Identification of DNA molecular markers by comparison of *Pinus densiflora* and *Pinus sylvestris* chloroplast genomes

Sang-Chul Kim Corresp., [1] , **Jei-Wan Lee** Corresp., [1] , **Seung-Hoon Baek** [1] , **Ji-Young Ahn** [1] , **Kyung-Nak Hong** [1]

[1] National Institute of Forest Science, Division of Forest Genetic Resources, Suwon, Republic of Korea

Corresponding Authors: Sang-Chul Kim, Jei-Wan Lee
Email address: majin01@korea.kr, leejeiwan@korea.kr

Background: Identifying and characterizing genetic variation can clarify the molecular basis of biological phenomena in plants. In particular, related or morphologically similar species can be distinguished by molecular markers. *Pinus densiflora* Siebold & Zucc. is a species that is distributed in the Korean peninsula, the Japanese archipelago, and China's Shandong and Manchu Provinces and has long been harvested for timber. However, it is difficult to distinguish *P. densiflora* from *Pinus sylvestris* L. both morphologically and phylogenetically. The complete chloroplast genome of *P. densiflora* has not yet been reported. In this study, we sequenced the *P. densiflora* chloroplast genome in order to identify the molecular markers that can be used to distinguish this species from *P. sylvestris.*

Methods: Genomic DNA was extracted from *P. densiflora* samples obtained from the clone bank of the National Forest Seed Variety Center and was sequenced on an Ion Torrent platform. Filtered sequences were assembled with *P. sylvestris* sequences used as a reference and gene annotation was performed. The chloroplast genome sequences of the two species were aligned and the number and location of forward, reverse, complement and palindromic matches were determined. Single nucleotide polymorphisms (SNPs) and insertion/deletion mutations (Indels) were identified and analyzed by PCR.

Results: The *P. densiflora* chloroplast genome consisted of circular double-stranded DNA with 119,835 bp compared to 119,758 bp for *P. sylvestris*. Between the two *Pinus* chloroplast genomes, we identified 73 SNPs and 171 Indels; two gene regions with amplification products ≤ 300 bp (*rpoC1* and *trnM-trnV*) were validated as molecular markers.

Discussion: PCR restriction fragment length polymorphism analysis revealed differences between *P. sylvestris* and *P. densiflora* at the molecular level. These differences can be used to distinguish between these two species, which is not possible by microscopy-based morphological examination.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26506v1 | CC BY 4.0 Open Access | rec: 12 Feb 2018, publ: 12 Feb 2018

Introduction

Pinaceae, comprising 11 genera and more than 200 species, is the largest extant family of gymnosperms. Many species of the pine family constitute the major forest elements in the northern temperate region (Wang, Tank & Sang. 2000). *Pinus* L. is one of 11 genera in Pinaceae, a monophyletic family among gymnosperms (Farjon, 2010). Approximately 110 species comprise 50% of Pinaceae, making it the largest genus of existing gymnosperms (Syring *et al*., 2005), most of which are distributed in the temperate zone of the Northern Hemisphere. *Pinus* is divided into two subgenera *Strobus* and *Pinus*, according to the number of fibrovascular bundles in the needle (Geada López, Kamiya & Harada, 2002). Approximately 20 species are native to or are cultivated in Korea (Korea National Arboretum and The Plant Taxonomic Society of Korea, 2007; Hong *et al*., 2014). *Pinus densiflora* Siebold & Zucc. is distributed throughout Korea and is one of the most economically important species sustaining forest ecosystems and is harvested for wood and fuel (Lee *et al*., 2004; Kim, Kim & Lim, 2017). *Pinus sylvestris* L. is the most abundant species in Europe and is found from Scotland and Spain to Siberia and northern Asia.

Chloroplasts a type of plastid in plants and algae, are intracellular organelles that carry out photosynthesis (Howe *et al*., 2003). They are presumed to have originated from an endosymbiotic event between cyanobacteria and non-photosynthetic host cells (Dyall, Brown & Johnson, 2004). Plastid genomes are stable in terms of structure, gene content, and gene order across land plants (Jansen *et al*., 2005). The chloroplast genome of higher plants consists of a circular double strand ranging from 120 to 210 kb that usually contains two inverted repeat (IR) regions (IRA and IRB) separated by large and small single-copy regions (LSC and SSC, respectively) (Ravi *et al*., 2008). Most plant genomes have 66–82 protein-coding genes, 29–32 genes encoding tRNAs, and four genes encoding rRNAs, With the exception of non-photosynthetic parasitic plants, gene composition, sequence, content, and orientation are highly conserved among seed plants (Jansen & Ruhlman, 2012). However, structural modifications such as loss of IR domains or entire genes and gene rearrangement have been reported in gymnosperms such as conifers (Lin *et al*., 2010; Wu *et al*., 2011; Wu & Chaw, 2014; Yi *et al*., 2016). The first complete sequences of *Pinus* cpDNA were reported in *Pinus thunbergii* Parl., with 4 rRNA genes and 32 tRNA genes, and the most striking feature is the loss of all 11 functional genes (*ndh* genes) for in subunits of a putative NADH dehydrogenase that are found in the chloroplast genomes of angliosperms and a bryophyte (Wakasugi *et al*., 1994). There are currently; 2,245 complete chloroplast genomes of seed plants in the National Center for Biotechnology Information (NCBI) Organelle Genome Resources database (http://www.ncbi.nlm.nih.gov/genomes/).

Plastid genome sequences are widely used for DNA barcoding, species conservation, genomic evolution, and molecular phylogenetic studies (Moore *et al*., 2007). Identifying and characterizing genetic variation can clarify the molecular basis the of biological phenomena in plants (Agarwal, Shrivastava & Padh, 2008) and provide insight into the mechanisms of evolution and natural selection. In particular, species that are difficult to differentiate morphologically can be distinguished using molecular markers. The complete chloroplast genome can be rapidly sequenced at a relatively low cost (Yi *et al*., 2016). Also, PCR restriction fragment length polymorphism (RFLP) analysis, also known as cleaved amplified polymorphic sequence (CAPS), is widely used to detect intra- and interspecies variation (Rasmussen, 2012).

P. sylvestris and *P. densiflora* belong to the subgenus *Pinus*, section *Pinus*, subsection *Pinus*. These trees are characterized by the shedding of bud-scales along with the leaves and by two cross-sectional vascular bundles in the leaves (Lee, 2003). Futher, molecular phylogenetic studies show that *P. sylvestris* and *P. densiflora* formseparate strongly supported groups, with common morphological features, including irregular cracking of 2-year-old bark (Wang *et al*., 1999; Gernandt, 2005; Hong *et al*., 2014). Thus, *P. sylvestris* and *P. densiflora* are difficult to distinguish morphologically and phylogenetically; as such, the timber of the two species is often combined or illegally substituted. A complete chloroplast genome is available for *P. sylvestris* but not for *P. densiflora.* In this study, we sequenced the chloroplast genome of *P. densiflora* and compared it with that of *P. sylvestris* in order to identify polymorphisms that can serve as molecular markers to distinguish between the two species by PCR-RFLP analysis.

