iForest
Biogeosciences and Forestry

Research Article
doi: 10.3832/ifor3913-015
vol. 15, pp. 141-147

Chloroplast DNA barcoding genes *matK* and *psbA-trnH* are not suitable for species identification and phylogenetic analyses in closely related pines

Sanna Olsson [(1)],
Guia Giovannelli [(2)],
Anne Roig [(2)],
Ilaria Spanu [(3)],
Giovanni Giuseppe Vendramin [(3)],
Bruno Fady [(2)]

The largest and most economically important conifer genus *Pinus* is widespread in the northern hemisphere. Comprehensive phylogenies relying on complete chloroplast gene sequences are now available for the entire genus. However, phylogenetic relationships remain unresolved for certain lineages. One such example, which is also inconsistent in terms of biogeography, is within the subsection *Pinus* and includes five taxa: *Pinus densiflora*, *P. nigra*, *P. resinosa*, *P. sylvestris* and *P. mugo / uncinata* species complex. In this study, we use this clade as an example to explain weak support in phylogenetic studies of closely related pine species and show that some of the most popular genetic markers, namely the chloroplast DNA barcoding sequences *matK*, *psbA-trnH* and *rbcL*, are not recommended for species identification purposes in European pines. In addition, we show that *matK* and *psbA-trnH* contain contradicting phylogenetic signals in some of the most economically important pine species.

Keywords: Gene Tree, Taxonomy, *Pinus*, GenBank

Introduction

Identification of individuals and multispecies phylogenies are usually based on established universal markers, so called DNA barcodes. DNA barcoding markers *matK* and *rbcL* are the standard markers for plants (Hollingsworth et al. 2009a). However, additional DNA regions are often needed to identify specimens at species level in plant groups that contain little variation among these markers, such as conifers (Hollingsworth et al. 2009b). In pines in particular, the most frequently used additional markers are the intergenic spacer *psbA-trnH* and more recently, the highly variable *ycf1* gene (Parks et al. 2011, Hernández-León et al. 2013, Dong et al. 2015).

The *psbA-trnH* region has been successfully used in a few phylogenetic Pinaceae studies (Hao et al. 2010, Ran et al. 2010) but due to technical problems related to the sequencing of the duplicated *psbA-trnH* region in some pines (Lidholm et al. 1991, Lidholm & Gustafsson 1991, Hernández-León et al. 2013), the use of this marker has been limited. In addition, observations that BLAST searches performed on *psbA-trnH* sequences do not always match conspecific specimens from GenBank (*P. cembra, P. nigra, P. mugo* and P. *Sylvestris* – Armenise et al. 2012) puts into question the reliability of this marker for species identification purposes. Within-species variability and high plasticity of the *psbA-trnH* intergenic spacer with frequent indels and inversions have been observed in angiosperms and bryophytes (Borsch & Quandt 2009).

The largest and most economically important conifer genus *Pinus* is widespread in the northern hemisphere. The number of species is currently estimated to be from about 110 to more than 200, but there exists disagreement about the exact numbers (Price et al. 1998, Gernandt et al. 2008, Eckenwalder 2009, Farjon & Filer 2013). Over the years, several classifications have been proposed for the genus *Pinus* based on morphological and anatomical traits (Gaussen 1960), isozyme loci (Karalamangala & Nickrent 1989), restriction patterns of chloroplast genome (Strauss & Doerksen 1990, Krupkin et al. 1996), sequence data from nuclear ITS (Internal Transcribed Spacer – Liston et al. 1999), mainly chloroplast organelle regions (Gernandt et al. 2005, Eckert & Hall 2006) as well as the full chloroplast genome (Parks et al. 2009). However, and although these numerous studies are based on different and complementary methods, some taxonomic subdivisions within each subgenus remain unresolved, especially among terminal taxa in some of the subsections (Eckert & Hall 2006, Parks et al. 2009, Gernandt et al. 2018).

One example of species relationships that has been difficult to resolve is within the section *Pinus*, subsection *Pinus* where a group of four Eurasian (*Pinus densiflora* Siebold & Zucc., *Pinus nigra* Arnold, *Pinus sylvestris* L. and *Pinus mugo* Turra) pines and the North American *Pinus resinosa* Aiton, are placed in the same clade (Gernandt et al. 2005, Eckert & Hall 2006). In particular, the sister-species relationships of *P. nigra* and *P. resinosa* first published by Eckert & Hall (2006), has aroused comments. In this phylogeny, the clade appear-

(1) Department of Forest Ecology & Genetics, Forest Research Centre, INIA-CSIC, Carretera de la Coruña km 7.5, E-28040 Madrid (Spain); (2) INRAE, UR629, Ecologie des Forêts Méditerranéennes - URFM, Domaine Saint Paul, F-84914 Avignon (France); (3) Institute of Biosciences and Bioresources, National Research Council, Division of Florence, I-50019 Sesto Fiorentino, FI (Italy)

@ Sanna Olsson (sanna.olsson@csic.es)

Received: Jun 25, 2021 - Accepted: Feb 15, 2022

Citation: Olsson S, Giovannelli G, Roig A, Spanu I, Vendramin GG, Fady B (2022). Chloroplast DNA barcoding genes *matK* and *psbA-trnH* are not suitable for species identification and phylogenetic analyses in closely related pines. iForest 15: 141-147. - doi: 10.3832/ifor3913-015 [online 2022-04-25]

