.019	0.084	0.095	0.019	0.002	0.051	0.027	0.066
	COR	0.067	0.071	0.083	0.094	0.111	0.176	0.000	0.027	0.014	0.015	0.018	0.036	0.057	0.047	0.028	0.154	0.113	0.048	0.066	0.146	0.132	0.178	0.152	0.087
	CPL	0.043	0.062	0.090	0.109	0.129	0.154	0.035	0.000	0.009	0.021	0.019	0.012	0.088	0.060	-0.001	0.066	0.055	0.020	0.037	0.071	0.049	0.068	0.050	0.022
	TOS	0.035	0.043	0.067	0.101	0.121	0.111	0.044	0.040	0.000	0.001	0.002	0.011	0.040	0.035	0.008	0.095	0.066	0.018	0.028	0.093	0.080	0.134	0.087	0.032
	VCL	0.052	0.044	0.083	0.094	0.127	0.161	0.035	0.032	0.020	0.000	0.005	0.033	0.046	0.035	0.022	0.153	0.116	0.048	0.070	0.155	0.138	0.181	0.136	0.055
	ABE	0.023	0.025	0.052	0.092	0.107	0.101	0.050	0.046	0.020	0.037	0.000	0.005	0.034	0.012	0.013	0.152	0.114	0.043	0.049	0.144	0.133	0.184	0.150	0.080
	BTR	0.023	0.039	0.073	0.122	0.129	0.132	0.058	0.056	0.030	0.046	0.022	0.000	0.072	0.043	0.002	0.117	0.087	0.036	0.035	0.111	0.094	0.148	0.120	0.073
	CER	0.030	0.055	0.083	0.121	0.134	0.104	0.091	0.067	0.045	0.073	0.019	0.048	0.000	0.007	0.073	0.228	0.168	0.083	0.078	0.210	0.211	0.281	0.244	0.150