Materials and Methods

Sample collection, DNA extraction, and sequencing

P. densiflora samples were obtained from the clone bank of the National Forest Seed Variety Center (Anmyeondo, Korea; elite tree: Gyeongbuk No. 4) and genomic DNA was isolated from fresh leaves using the Plasmid SV mini kit (GeneAll, Seoul, Korea). DNA samples from plants used in this study are now stored in the DNA Bank of the Forest Genetic Resources Department of the National Institute of Forest Science. Total genomic DNA was extracted from 10 g of fresh leaves using a Plasmid SV mini kit. Whole genome sequencing was performed on the Ion Torrent platform (Life Technologies, Carlsbad, CA, USA). Libraries were sequenced on Ion Proton using the Ion PI Chip kit v3 deposited at full density according to the protocol for 200 bp sequencing supplied by the manufacturer.

Chloroplast genome assembly and annotation

Filtered sequences were assembled using Bowtie2 v. 2.2.3 software (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml; Langmead & Salzberg, 2012) with *P. sylvestris* sequence (GenBank: NC035069) as a reference. In total, 1,449,103 reads were mapped to the reference sequence with an average coverage of 851.1X. Finally, the contigs were assembled using Geneious 10.2.3 (Biomatters, Auckland, New Zealand; Kearse *et al*., 2012). Gene annotation was performed using the Basic Local Alignment Search Tool (BLAST and BLASTX) available on the NCBI website. All tRNA sequences were confirmed using the web-based tool, tRNAScan-SE (Schattner, Brooks & Lowe, 2005) with default settings to corroborate tRNA boundaries identified by Geneious. Genome maps were generated using OrganellarGenomeDRAW (Lohse, Drechsel & Bock, 2007), followed by manual modification.

Comparison of Pinus chloroplast genome sequences

Simple sequence repeats (SSRs) were analyzed using Phobos v. 3.3.12 (Mayer, 2010), with thresholds of eight repeat units for mononucleotide SSRs, four for di- and trinucleotide SSRs, three for tetra- and pentanucleotide SSRs, and two for hexanucleotide SSRs. All detected repeats were manually verified, and redundant results were removed. We aligned the plastid genome sequences of the two *Pinus* species using MAFFT (Katoh *et al*., 2002). For long repeat sequences, the REPuter program was used to assess the number and location of forward, reverse, complement and palindromic matches (Kurtz *et al*., 2001). Repeat identity and size were limited to > 90% and ≥ 25 bp, respectively.

Identification of molecular markers to distinguish between Pinus species

Single nucleotide polymorphisms (SNPs) and insertion/deletion mutations (Indels) were identified using Geneious 10.2.3 and analyzed by PCR. Primer3 software was used to design primers ranging in size from 18 to 22 mer. The temperature ranged from 57 °C to 63 °C. There was one GC clamp, and primer amplification products ranged from 250 to 350 bp. Each thirty individuals of both *Pinus* species were tested for the species-specific DNA markers. The individuals of *P. densiflora* were sampled from thirteen populations in South Korea (Ahn *et al*., 2015), and the individuals of *P. sylvestris* were sampled from the 22 provenance in Sweden, which were introduced into Korea in order to select superior provenances that are well adapted to Korean environment (Ryu *et al*., 2013). It is also stored in the DNA Bank of Forest Genetic Resources Department (NIFS_122059323 to 122059343). PCR reaction mixtures contained 10 pmol of each primers pairs, 25 ng total DNA, 0.5 µl of 10 mM dNTPs, 2.5 µl of 10× reaction buffer (2.5 mM $MgCl_2$, 20 mM Tris-HCl [pH 8.4], 50 mM KCl), and 1 U Taq DNA polymerase (BioFACT, Daejeon, Korea). Reactions were performed on a GeneAmp PCR System 9700 thermal cycler (Applied Biosystems, Foster City, CA, USA) under the following conditions: 94 °C for 5 min; 45 cycles of 94 °C for 30 s, 60 °C for 30 s, and 72 °C for 30 s; and 72 °C for 10 min. PCR products were confirmed on a 2 % agarose gel, and were digested using 5 U of *Hinf*I, *BsaW*I and *Sph*I followed by electrophoresis on a 2 % acrylamide gel. DNA fragment sizes were estimated by comparison with 100 bp Plus Ladder (Thermo Fisher Scientific, Waltham, MA, USA).

Results

Chloroplast genome assembly and features

The complete chloroplast genome of *P. densiflora* was determined to be a circular double-stranded DNA sequence of 119,835 bp (GenBank accession number: MF990371). The genome showed a typical quadripartite structure including LSC (65,896 bp) and SSC (53,219 bp) regions and IRs (360 bp) (Fig. 1). The genome had a similar GC content to that of *P. sylvestris*. The GC content was the highest in the SSC region (39.4%), moderate in the LSC region (37.8%), and the lowest in the IR region (34.7%). We identified 113 genes including 74 encoding proteins and 36 and four encoding tRNA and rRNA, respectively. Twelve genes (six protein-coding and six tRNA genes) contained one intron, while two of the protein-coding genes (*ycf3* and *rps12*) had two introns. We also confirmed that *trnS*-GCU and *psaM* were duplicated in two chloroplasts (Table 1).

Analyses of repetitive sequences

We detected SSRs > 8 bp in *P. densiflora* and *P. sylvestris* chloroplast genomes according to a previously published method (Qian *et al*., 2013). We set the threshold based on the fact that SSRs > 8 bp are prone to strand slippage and mispairing, which is thought to be the primary mechanism underlying the high rate of polymorphism. In our analysis, there were 103 and 106 SSRs accounting for 1,236 bp in *P. densiflora* and 1,254 bp in *P. sylvestris*, respectively. These included 18 mono-, five di-, one tri-, four tetra-, and 75 hexa-nucleotide repeats for *P. densiflora* and 21 mono-, seven di-, one tri-, four tetra-, and 73 hexanucleotide repeats for *P. sylvestris*. Hexanucleotide repeats accounted for 72.8% and 68.9% of total SSRs in *P. densiflora* and *P. sylvestris*, respectively. The majority of mononucleotide SSRs were thymine and adenine. The majority of the identified repeats were located in the non-coding regions (intergenic spacers and introns) except 17 protein-coding genes (*matK, atpH, atpI, rpoC2, rpoB, petL, psbL, petA, rbcL, atpB, rpl22, psbC, rrn16, rrn23, ycf1, rpl32*, and *ycf2*). In total, 35 long repeat sequences > 25 bp were identified in the *P. densiflora* chloroplast genome, including 21 forward and 14 palindromic matches. Thirty long repeat sequences were identified in *P. sylvestris*, including 16 forward, one reverse, and 13 palindromic matches (Fig. 2).