Communicated by: Alberto Santini

ed during the Oligocene approx. 30 million years ago and the divergence between the North American *P. resinosa* and the Eurasian *P. nigra* dates from much later, approx. 4 million years ago, in clear contradiction with geological events dating the opening of the Atlantic Ocean at 65 million years ago (Tiffney 1985). The authors themselves interpreted the close relationship of these two species to be an incorrect inference of topology. However, this relationship reappears in other studies based on comprehensive taxon sampling and eight sequenced chloroplast markers (Gallien et al. 2016, Saladin et al. 2017). The sequences used in these studies came from independent sources: the study by Eckert & Hall used sequences originally published by Wang et al. 1999 (*P. nigra*: *matK* AB019854, *rbcL* AB019817) and Geada López et al. 2002 (*P. resinosa: matK* AB063516, *rbcL*: AB063384), while Gallien et al. (2016) and Saladin et al. (2017) used the whole chloroplast sequences produced by Parks et al. (2009, 2012). It is noteworthy that Parks et al. (2012) published a tree based on the whole chloroplast sequences that disagrees with the *P. resinosa*-*P. nigra* relationship, as do several other studies. In Gernandt et al. (2005), who also used *matK* and *rbcL* but mainly newly produced sequences (*P. nigra*: *matK* AB084498, *rbcL* AB019817; *P. resinosa*: *matK* AY497288, *rbcL* AY497252), these species are placed in unresolved position within the subsection *Pinus*. Using nuclear genes (Palmé et al. 2009) or a combination of molecular and morphological data (Grotkopp et al. 2004), *P. nigra* was more closely related to *P. sylvestris* than to *P. resinosa*.

We wanted to study if differences in gene tree topology could be the explanation for incongruent results in the phylogenetic relationships of certain taxa between studies. As an alternative, we wanted to test if within-species variation in *P. nigra* was a reason for incongruent phylogenetic placements. *P. nigra* has a large distribution area where five subspecies are generally described (Von Raab-Straube 2014, Euro+Med PlantBase - http://www.emplantbase.org/), although up to six different lineages are recognized by the most recent genetic study on this group (Scotti-Saintagne et al. 2019). Yet another possibility was that the material in GenBank includes hybrids or introgressed individuals, since hybridization is quite common in conifers, and not always visible from a morphological point of view (Vasilyeva & Goroshkevich 2019). Errors in labeling and misidentification of specimens in sequences deposited in GenBank could not be ruled out either. Especially young trees might be difficult to identify. Therefore, we re-sequenced the chloroplast markers *matK, psbA-trnH* and *rbcL* from several new samples, including different subspecies of *P. nigra*, and compared them to sequences deposited in GenBank in order to provide a critical assessment on these barcoding genes and the reliability of sequence information obtained from GenBank. We discuss why inconsistencies may be frequent in phylogenies of the genus *Pinus* and possibly other conifers, and propose solutions to this problem.

Material and methods

Taxon sampling and sequence data

During preliminary analyses of the sequence data produced for this study, including the five pine species *P. densiflora, P. nigra, P. resinosa, P. sylvestris* and *P. mugo / uncinata*, we noticed that there existed multiple types of DNA sequences in samples from the same species, both among our and the downloaded sequences. To understand the reasons for inconsistent sequences in these markers, all available *matK, rbcL* and *trnH-psbA* sequences for these species based on Blast search results were downloaded from GenBank. *P. halepensis* was used as outgroup in the analyses. To minimize the possibility of mislabeled samples, we re-sequenced several additional samples focusing on the emblematic clade of Eurasian and New World species that include *P. densiflora, P. nigra, P. resinosa, P. sylvestris* and *P. mugo / uncinata* species complex. Sequences from Scotti-Saintagne et al. (2019) were obtained, consisting of 21-34 *P. nigra* individuals representing a total of 20 different populations (the number of individuals and populations varied among markers) and the five *P. nigra* subspecies (as defined in Euro+Med PlantBase – Von Raab-Straube 2014). They were complemented with additional sequences (*matK*: MK028114 – MK028128; *rbcL*: MK092816 – MK092835; *trnH-psbA*: LR590646 – LR590656). The number of individuals sequenced in this study ranged, per species, from a minimum of two to a maximum of fifteen (Tab. 1).

DNA was extracted from leaf tissue using the DNeasy® 96 Plant Kit (QIAGEN, Germany) at the INRA molecular biology laboratory in Avignon, France. Primers were obtained from Kress & Erickson (2007 – *matK* and *rbcL*) and Kress et al. (2005 – *trnH-psbA*). Polymerase chain reactions (PCR) were performed according to the following conditions: denaturation at 95 °C for 5 min, followed by 35 cycles at 94 °C for 30 sec, 48 °C or 53 °C (*matK* or *rbcL* and *trnH-psbA*) for 30 sec and 72 °C for 45 sec with a final 10 min extension step at 72 °C. PCR final volume was optimized to 30 µl and the PCR mix contained: 0.2 mM of each dNTP, 2.5 mM of $MgCl_2$, 0.3 µM of each primers, 1X GoTaq® Flexi Buffer, 1.25 U of GoTaq® DNA Polymerase (Promega, USA) and 30 ng DNA. PCR products were quality-checked on 1.5% agarose gel stained with Ethidium Bromide (EtBr) and successfully amplified samples were sent to the French Genomics Institute "Genoscope" for Sanger sequencing. Sequences were quality checked and edited using CodonCode Aligner® v. 3.7.1 (CodonCode Co., MA, USA); low quality sequences were trimmed and a consensus sequence for each individual using both forward and reverse sequences were obtained, when possible.

The alignments were edited with PhyDE® v. 1.0 (Müller et al. 2005). The re-sequenced and downloaded sequences were organized in groups of identical sequences, *i.e.*, haplotypes, and the species together with their number of occurrence was recorded.