	NER	0.021	0.031	0.038	0.073	0.091	0.099	0.056	0.045	0.025	0.039	0.009	0.031	0.019	0.000	0.051	0.207	0.158	0.073	0.070	0.189	0.188	0.238	0.216	0.135
	PIG	0.012	0.028	0.044	0.082	0.092	0.095	0.050	0.044	0.020	0.037	0.006	0.014	0.021	0.005	0.000	0.080	0.056	0.018	0.028	0.083	0.064	0.100	0.081	0.040
	CIL	0.085	0.116	0.137	0.197	0.182	0.028	0.139	0.116	0.078	0.120	0.072	0.105	0.073	0.073	0.067	0.000	0.005	0.052	0.062	0.003	-0.001	0.050	0.026	0.052
	GAM	0.097	0.135	0.138	0.201	0.182	0.044	0.157	0.148	0.099	0.144	0.083	0.118	0.077	0.076	0.070	0.018	0.000	0.022	0.024	0.002	0.007	0.078	0.050	0.052
	LAU	0.084	0.113	0.121	0.169	0.166	0.044	0.118	0.101	0.079	0.112	0.077	0.104	0.080	0.070	0.068	0.022	0.036	0.000	0.000	0.037	0.045	0.121	0.085	0.047
	SIL	0.077	0.108	0.115	0.180	0.161	0.036	0.136	0.116	0.075	0.121	0.066	0.094	0.063	0.060	0.055	0.014	0.011	0.028	0.000	0.038	0.048	0.134	0.104	0.071
	SSB	0.079	0.111	0.132	0.192	0.186	0.043	0.141	0.116	0.083	0.127	0.062	0.107	0.057	0.062	0.062	0.011	0.013	0.030	0.007	0.000	0.006	0.090	0.064	0.081