Comparison of indels and SNPs in Pinus species

In total, 171 indels were found to differ between *P. densiflora* and *P. sylvestris*, most of which were located in intergenic spacer regions (91.2%), with 69.6% and 30.4% in the LSC and SSC, respectively. The average Indel length was 6 bp, and the longest was located in *cemA-ycf4* and *trnE-clpP*. The frequency of 1 bp Indels was 36.7%, while 30% was > 8 bp. In addition, Indels were detected in two coding genes of both species (*psaM* and *ycf2*; Table 2). Seventy-three SNPs differed between *P. densiflora* and *P. sylvestris*, of which 46 were transversions (63%). In total, 34 (46.6%) SNPs were located in coding regions, whereas 39 (53.4%) were in intergenic spacer regions or introns (Table 3).

The *rpoC1* and *trnM-trnV* gene regions were amplified by PCR (using the primers shown in Table 4) in order to validate their capacity to distinguish between the two *Pinus* species. The amplification product obtained using the Pdest-cp1 primer was digested with *Hinf*I and visualized by agarose gel electrophoresis. Approximately 200 bp fragment was observed in *P. densiflora* but not in *P. sylvestris* (Fig. 3). On the other hand, digestion of the Pidest-cp2 amplification product with *BsaW*I yielded approximately 200 bp fragment that was observed in *P. sylvestris* but not in *P. densiflora* (Fig. 4). In addition, digesting the Pidest-cp2 amplification product with *Sph*I produced a fragment of approximately 200 bp that was observed in *P. densiflora* but not in *P. sylvestris* (Fig. 5).

Discussion

Comparison of the complete plastid genomes of P. densiflora and P. sylvestris

IRs are known to stabilize the plastid genome because of its low base exchange rate and high copy-correcting activity. Thus, the loss of IRs can result in the shortening of intergenic spaces, gene loss, and structural variations in plastids. Reductions in Irs have been observed in most gymnosperm and in some legumes. In *P. densiflora*, the total chloroplast genome was 77 bp longer than that of *P. sylvestris,* while the gene content, order, and orientation were similar to those of ther *Pinus* chloroplast genomes (Wakasugi *et al*.,1994; Duan *et al*., 2016; Fang *et al*., 2016; Celiński *et al*., 2017; Ni *et al*., 2017).

We also found that SSRs of 1–6 bp per unit—known as, microsatellites— were distributed throughout the *P. densiflora* chloroplast genome. SSRs are important molecular markers of genomic variation within species or populations due to their high polymorphism, and have been extensively used to analyze plant population structure, diversity, differentiation and fertility (Kim & Kim, 2016). The SSRs detected in the present study will provide basic information for future analyses of genetic diversity in Pinaceae.

Identification molecular markers for distinguish between Pinus species based on SNPs

P. sylvestris and *P. densiflora* are difficult to distinguish because they form separate strongly supported groups in molecular phylogenetic studies (Wang *et al*., 1999; Gernandt, 2005; Hong *et al*., 2014) and also have very similar morphology (Lee, 2003; Hong *et al*., 2014).

Complete chloroplast genome sequences (plastomes) have been very useful for understanding phylogenetic relationships in angiosperms at the family level and above and have been used to resolve previously recalcitrant nodes (Barrett *et al.*, 2016).

In order to identify molecular markers that can be used to distinguish the two *Pinus* species, we compared their chloroplast genomes and found a total of 171 indels and 73 SNPs. We amplified the *rpoC1* and *trnM-trnV* gene regions by PCR-RFLP and found that SNPs in these two regions could be clearly distinguished by restriction enzyme digestion.

The differences between individuals can be clearly detected by separating differently sized fragments based on the single nucleotide and Indel polymorphisms of the restriction enzyme sites rather than by using dominant markers (Lee *et al*., 2012). In this context, the co-dominant CAPS markers developed in this study provide a means of unambiguously identifying *P. densiflora* and *P. sylvestris*.

It is likely that the SNP loci found in this study exist in other *Pinus* species. These markers have many research and commercial applications, including studies of genetic diversity, breeding, and species identification for the timber market.

Conclusions

This study provides the complete chloroplast sequences of *P. densiflora*. These sequences revealed significant similarity in the structural organization of the chloroplast genomes in Pinaceae, such as loss of IR and reduction of *ndh* genes, i.e., all *ndh* genes were transferred to the nucleus or that NADH dehydrogenase is not essential in pine chloroplasts. In addition, molecular markers that can distinguish between the phylogenetically and morphologically similar *P. sylvestris* and *P. densiflora* were identified, providing a more objective and reliable method of identification than that by conventional visual identification methods. Further, the data generated here can be used to develop additional molecular markers for comparing with other *Pinus* species. These markers can also have research and commercial applications such as in genetic diversity studies, breeding, and identification of species for the timber market.

Further research is necessary to determine whether the restriction site differences occur across the entire geographic range of both species, given that additional mutations may have caused one or more of the restriction sites to disappear in some populations.

References

Agarwal M, Shrivastava N, Padh H. 2008. Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Reports*, 27: 617–631. DOI: 10.1007/s00299-008-0507-z

Ahn, J. Y., Hong, K. N., Lee, J. W., Hong, Y. P., & Kang, H. 2015. Genetic Variation of *Pinus densiflora* Populations in South Korea Based on ESTP Markers. *Korean Journal of Plant Resources*, *28*(2), 279-289.

Celiński K, Kijak H, Barylski J, Grabsztunowicz M, Wojnicka-Półtorak A, Chudzińska E. 2017. Characterization of the complete chloroplast genome of *Pinus uliginosa* (Neumann) from the *Pinus mugo* complex. *Conservation Genetics Resources*, 9: 209–212. DOI: 10.1007/s12686-016-0652-6.

Duan RY, Yang LM, Lv T, Wu GL, Huang MY. 2016. The complete chloroplast genome sequence of *Pinus dabeshanensis*. *Conservation Genetics Resources*, 8: 395–397. DOI: 10.1007/s12686-016-0567-2

Dyall SD, Brown MT, Johnson PJ. 2004. Ancient invasions: from endosymbionts to organelles. *Science*, 304: 253–257. DOI: 10.1126/science.1094884

Farjon, A. (2010). *A Handbook of the World's Conifers (2 vols.)*(Vol. 1). Brill.

Fang MF, Wang YJ, Zu YM, Dong WL, Wang RN, Deng TT, Li ZH. 2016. The complete chloroplast genome of the Taiwan red pine *Pinus taiwanensis* (Pinaceae). *Mitochondrial DNA Part A*, 27: 2732–2733. DOI: 10.3109/19401736.2015.1046169

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, D. S., G. Lopez, S. O. Garcia and A. Liston. 2005. Phylogeny and classification of *Pinus*. *Taxon* 54: 29-42.