Phylogenetic analyses

Phylogenetic trees using both Bayesian and maximum likelihood approaches were built for each marker. Bayesian analyses were performed using MrBayes v. 3.2.3 (Ronquist et al. 2012), applying the best-fit substitution model selected using the AIC criterion in jModeltest v. 2.1.10 (Darriba et al. 2012). Four runs with four chains (10^6 iterations each) were run simultaneously. Chains were sampled every 1000 iterations and the respective trees written to a tree file. The burn-in was set at 250,000 generations. Tracer v. 1.6 (Rambaut et al. 2014) was used for the output of the model parameters to examine the sampling and convergence results. In addition, the concatenated data matrix was analyzed by maximum likelihood (ML) after automatic model selection using ModelFinder (Kalyaanamoorthy et al. 2017) implemented in IQ-Tree v. 1.4.2 (Nguyen et al. 2015) applying 1000 ultrafast bootstrap replicates. Consensus topologies and support values from the different methodological approaches were compiled and drawn using TreeGraph2 (Stöver & Müller 2010).

Tab. 1 - Number of sequences generated for this study. The number of individuals, country of origin and number of sequences per DNA region are shown.

Species	No. ind	Country	matK	rbcL	trnH-psbA
Pinus halepensis	13	France/ Israel/ Morocco/ Turkey	12	13	13
Pinus nigra	34	Algeria/ Austria/ Bulgaria/ Croatia/ Cyprus/ France/ Greece/ Italy/ Morocco/ Romania/ Serbia/ Spain/ Turkey/ Ukraine	21	25	34
Pinus sylvestris	7	France	7	4	6
Pinus uncinata	3	France	3	2	2

Results

Within-species polymorphism detected in conservative chloroplast markers

All three sequence regions among these closely related pines are conserved and contain a low number of variable and parsimony informative (present in at least two samples) sites (Tab. 2). The three alignments were deposited in Zenodo at https://zenodo.org/deposit/4889916.

Our results show that the counter-intuitive grouping of *P. resinosa* with *P. nigra* is not due to the representation of a specific subspecies by the selected *P. nigra* sample. No variation was detected at barcoding cytoplasmic genes within the *P. nigra* subspecies despite their high morphological, physiological and ecological variability as well as high diversity at other nuclear genes (Scotti-Saintagne et al. 2019). However, when comparing the newly produced sequences with sequences available in GenBank, different haplotypes of *P. nigra* were detected for each marker (Tab. 3). Likewise, the *P. halepensis* and *P. uncinata* sequences produced for this study were monomorphic but different haplotypes were retrieved from GenBank. The re-sequenced *P. sylvestris* samples represented two different haplotypes, complemented with further downloaded haplotypes (see Tab. S1 in Supplementary material for accession information and references for all sequences used in this study). Since several haplotypes were shared across species, the selection of one sample or another would affect the results of phylogenetic analyses based on these markers. The inclusion of all haplotypes in the analyses performed to resolve taxonomic relationships among the five pine species (*P. densiflora, P. nigra, P. resinosa, P. sylvestris* and *P. uncinata*) resulted in ambiguous placements of several species, not consistent with species delimitations (Fig. 1, Fig. 2, Fig. 3). Due to the detected within-species variability and low number of parsimony-informative sites, the taxonomic position of the studied species based on only these markers could not be reliable.

Variation in matK causes species non-monophyly

There were nine haplotypes present in the *matK* alignment. Only the *P. halepensis* sequences were species-specific and monomorphic (except for one unresolved base in JN854197). The most frequent haplotypes were shared by different species: the densiflora-mugo-uncinata-sylvestris haplotype was present in 36 samples from several different studies, while the nigra-resinosa haplotype contained mainly sequences from *P. nigra* and two *resinosa* samples from two different studies. The densiflora-sylvestris haplotype contained six *densiflora* samples and one *sylvestris*, all from the same study.

It was not possible to build a well-resolved phylogenetic tree based on the *matK* alignment. In addition to lack of support for the nodes, the branch lengths were short. Noteworthy, *P. sylvestris* and *P. resinosa* were resolved as non-monophyletic. The expected outgroup *P. halepensis* was grouped together with a *P. sylvestris* haplotype (sylvestris 2) represented by only one sample. This *P. sylvestris* sample (MK036240) was produced within the current study and a Blast search showed that the compared sequence parts were identical with other pine species (*P. canariensis, P. pinea, P. pinaster* and *P. roxburghii*) instead of *P. Sylvestris*.

Random distribution of the psbA-trnH intergenic spacer sequences

Again, the phylogenetic tree based on the *psbA-trnH* intergenic spacer is not likely to represent a satisfactory hypothesis of rela-

Tab. 2 - Alignment statistics. Length of alignment (bp), number of variable sites, percentage of variable sites, number of parsimony informative sites and percentage of parsimony informative sites are shown.

Region	Bp	variable	%variable	informative	%informative
matK	510	19	3.7	12	2.4
trnH-psbA	541	22	4.1	17	3.1
rbcL	627	4	0.6	2	0.3

Tab. 3 - A list of *matK, psbA-trnH* and *rbcL* haplotypes. The haplotypes are named according to the species they occurred in. The column nb indicates the total number of individuals represented by each haplotype. The last column indicates the number of individuals per species.