	TNP	0.076	0.095	0.103	0.162	0.152	0.034	0.114	0.095	0.051	0.093	0.052	0.084	0.059	0.051	0.048	0.012	0.023	0.022	0.006	0.015	0.000	0.043	0.022	0.047
	BLG	0.099	0.124	0.137	0.194	0.179	0.103	0.139	0.116	0.070	0.099	0.081	0.104	0.081	0.072	0.074	0.056	0.067	0.070	0.056	0.070	0.038	0.000	0.025	0.064
	ROM	0.130	0.137	0.151	0.203	0.209	0.137	0.159	0.153	0.075	0.113	0.110	0.131	0.129	0.102	0.096	0.097	0.115	0.117	0.096	0.115	0.066	0.050	0.000	0.014
	SER	0.117	0.129	0.148	0.193	0.205	0.134	0.143	0.126	0.068	0.093	0.101	0.120	0.108	0.095	0.088	0.092	0.109	0.105	0.091	0.109	0.065	0.035	0.019	0.000

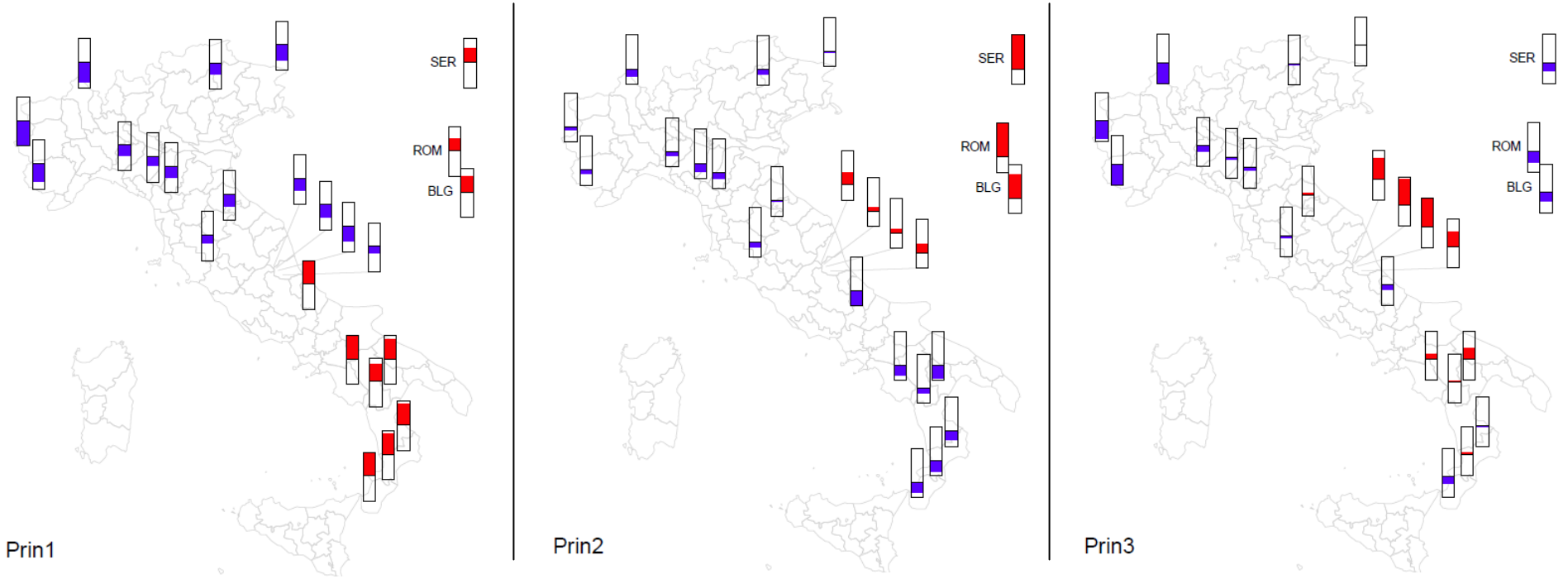

Figure 1: Scores of the 3 principal components for each population. Each colored barplot is proportional to the score. Red barplots indicate positive values, blue barplots indicate negative values

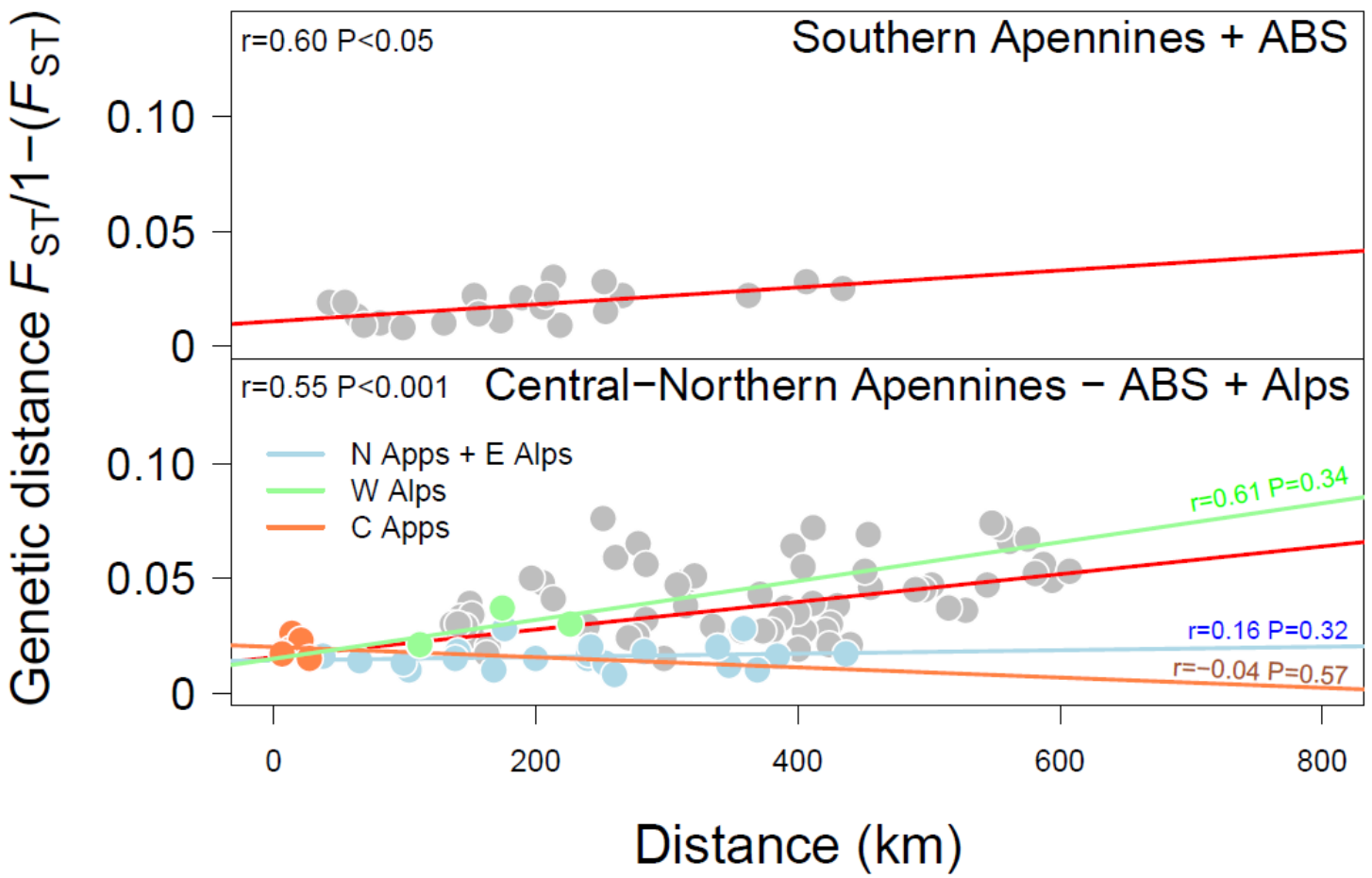

Figure 2: Relationship between geographic and genetic distance in different groups of population, tested using Mantel test.