Hong JK, Yang JC, Lee YM, Kim JH. 2014. Molecular phylogenetic study of *Pinus* in Korea based on chloroplast DNA *psbA-trnH* and *atpF-H* sequences data. *Korean Journal of Plant Taxonomy*, 44: 111–118. DOI: 10.11110/kjpt.2014.44.2.111

Howe CJ, Barbrook AC, Koumandou VL, Nisbet RER, Symington HA, Wightman TF. 2003. Evolution of the chloroplast genome. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 358: 99-107. DOI: 10.1098/rstb.2002.1176

Jansen RK, Ruhlman TA. 2012. Plastid genomes of seed plants. In: Bock R, Knoop V, eds. *Genomics of chloroplasts and mitochondria*. Dordrecht: Springer Netherlands. 103–126. DOI: 10.1007/978-94-007-2920-9_5

Jansen RK, Raubeson LA, Boore JL, Chumley TW, Haberle RC, Wyman SK, Kuehl JV. 2005.

Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods in Enzymology* 395: 348–384. DOI: 10.1016/S0076-6879(05)95020-9

Katoh K, Misawa K, Kuma KI, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30: 3059–3066. DOI: 10.1093/nar/gkf436

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Thierer T. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647–1649. DOI: 10.1093/bioinformatics/bts199

Kim JB, Kim ES, Lim JH. 2017. Topographic and meteorological characteristics of *Pinus densiflora* dieback areas in Sogwang-Ri, Uljin. *Korean Journal of Agricultural and Forest Meteorology* 19: 10–18. DOI: 10.5532/KJAFM.2017.19.1.10

Kim SC, Kim JS, Kim JH. 2016. Insight into infrageneric circumscription through complete chloroplast genome sequences of two *Trillium* species. *AoB Plants* 8. DOI: 10.1093/aobpla/plw015

Korea National Arboretum and The Plant Taxonomic Society of Korea. 2007. A Synonymic List of Vascular Plants in Korea. Korea National Arboretum. Pocheon. (in Korean)

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research* 29: 4633–4642. DOI: 10.1093/nar/29.22.4633

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9: 357–359. DOI: 10.1038/nmeth.1923

Lee CS, Kim JH, Yi H, You YH. 2004. Seedling establishment and regeneration of Korean red pine (*Pinus densiflora* S. et Z.) forests in Korea in relation to soil moisture. *Forest Ecology and Management* 199: 423–432. DOI: 10.1016/j.foreco.2004.05.053

Lee JW, Bang KH, Kim YC, Seo AY, Jo IH, Lee JH, Cho JH. 2012. CAPS markers using mitochondrial consensus primers for molecular identification of *Panax* species and Korean ginseng cultivars (*Panax ginseng* CA Meyer). *Molecular Biology Reports* 39: 729–736. DOI: 10.1007/s11033-011-0792-4

Lee TB. 2003. Coloured Flora of Korea. Hyangmunsa, Seoul (2003) (in Korean)

Lin CP, Huang JP, Wu CS, Hsu CY, Chaw SM. 2010. Comparative chloroplast genomics reveals the evolution of Pinaceae genera and subfamilies. *Genome Biology and Evolution* 2: 504–517. DOI: 10.1093/gbe/evq036

Lohse M, Drechsel O, Bock R. 2007. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics* 52: 267–274. DOI: 10.1007/s00294-007-0161-y

Mayer C. 2010. Phobos Version 3.3.12. A tandem repeat search program, 20.

Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences of the United States of America* 104: 19363–19368. DOI: 10.1073/pnas.0708072104

Ni Z, Ye Y, Bai T, Xu M, Xu LA. 2017. Complete chloroplast genome of *Pinus massoniana* (Pinaceae): Gene rearrangements, loss of ndh genes, and short inverted repeats contraction, expansion. *Molecules* 22: 1528. DOI: 10.3390/molecules22091528

Qian J, Song J, Gao H, Zhu Y, Xu J, Pang X, Liu J. 2013. The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza. PLoS One* 8: e57607. DOI: 10.1371/journal.pone.0057607

Rasmussen HB. 2012. Restriction fragment length polymorphism analysis of PCR-amplified fragments (PCR-RFLP) and gel electrophoresis-valuable tool for genotyping and genetic fingerprinting. In: Magdeldin S, ed. *Gel Electrophoresis – Principles and Basics.* InTech. DOI: 10.5772/37724

Ravi V, Khurana JP, Tyagi AK, Khurana P. 2008. An update on chloroplast genomes. *Plant Systematics and Evolution* 271: 101–122. DOI: 10.1007/s00606-007-0608-0

Ryu, K. O., Han, M. S., Kim, I. S., Lee, J. H., & Lee, J. C. 2013. Adaptation Test of Scotch Pine (*Pinus sylvestris* L.) in Korea-Thirty-six-year-old Growth Performance of Twenty-two Provenances. *Korean Journal of Plant Resources*, *26*(1), 26-35.

Schattner P, Brooks AN, Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research* 33: W686–W689. DOI: 10.1093/nar/gki366

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

Wakasugi, T., Tsudzuki, J., Ito, S., Nakashima, K., Tsudzuki, T., & Sugiura, M. 1994 Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. *Proceedings of the National Academy of Sciences*, *91*(21), 9794-9798.

Wang, X. Q., Tank, D. C., & Sang, T. (2000). Phylogeny and divergence times in Pinaceae: evidence from three genomes. *Molecular Biology and Evolution*, 17(5), 773-781.

Wang, X. R., Y. Tsumura, H. Yoshimaru, K. Nagasaka and A. E. Szmidt. 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: 1742-1753.

Wu CS, Chaw SM. 2014. Highly rearranged and size□variable chloroplast genomes in conifers II clade (cupressophytes): evolution towards shorter intergenic spacers. *Plant Biotechnology*

Journal 12: 344–353. DOI: 10.1111/pbi.12141

Wu CS, Lin CP, Hsu CY, Wang RJ, Chaw SM. 2011. Comparative chloroplast genomes of Pinaceae: insights into the mechanism of diversified genomic organizations. *Genome Biology and Evolution* 3: 309–319. DOI: 10.1093/gbe/evr026

Yi DK, Choi K, Joo M, Yang JC, Mustafina FU, Han JS, Lee YM. 2016. The complete chloroplast genome sequence of *Abies nephrolepis* (Pinaceae: Abietoideae). *Journal of Asia-Pacific Biodiversity* 9: 245–249. DOI: 10.1016/j.japb.2016.03.014

Figure 1

Gene maps and summary of the *Pinus densiflora* S. *et* Z. chloroplast genome.

Genes lying outside the circle are transcribed in a clockwise direction, whereas genes inside are transcribed in a counterclockwise direction. Different colors denote known functional groups. The GC and AT contents of the genome are denoted by dashed darker and lighter gray in the inner circle. LSC, SSC, and IR indicate large single-copy, small single-copy, and inverted repeat regions, respectively.