Region	Haplotype	nb	Species
matK	halepensis_1	14	-
	densiflora-sylvestris	7	*densiflora*: 6, *sylvestris*: 1
	densiflora-mugo-uncinata-sylvestris	36	*densiflora*: 11, *mugo*: 10, *uncinata*: 3, *sylvestris*: 12
	nigra-resinosa	37	*nigra*: 35, *resinosa*: 2
	nigra_1	1	-
	resinosa_1	3	-
	sylvestris_1	1	-
	sylvestris_2	1	-
	uncinata_1	1	-
psbA-trnH	halepensis_1	25	-
	halepensis_2	1	-
	halepensis_3	1	-
	densiflora_1	3	-
	densiflora-sylvestris	30	*densiflora*: 16, *sylvestris*: 14
	mugo	1	1
	mugo-sylvestris-uncinata	13	*mugo*: 5, *sylvestris*: 1, *uncinata*: 7
	nigra_1	27	-
	nigra_2	1	-
	nigra_3	1	-
	nigra_4	1	-
	resinosa_1	1	-
	resinosa_2	1	-
	sylvestris_1	6	-
	sylvestris_2	1	-
rbcL	halepensis-resinosa	20	*halepensis*: 14, *resinosa*: 6
	densiflora-nigra-resinosa-sylvestris	39	*densiflora*: 16, *nigra*: 4, *resinosa*: 1, *sylvestris*: 18
	densiflora-nigra	5	*densiflora*: 4, *nigra*: 1
	mugo_1	1	-
	mugo_2	1	-
	nigra-sylvestris	24	*nigra*: 20, *sylvestris*: 4
	uncinata_1	2	-

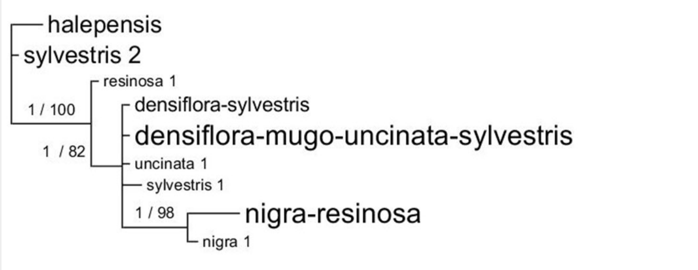

Fig. 1 - Phylogram based on *matK* sequences. The haplotypes are named according to the species they occurred in, with a font size scaled according to the frequency (bigger font for more common haplotype). The tree represents the majority consensus of trees sampled after stationarity in the Bayesian analysis. Posterior probability values from the Bayesian inference are indicated first and the corresponding bootstrap values of the maximum likelihood analysis are shown after when applicable. Only bootstrap values ≥ 50 are indicated.

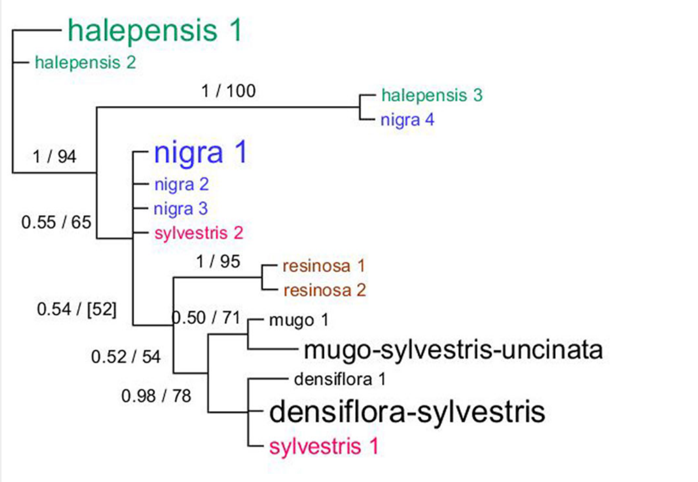

Fig. 2 - Phylogram based on *trnH-psbA* sequences. The haplotypes are named according to the species they occurred in, with a font size scaled according to the frequency (bigger font for more common haplotype). Different haplotypes from same species are colored with the same color. The tree represents the majority consensus of trees sampled after stationarity in the Bayesian analysis. Posterior probability values from the Bayesian inference are indicated first and the corresponding bootstrap values of the maximum likelihood (ML) analysis are shown after when applicable. Only bootstrap values ≥ 50 are indicated. Square brackets indicate topological conflict between Bayesian and ML tree.

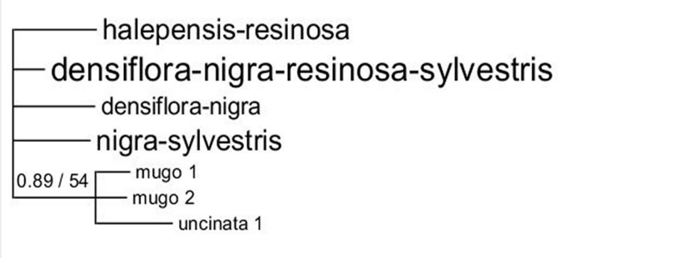

Fig. 3 - Phylogram based on *rbcL* sequences. The haplotypes are named according to the species they occurred in, with a font size scaled according to the frequency (bigger font for more common haplotype). The tree represents the majority consensus of trees sampled after stationarity in the Bayesian analysis. Posterior probability values from the Bayesian inference are indicated first and the corresponding bootstrap values of the maximum likelihood analysis are shown after when applicable. Only bootstrap values ≥ 50 are indicated.

tionships (Fig. 2). The sequences did not group according to species, as can be observed both from the aligned sequences and the phylogeny based on them. The resolution and support values were low for many branches in both Bayesian and ML analyses. Since incongruences at nodes with high support were not present, only the Bayesian consensus tree is shown (Fig. 2). The only species that appeared monophyletic based on this data set is *P. resinosa*. The position of *P. halepensis* as outgroup was not supported. One *P. nigra* sequence (nigra 4 haplotype, EU531715) was almost identical (only 1 bp difference) to a sequence from *P. halepensis* (halepensis 3, FN689388), including a 10 bp inversion at the beginning of the sequences, and both were shown by a BLAST search to share 100% similarity with sequences from *P. pinaster*. The most common haplotypes of both *P. halepensis* (halepensis_1) and *P. nigra* (nigra_1) represented numerous samples (25 and 27, respectively). The most common *psbA-trnH* haplotype was the densiflora-sylvestris haplotype with sequences shared by 30 samples. Although the level of variation was generally low in these closely related species (Tab. 2), several small repetitions, deletions and inversions were observed in the alignments with an apparently random distribution across species.