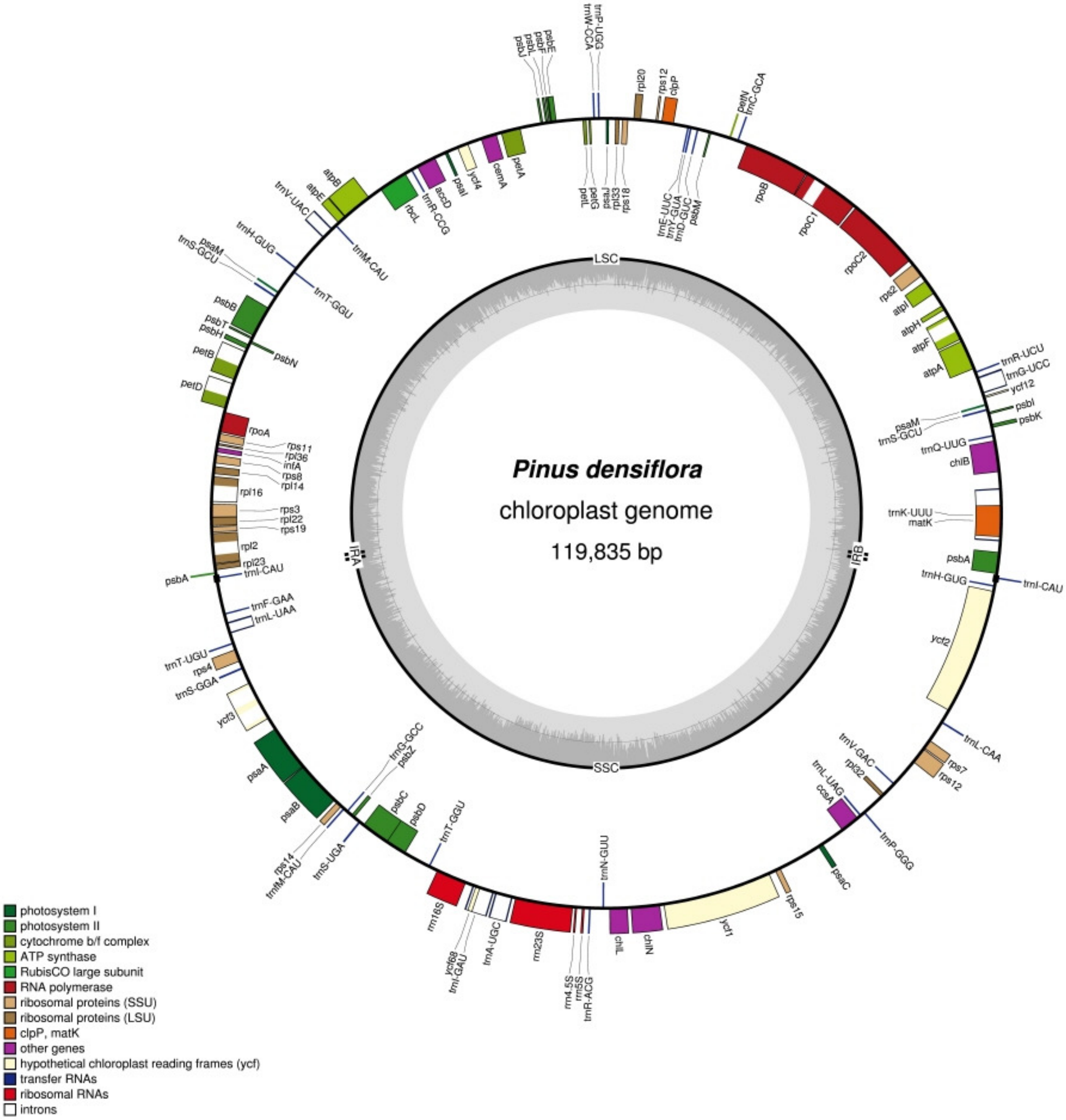
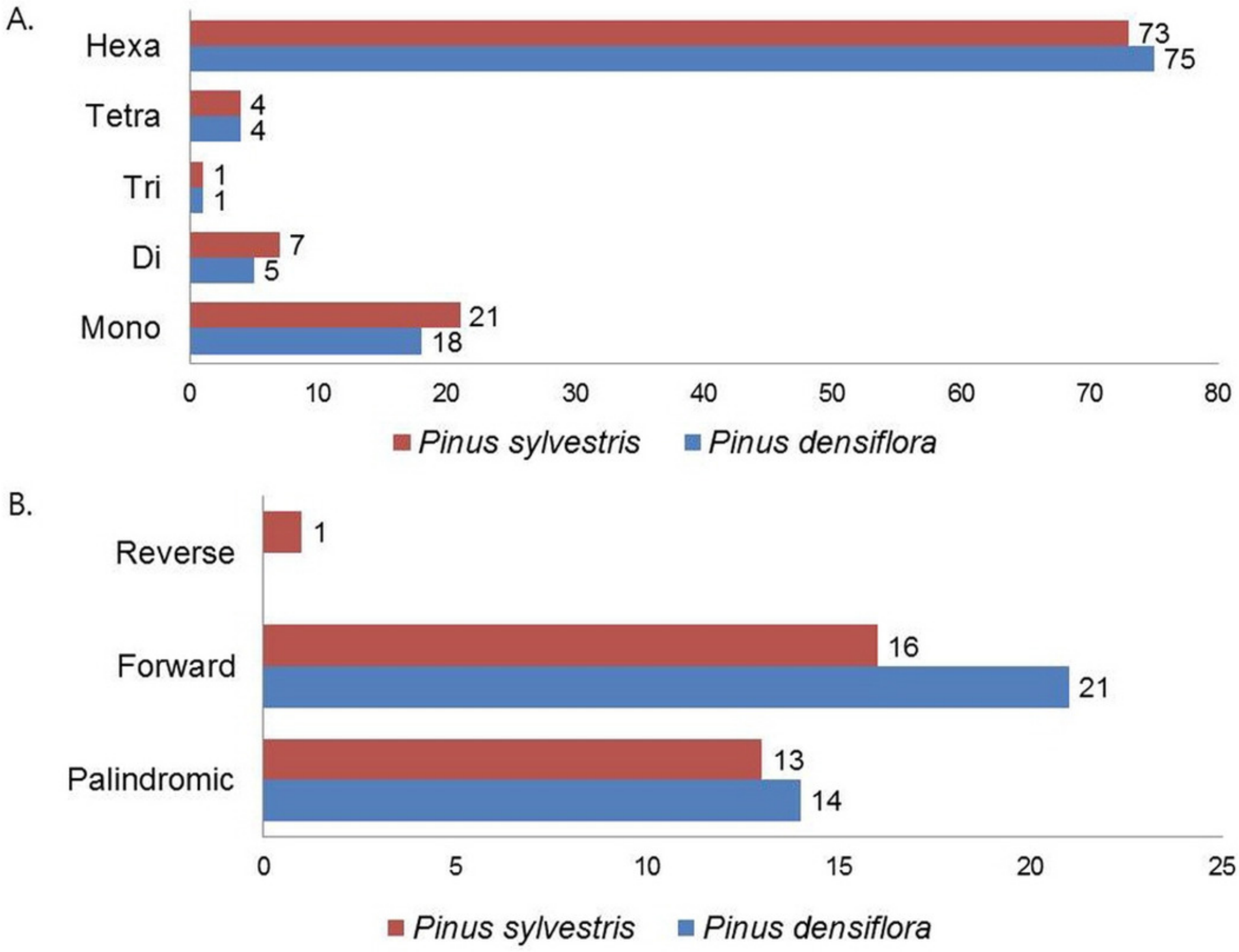

Pinus densiflora
chloroplast genome
119,835 bp
LSC
IRB
SSC
IRA
psbJ
psbL
psbF
psbE
trnW-CCA
trnP-UGG
rpl20
rps12
clpP
petN
trnC-GCA
petL
petG
psaJ
rpl33
rps18
trnE-UUC
trnY-GUA
trnD-GUC
psbM
rpoB
rpoC1
rpoC2
rps2
atpI
atpH
atpF
atpA
trnR-UCU
trnG-UCC
ycf12
psbI
psbK
psaM
trnS-GCU
trnQ-UUG
chlB
trnK-UUU
matK
psbA
trnH-GUG
trnI-CAU
ycf2
trnL-CAA
rps7
rps12
trnV-GAC
rpl32
trnL-UAG
ccsA
trnP-GGG
psaC
rps15
ycf1
chlN
chlL
trnN-GUU
trnR-ACG
rrn5S
rrn4.5S
rrn23S
trnA-UGC
trnI-GAU
ycf68
rrn16S
trnT-GGU
psbD
psbC
psbZ
trnG-GCC
trnS-UGA
trnfM-CAU
rps14
psaB
psaA
ycf3
trnS-GGA
rps4
trnT-UGU
trnL-UAA
trnF-GAA
trnI-CAU
psbA
rpl23
rpl2
rps19
rpl22
rps3
rpl16
rpl14
rps8
infA
rpl36
rps11
rpoA
petD
petB
psbN
psbH
psbT
psbB
trnS-GCU
psaM
trnH-GUG
trnT-GGU
trnM-CAU
trnV-UAC
atpE
atpB
rbcL
trnR-CCG
accD
psaI
ycf4
cemA
petA
photosystem I
photosystem II
cytochrome b/f complex
ATP synthase
RubisCO large subunit
RNA polymerase
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
other genes
hypothetical chloroplast reading frames (ycf)
transfer RNAs
ribosomal RNAs
introns

Figure 2

Distribution of repeats present in *Pinus* chloroplast genomes. A) Distribution of SSRs present in *Pinus* chloroplast genomes. B) Distribution of long repeat sequences present in *Pinus* chloroplast genomes.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26506v1 | CC BY 4.0 Open Access | rec: 12 Feb 2018, publ: 12 Feb 2018

Figure 3

Relevant part of the SNP multiple sequence alignment in the *trnM-trnV* gene regions.

The recognition site of *Hinf*I restriction enzyme (G/ANTC) is altered by one SNP at position 84 (A/G transition). A fragment degraded to 198 bp was observed In *P. densiflora* (1-9), but not in *P. sylvestris* (10-21).

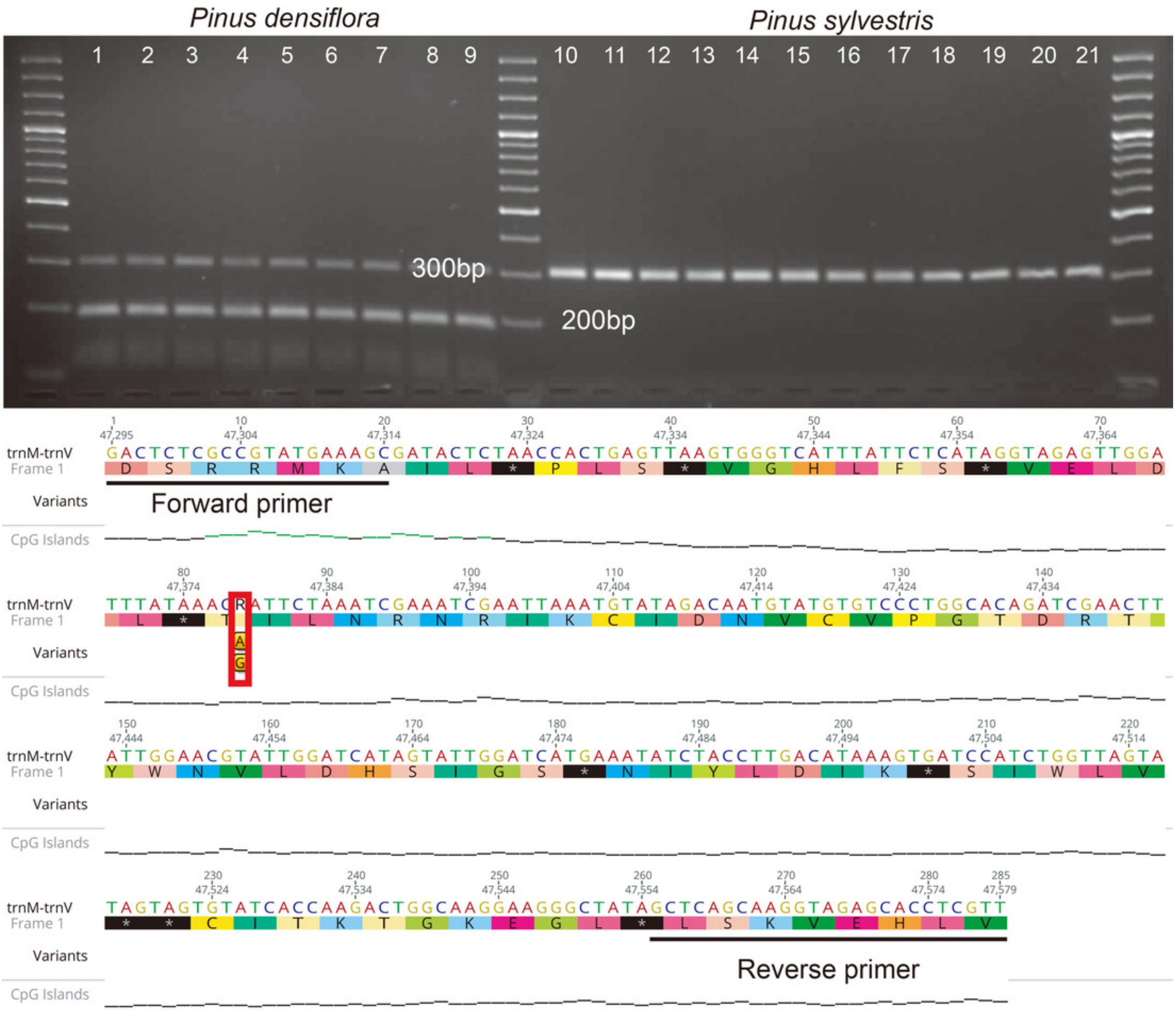


Figure 4

Relevant part of the SNP multiple sequence alignment in the *rpoC1* gene.

The recognition site of *BsaW*I restriction enzyme (W/CCGGW) is altered by one SNP at position 184 (A/G transition). A fragment degraded to 179 bp was observed In *P. sylvestris* (10-21), but not in *P. densiflora* (1-9).