Conservative rbcL region

The *rbcL* region contained only four variable sites of which two were parsimony informative. In addition, two of the variable sites were located at the ends of the alignment, within the first or last 5 bp, which might affect their reliability due to bad quality sequences in the ends of sequences. The most common haplotypes were shared among species: the densiflora-nigra-resinosa-sylvestris type was detected in 39 samples, the nigra-sylvestris type in 24 samples, the halepensis-resinosa in 20 samples and the densiflora-nigra in 5 samples. The low level of variation within this region was insufficient to resolve phylogenetic relationships in the species included in this study (Fig. 3).

Discussion

Evaluation of barcoding markers in pines from subsection Pinus

The chloroplast DNA markers *matK*, *psbA-trnH* and *rbcL* have been shown to be poor barcoding genes for hard pines (Hernández-León et al. 2013) and the *P. mugo* complex (Celinski et al. 2017), as well as Cycads (Sass et al. 2007) and closely related angiosperm species complexes (Roy et al. 2010, Von Cräutlein et al. 2011, Yan et al. 2018). Nonetheless, they are still commonly used for taxonomic identification and phylogenetic inference in all land plants. Our results confirmed that none of them are well suited for species identification and phylogenetic inference in the group of pines we studied, and in addition, pointed out inconsistent within-species variability.

It is difficult to conclude with certainty whether the variability is due to real variation or independent errors caused by technical artifacts, wrong identification and mislabeling among the studied samples. In general, very few studies report error rates on sequencing and labeling. As an exception, the thorough study by De Vere et al. (2012) mentioned an error rate of 4.8% for *rbcL* and 3.8% for *matK* region. This error rate was reported on sequence production, and because of verification of results and correction of identified errors, the error rates in sequences that are actually used in research and submitted to GenBank can be assumed to be lower. Attention must naturally be paid when using material downloaded from external sources, but if human error is assumed to be the source of all variability in the studied sequences, then the error rate among this particular group of organisms would be exceptionally high and, in our opinion, unlikely. However, if the detected variation was real, there might be several biological reasons for this.

Wind pollination and predominantly paternally inheritance of chloroplast genome, large effective population sizes, genetically variable populations, hybridization and chloroplast capture are factors that cause incomplete lineage sorting and therefore interfere with species division based on genetic markers in conifers. Indeed, local chloroplast capture has been documented in the genus (Hong et al. 1993, Matos & Schaal 2000, Liston et al. 2007, Willyard et al. 2009, Hernández-León et al. 2013). Also, Heuertz et al. (2010) reported that the genetic variation based on chloroplast SSR data in the *P. mugo / uncinata* complex was not structured according to morphology, but according to geography, and that some haplotypes were shared between *P. mugo* and the closely related *P. sylvestris*.

Chloroplast capture is possible only where species are sympatric, either current-day or sometime in the past. In this study we did not observe clear differentiation by geographic region, though a tendency of finding the same haplotypes within the same published study can be observed. Again, mix-up of samples within the study and consistent artifacts (*i.e.*, caused by bad quality sequences) usually in the ends of the sequences, when the same primers and sequencing methods are used, might explain some of these differences. In general, results based on regions with low variability and especially variation at sequence ends should be considered with caution since they could be low-quality sequence induced errors. Consistent differences were detected in the middle of the sequences of several samples, and we therefore believe that real biological differences exist, in addition to possible artifacts and human errors.

Perspectives and alternative molecular markers

We confirm that the chloroplast DNA markers *matK*, *psbA-trnH* and *rbcL* should not be used for reliable phylogenetic and taxonomic inference in pines. We also suggest that the reliability of the studied markers need to be analyzed in other conifer species. Within-species variation might go unnoticed in multi-species phylogenetic analyses in which it is often common practice to use only one specimen per species due to limited resources. Furthermore, when different gene variants exist and one of the forms is locally more common, even the inclusion of several samples would not make an obvious difference. The fast evolving marker *ycf1* harbors a large amount of variable sites but it might not reflect species relationships correctly (Saladin et al. 2017). In addition, its usefulness for species identification using a large amount of samples still needs to be studied.

The use of nuclear genes will be useful for evolutionary and phylogenetic analyses. While chloroplast capture often occurs in the absence of nuclear introgression (Rieseberg & Soltis 1991), hybridization also affects nuclear genes. On one hand, nuclear genes are exchanged less freely between species (Whittemore & Schaal 1991). On the other hand, nuclear genes are likely to show non-monophyly in taxa sharing similar life history traits due to the absence of allelic coalescence (Syring et al. 2005). It will be important to employ both chloroplast and nuclear markers to study phylogenetic relationships and evolutionary history, including possible hybridization and chloroplast capture in these species. Currently the only nuclear barcoding marker is the widely used internal transcribed spacer (ITS), but its complex and unpredictable evolutionary behavior has been shown to reduce its utility in phylogenetic analyses (Alvaréz & Wendel 2003), being especially puzzling in pines (Liston et al. 1999). Recent studies have provided large amounts of nuclear sequence regions, *e.g.*, using Hyb-Seq (Gernandt et al. 2018) or identification of orthologous genes (Olsson et al. 2020) for specific pine groups, but it remains to test which of these potential markers are suitable to be used as general DNA barcoding markers in all pines.