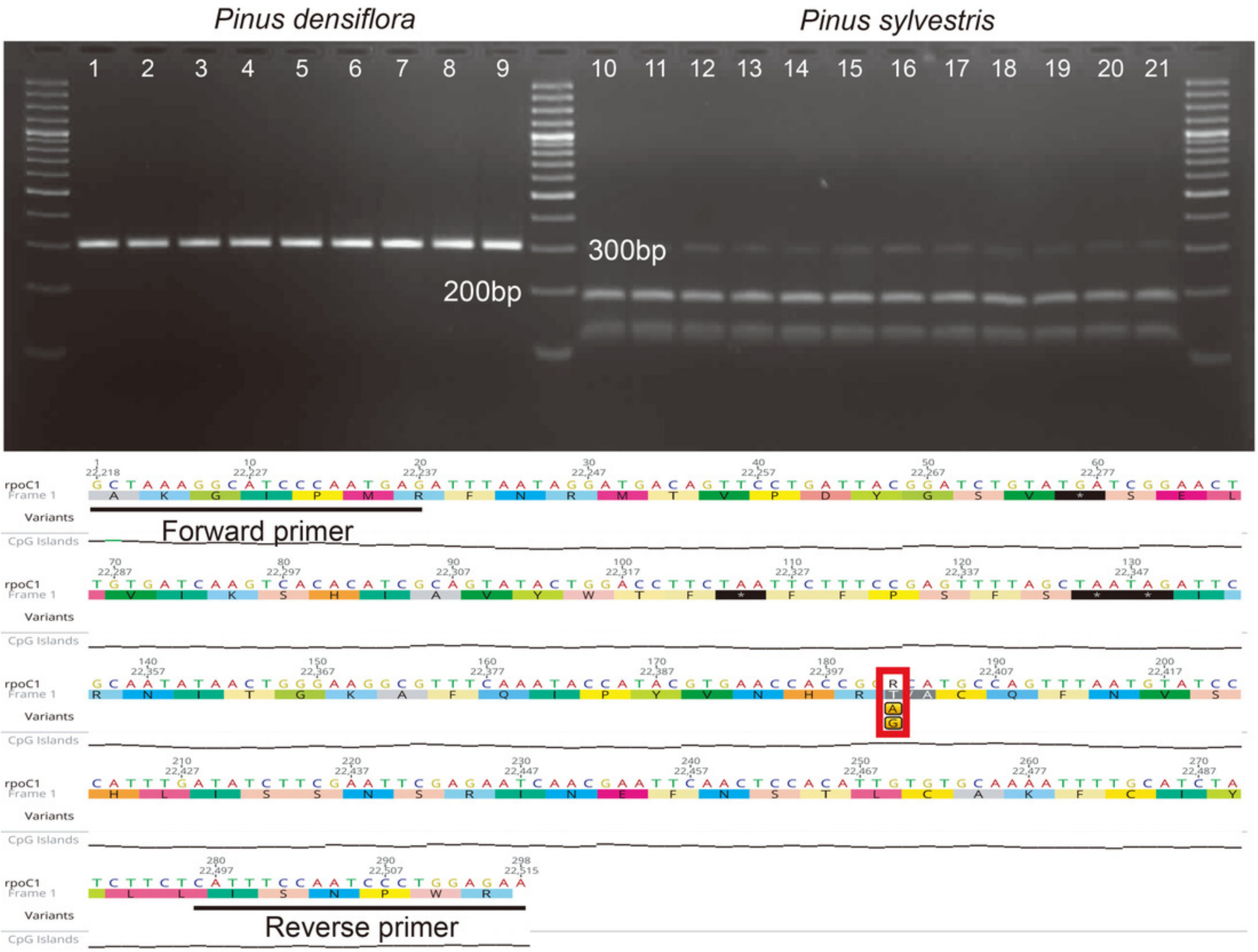

Pinus densiflora
Pinus sylvestris
1 2 3 4 5 6 7 8 9
10 11 12 13 14 15 16 17 18 19 20 21
300bp
200bp
rpoC1
Frame 1
Variants
CpG Islands
GCTAAAGGCATCCCAATGAGATTTAATAGGATGACAGTTCCTGATTACGGATCTGTATGATCGGAACT
TGTGATCAAGTCACACATCGCAGTATACTGGACCTTCTAATTCTTTCCGAGTTTTAGCTAATAGATTC
GCAATATAACTGGGAAGGCGTTTCAAATACCATACGTGAACCACCGG R CATGCCAGTTTAATGTATCC
A
G
CATTTGATATCTTCGAATTCGAGAATCAACGAATTCAACTCCACATTGTGTGCAAAATTTTGCATCTA
TCTTCTCATTTCCAATCCCTGGAGAA
Forward primer
Reverse primer

Figure 5

Relevant part of the SNP multiple sequence alignment in the *rpoC1* gene.

The recognition site of *Sph*I restriction enzyme (GCATG/C) is altered by one SNP at position 184 (A/G transition). A fragment degraded to 184 bp was observed In *P. densiflora* (1-9), but not in *P. sylvestris* (10-21).