Acknowledgements

We thank Denis Vauthier (INRAE UEFM, Avignon, France) for sample collection. We appreciate comments from four reviewers on the current and an earlier version of the manuscript.

This work was supported by: The French *Office National des Forêts* "*Programme global de conservation des populations françaises de pin de Salzmann*"; the French DNA barcoding program "*Bibliothèque du Vivant*"; the Spanish Ministry of Economy and Competitiveness (MINECO) contract to SO [PTA2015-10836-I]; the Aix-Marseille Université, France (Ecole Doctorale ED251) grant to GG; the short-term scientific mission program of the COST Action FP1202 grant to GG.

References

Alvaréz I, Wendel JF (2003). Ribosomal ITS sequences and plant phylogenetic inference. Molecular Phylogeny and Evolution 29: 417-434. - doi: 10.1016/S1055-7903(03)00208-2

Armenise L, Simeone MC, Piredda R, Schirone B (2012). Validation of DNA barcoding as an efficient tool for taxon identification and detection of species diversity in Italian conifers. European Journal of Forest Research 131: 1337-1353. - doi: 10.1007/s10342-012-0602-0

Borsch T, Quandt D (2009). Mutational dynamics and phylogenetic utility of noncoding chloroplast DNA. Plant Systematics and Evolution 282: 169-199. - doi: 10.1007/s00606-009-0210-8

Celinski K, Kijak H, Wojnicka-Póltorak A, Buczkowska-Chmielewska K, Sokolowska J, Chudzinska E (2017). Effectiveness of the DNA barcoding approach for closely related conifers discrimination: a case study of the *Pinus mugo* complex. Comptes Rendus Biologies 340: 339-348. - doi: 10.1016/j.crvi.2017.06.002

Darriba D, Taboada GL, Doallo R, Posada D (2012). jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9: 772-772. - doi: 10.1038/nmeth.2109

De Vere N, Rich TCG, Ford CR, Trinder SA, Long C, Moore CW, Sattertwaite D, Davies H, Allainguillaume J, Ronca S, Tatarinova T, Garbett H, Walker K, Wilkinson MJ (2012). DNA Barcoding the Native Flowering Plants and Conifers of Wales. PLoS One 7: e37945. - doi: 10.1371/journal.pone.0037945

Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J, Zhou S (2015). *ycf1*, the most promising plastid DNA barcode of land plants. Scientific Reports 5 (1): 313. - doi: 10.1038/srep08348

Eckenwalder JE (2009). Conifers of the world: the complete reference. Timber Press, Portland, OR, USA, pp. 720.

Eckert AJ, Hall BD (2006). Phylogeny, historical biogeography, and patterns of diversification for *Pinus* (Pinaceae): phylogenetic tests of fossil-based hypotheses. Molecular Phylogeny and Evolution 40: 166-182. - doi: 10.1016/j.ympev.2006.03.009

Farjon A, Filer D (2013). An atlas of the world's conifers: an analysis of their distribution, biogeography, diversity and conservation status. Brill, Leiden, Netherlands, pp. 512.

Gallien L, Saladin B, Boucher FC, Richardson DM, Zimmermann NE (2016). Does the legacy of historical biogeography shape current invasiveness in pines? New Phytologist 209: 1096-105. - doi: 10.1111/nph.13700

Gaussen H (1960). Les Gymnospermes actuelles et fossiles. Fascicule VI Chapitre XI, Généralités, genre *Pinus* [Current and fossil Gymnosperms. Fascicle VI Chapter XI, General, genus *Pinus*]. Travaux du Laboratoire Forestier de Toulouse, Faculté des Sciences, Toulouse, France. [in French]

Geada López G, Kamiya K, Harada K (2002). Phylogenetic relationships of *Diploxylon* pines (Subgenus *Pinus*) based on plastid sequence data. International Journal of Plant Sciences 163: 737-747. - doi: 10.1086/342213

Gernandt DS, López G, García S, Liston A (2005). Phylogeny and classification of *Pinus*. Taxon 4: 29-42. - doi: 10.2307/25065300

Gernandt DS, Magallón S, López G, Flores OZ, Willyard A, Liston A (2008). Use of simultaneous analyses to guide fossil-based calibrations of Pinaceae phylogeny. International Journal of Plant Sciences 169: 1086-1099. - doi: 10.1086/590472

Gernandt DS, Aguirre X, Vázquez-Lobo A, Willyard A, Moreno A, Pérez De la Rosa JA, Piñero D, Liston A (2018). Multi-locus phylogenetics, lineage sorting, and reticulation in Pinus subsection Australes. American Journal of Botany 105: 1-15. - doi: 10.1002/ajb2.1015

Grotkopp E, Rejmánek M, Danderson MJ, Rost TL (2004). Evolution of genome size in pines (*Pinus*) and its life-history correlates: supertree analyses. Evolution 58: 1705-1729. - doi: 10.1111/j.0014-3820.2004.tb00456.x

Hao DC, Chen SL, Xiao PG (2010). Sequence characteristics and divergent evolution of the chloroplast psbA-trnH noncoding region in gymnosperms. Journal of Applied Genetics 51: 259-273. - doi: 10.1007/BF03208855

Hernández-León S, Gernandt D, Pérez De la Rosa J, Jardón-Barbolla L (2013). Phylogenetic relationships and species delineation in *Pinus* Section *Trifoliae* inferred from plastid DNA. PLoS One 8 (7): e70501. - doi: 10.1371/journal.pone.0070501

Heuertz M, Teufel J, González-Martínez SC, Soto A, Fady B, Alía R, Vendramin GG (2010). Geography determines genetic relationships between species of mountain pine (*Pinus mugo* complex) in Western Europe. Journal of Biogeography 37: 541-556. - doi: 10.1111/j.1365-2699.2009.02223.x

Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL, Fazekas AJ, Graham SW, James KE, Kim KJ, Kress WJ, Schneider H, van Alphenstahl J, Barrett SC, van den Berg C, Bogarin D, Burgess KS, Cameron KM, Carine M, Chacón J, Clark A, Clarkson JJ, Conrad F, Devey DS, Ford CS, Hedderson TA, Hollingsworth ML, Husband BC, Kelly LJ, Kesanakurti PR, Kim JS, Kim YD, Lahaye R, Lee HL, Long DG, Madriñán S, Maurin O, Meusnier I, Newmaster SG, Park CW, Percy DM, Petersen G, Richardson JE, Salazar GA, Savolainen V, Seberg O, Wilkinson MJ, Yi DK, Little DP (2009a). A DNA barcode for land plants. Proceedings of the National Academy of Sciences USA 106 (31): 12794-12797. - doi: 10.1073/pnas.0905845106

Hollingsworth ML, Clark AA, Forrest LL, Richardson J, Pennington RT, Long DG, Cowan R, Chase MW, Gaudeul M, Hollingsworth PM (2009b). Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Molecular Ecology Resources 9: 439-457. - doi: 10.1111/j.1755-0998.2008.02439.x

Hong Y-P, Krupkin AB, Strauss SH (1993). Chloroplast DNA transgresses species boundaries and evolves at variable rates in the California closecone pines (*Pinus radiata*, *P. muricata* and *P. attenuata*). Molecular Phylogeny and Evolution 2 (4): 322-329. - doi: 10.1006/mpev.1993.1031

Kalyaanamoorthy S, Minh BQ, Wong TK, Von Haeseler A, Jermiin LS (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14 (6): 587-589. - doi: 10.1038/nmeth.4285

Karalamangala RR, Nickrent DL (1989). An electrophoretic study of representatives of subgenus *Diploxylon* of *Pinus*. Canadian Journal of Botany 67: 1750-1759. - doi: 10.1139/b89-222

Kress WJ, Erickson DL (2007). A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. PLoS One 2 (6): e508. - doi: 10.1371/journal.pone.0000508

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005). Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences USA 102 (23): 8369-8374. - doi: 10.1073/pnas.0503123102

Krupkin AB, Liston A, Strauss SH (1996). Phylogenetic analysis of the hard pines (*Pinus* subgenus *Pinus*, Pinaceae) from chloroplast DNA restriction site analysis. American Journal of Botany 83: 489-498. - doi: 10.1002/j.1537-2197.1996.tb12730.x

Lidholm J, Szmidt A, Gustafsson P (1991). Duplication of the *psbA* gene in the chloroplast genome of two *Pinus* species. Molecular and General Genetics 226: 345-352. - doi: 10.1007/BF00260645

Lidholm J, Gustafsson P (1991). A three-step model for the rearrangement of the chloroplast *trnK-psbA* region of the gymnosperm *Pinus contorta*. Nucleic Acids Research 19 (11): 2881-2887. - doi: 10.1093/nar/19.11.2881

Liston A, Robinson WA, Piñero D, Alvarez-Buylla ER (1999). Phylogenetics of *Pinus* (Pinaceae) based on nuclear ribosomal DNA internal transcribed spacer region sequences. Molecular Phylogeny and Evolution 11: 95-109. - doi: 10.1006/mpev.1998.0550

Liston A, Parker-Defeniks M, Syring JV, Willyard A, Cronn R (2007). Interspecific phylogenetic analysis enhances intraspecific phylogeographical inference: a case study in *Pinus lambertiana*. Molecular Ecology 16: 3926-3937. - doi: 10.1111/j.1365-294X.2007.03461.x

Matos JA, Schaal BA (2000). Chloroplast evolution in the *Pinus montezumae* complex: a coalescent approach to hybridization. Evolution 54: 1218-1233. - doi: 10.1111/j.0014-3820.2000.tb00556.x

Müller KF, Quandt D, Müller J, Neinhuis C (2005). PhyDE® 0.995: phylogenetic data editor. Web site. [online] URL: http://www.phyde.de

Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution 32: 268-274. - doi: 10.1093/molbev/msu300

Olsson S, Grivet D, Cattonaro F, Vendramin V, Giovannelli G, Scotti-Saintagne C, Vendramin GG, Fady B (2020). Evolutionary relevance of lineages in the European black pine (*Pinus nigra*) in the transcriptomic era. Tree Genetics and Genomes 16 (2): 508. - doi: 10.1007/s11295-020-1424-8

Palmé AE, Pyhäjärvi T, Wachowiak W, Savolainen O (2009). Selection on nuclear genes in a *Pinus* phylogeny. Molecular Biology and Evolution 26: 893-905. - doi: 10.1093/molbev/msp010

Parks M, Cronn R, Liston A (2009). Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biology 7 (1): 19363. - doi: 10.1186/1741-7007-7-84

Parks M, Liston A, Cronn R (2011). Newly developed primers for complete *ycf1* amplification in *Pinus* (Pinaceae) chloroplasts with possible family-wide utility. American Journal of Botany 98: 185-188. - doi: 10.3732/ajb.1100088

Parks M, Cronn R, Liston A (2012). Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenomic data set from *Pinus* L. (Pinaceae). BMC Evolutionary Biology 12 (1): 100. - doi: 10.1186/1471-2148-12-100

Price RA, Liston A, Strauss SH (1998). Phylogeny and systematics of *Pinus*. In: "Ecology and Biogeography of *Pinus*" (Richarson DM ed). Cambridge University Press, Cambridge, UK, pp. 49-68. [online] URL: http://books.google.com/books?id=YawYOzQmcHEC

Rambaut A, Suchard MA, Xie Drummond D AJ (2014). Tracer v1.6. Web site. [online] URL: http://beast.bio.ed.ac.uk/Tracer

Ran JH, Wang PP, Zhao HJ, Wang XQ (2010). A test of seven candidate barcode regions from the plastome in *Picea* (Pinaceae). Journal of Integrative Plant Biology 52 (12): 1109-1126. - doi: 10.1111/j.1744-7909.2010.00995.x

Rieseberg LH, Soltis DE (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. Evolutionary Trends in Plants 5 (1): 65-84. [online] URL: http://www.researchgate.net/publication/262005952

Ronquist F, Teslenko M, Van Der Mark P, Ayres D, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61: 539-542. - doi: 10.1093/sysbio/sys029

Roy S, Tyagi A, Shukla V, Kumar A, Singh UM, Chaudhary LB, Datt B, Bag SK, Singh PK, Nair NK, Husain T, Tuli R (2010). Universal plant DNA barcode loci may not work in complex groups: a case study with Indian *Berberis* species. PLoS One 5 (10): e13674. - doi: 10.1371/journal.pone.0013674

Saladin B, Leslie AB, Wüest RO, Litsios G, Conti E, Salamin N, Zimmermann NE (2017). Fossils matter: improved estimates of divergence times in *Pinus* reveal older diversification. BMC Evolutionary Biology 17 (1): 49. - doi: 10.1186/s12862-017-0941-z

Sass C, Little DP, Stevenson DW, Specht CD (2007). DNA barcoding in the Cycadales: testing the potential of proposed barcoding markers for species identification of Cycads. PLoS One 11: e1154. - doi: 10.1371/journal.pone.0001154

Scotti-Saintagne C, Giovannelli G, Scotti I, Roig A, Spanu I, Vendramin GG, Guibal F, Fady B (2019). Recent, late Pleistocene fragmentation shaped the phylogeographic structure of the European black pine (*Pinus nigra* Arnold). Tree Genetics and Genomes 15 (5): 355. - doi: 10.1007/s11295-019-1381-2

Strauss SH, Doerksen AH (1990). Restriction fragment analysis of pine phylogeny. Evolution 4: 1081-1096. - doi: 10.1111/j.1558-5646.1990.tb03827.x

Stöver BC, Müller KF (2010). TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics 11: 7. - doi: 10.1186/1471-2105-11-7

Syring J, Willyard A, Cronn R, Liston A (2005). Evolutionary relationships among *Pinus* (Pinaceae) subsections inferred from multiple low-copy nuclear loci. American Journal of Botany 92: 2086-2100. - doi: 10.3732/ajb.92.12.2086

Tiffney BH (1985). The Eocene North Atlantic land bridge: its importance in Tertiary and modern phytogeography of the Northern Hemisphere. Journal of the Arnold Arboretum 66: 243-273. - doi: 10.5962/bhl.part.13183

Vasilyeva G, Goroshkevich S (2019). Artificial crosses and hybridization frequency in five-needle pines. Dendrobiology 80: 123-130. - doi: 10.12657/denbio.080.012

Von Cräutlein M, Korpelainen H, Pietiläinen M, Rikkinen J (2011). DNA barcoding: a tool for improved taxon identification and detection of species diversity. Biodiversity and Conservation 20: 373-389. - doi: 10.1007/s10531-010-9964-0

Von Raab-Straube E (2014). Gymnospermae. In: "Euro+Med Plantbase - The Information Resource For Euro-mediterranean Plant Diversity". Web site. [online] URL: http://ww2.bgbm.org/EuroPlusMed/

Wang XR, Tsumura Y, Yoshimaru H, Nagasaka K, Szmidt AE (1999). Phylogenetic relationships of Eurasian pines (*Pinus*, Pinaceae) based on chloroplast *rbcL*, *MATK*, *RPL20-RPS18* spacer, and *TRNV* intron sequences. American Journal of Botany 86 (12): 1742-1753. - doi: 10.2307/2656672

Whittemore AT, Schaal BA (1991). Interspecific gene flow in sympatric oaks. Proceedings of the National Academy of Sciences USA 88 (6): 2540-2544. - doi: 10.1073/pnas.88.6.2540

Willyard A, Cronn R, Liston A (2009). Reticulate evolution and incomplete lineage sorting among the ponderosa pines. Molecular Phylogeny and Evolution 52: 498-511. - doi: 10.1016/j.ympev.2009.02.011

Yan M, Xiong Y, Liu R, Deng M, Song J (2018). The application and limitation of universal chloroplast markers in discriminating East Asian evergreen oaks. Frontiers in Plant Science 9: 235. - doi: 10.3389/fpls.2018.00569

Supplementary Material

Tab. S1 - Accession numbers of sequences used in this study.

Link: Olsson_3913@suppl001.pdf