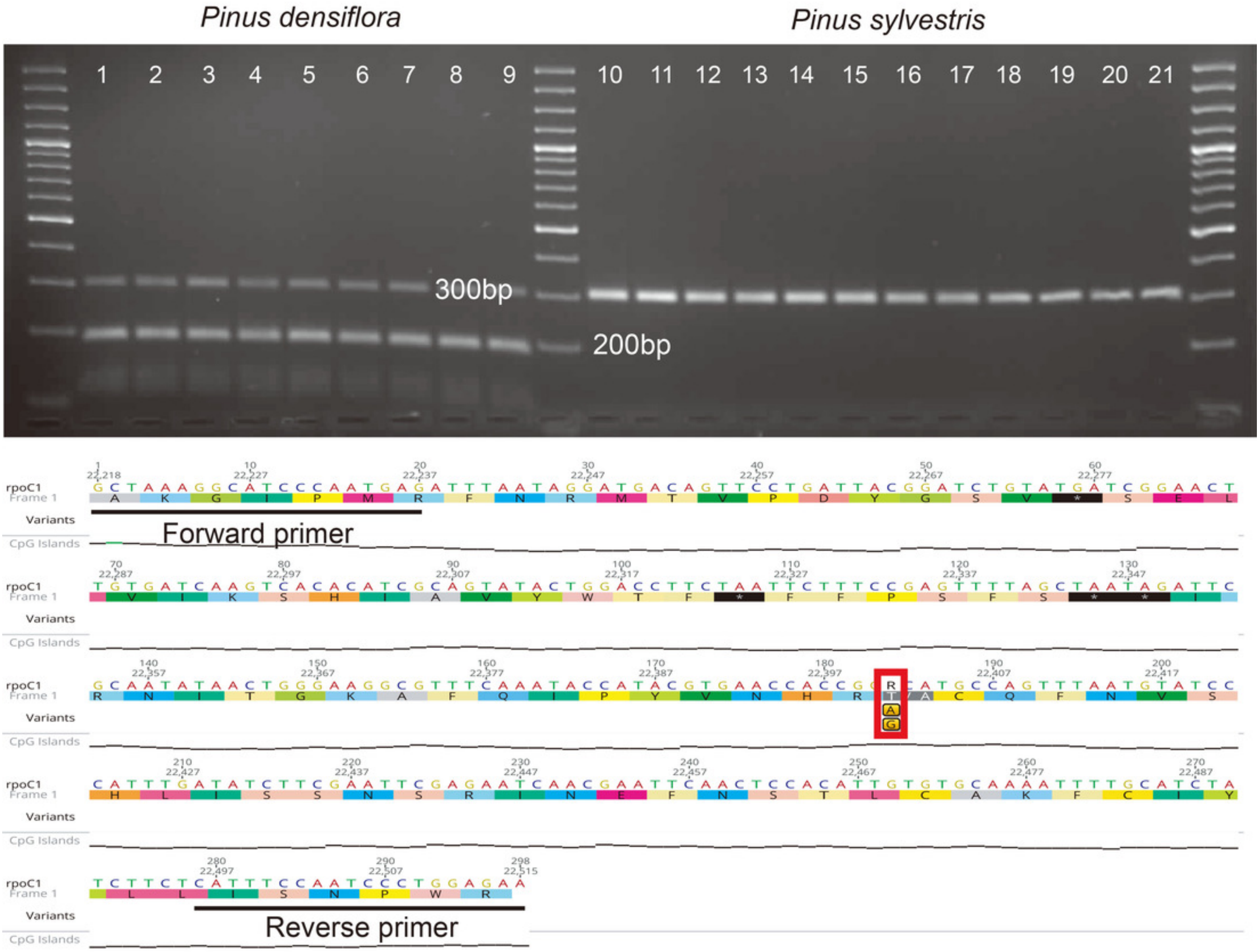

Pinus densiflora
Pinus sylvestris
1 2 3 4 5 6 7 8 9
10 11 12 13 14 15 16 17 18 19 20 21
300bp
200bp
rpoC1
Frame 1
Variants
CpG Islands
GCTAAAGGCATCCCAATGAGAGATTTAATAGGATGACAGTTCCTGATTACGGATCTGTATGATCGGAACT
TGTGATCAAGTCACACATCGCAGTATACTGGACCTTCTAATTCTTTCCGAGTTTTAGCTAATAGATTC
GCAATATAACTGGGAAGGCGTTTCAAATACCATACGTGAACCACCG R ATGCCAGTTTAATGTATCC
A
G
CATTTGATATCTTCGAATTCGAGAATCAACGAATTCAACTCCACATTGTGTGCAAAATTTTGCATCTA
TCTTCTCATTTCCAATCCCTGGAGAA
Forward primer
Reverse primer

Table 1(on next page)

List of genes encoded by the *P. densiflora* chloroplast genome.

Table 1. List of genes encoded by the *P. densiflora* chloroplast genome.

Gene types	Gene products	
Ribosomal RNAs	*rrn4.5, rrn5, rrn16, rrn23*	4
Transfer RNAs	*trnA*-UGC[a], *trnC*-GCA, *trnD*-GUC, *trnE*-UUC, *trnF*-GAA, *trnfM*-CAU, *trnG*-GCC, *trnG*-UCC[a], *trnH*-GUG, *trnH*-GUG, *trnI*-CAU(x2), *trnI*-GAU[a], *trnK*-UUU[a], *trnL*-CAA, *trnL*-UAA[a], *trnL*-UAG, *trnM*-CAU, *trnN*-GUU, *trnP*-GGG, *trnP*-UGG, *trnQ*-UUG, *trnR*-ACG, *trnR*-CCG, *trnR*-UCU, *trnS*-GCU, *trnS*-GCU, *trnS*-GGA, *trnS*-UGA, *trnT*-GGU, *trnT*-GGU, *trnT*-UGU, *trnV*-GAC, *trnV*-UAC[a], *trnW*-CCA, *trnY*-GUA	36
Photosystem I	*psaA, psaB, psaC, psaI, psaJ, psaM*(x2)	7
Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ*	15
Cytochrome b/f complex	*petA, petB*[a], *petD*[a], *petG, petL, petN*	6
ATP synthase	*atpA, atpB, atpE, atpF*[a], *atpH, atpI*	6
Large subunit of rubisco	*rbcL*	1
Chloroplast envelope membrane protein	*cemA*	1
Large subunit ribosomal proteins	*rpl2*[a], *rpl14, rpl16*[a], *rpl20, rpl22, rpl23, rpl32, rpl33, rpl36*	9
Small subunit ribosomal proteins	*rps2, rps3, rps4, rps7, rps8, rps11, rps12*[b], *rps14, rps15, rps18, rps19*	11
RNA polymerase	*rpoA, rpoB, rpoC1*[a], *rpoC2*	4
Translational initiation factor	*infA*	1
Subunit of acetyl-CoA-carboxylase	*accD*	1
C-type cytochrome synthesis gene	*ccsA*	1
Maturase	*matK*	1
Chlorophyll biosynthesis	*chlB, chlL, chlN*	3
ATP-dependent protease	*clpP*	1
Conserved open reading frames	*ycf1, ycf2, ycf3*[b], *ycf4, ycf12, ycf68*	6
Total		114

[a]: Gene containing a single intron.

[b]: Gene containing two introns.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26506v1 | CC BY 4.0 Open Access | rec: 12 Feb 2018, publ: 12 Feb 2018

3

Table 2(on next page)

Distribution of Indels in *Pinus* chloroplast genomes

P. densiflora	*P. sylvestris*	Minimum	Maximum	Length	Loci
	T	1376	1376	1	*psbA-trnK*
CAG		8212	8214	3	*psaM*
	GGG	9850	9852	3	*prnG-trnR*
	CC	15202	15203	2	*atpI-rps2*
	TCTA	15236	15239	4	*atpI-rps2*
CTATTTCTCAAGA		16150	16162	13	*rps2-rpoC2*
A		26128	26128	1	*rpoB-trnC*
ATTTAAATAATTTTGATAATTTTAATT		29221	29247	27	*trnE-clpP*
	G	30329	30329	1	*clpP-rps12*
TATTTTCTTC		36493	36502	10	*psbJ-petA*
AATTTCAATAAATATTTCATTGTATGAAAATGG		39594	39626	33	*cemA-ycf4*
T		41146	41146	1	*psaI-accD*
	TTTTTTTATTT	45123	45133	11	*rbcL-atpB*
CTG		51621	51623	3	*psaM*
T		51963	51963	1	*trnS-psbB*
	T	63817	63817	1	*rps19-rpl2*
	GC	65871	65872	2	*psbA-trnI*
	AA	68122	68123	2	*trnF-trnL*
CTCCCCTTCT		68911	68920	10	*trnL-trnT*
TTTTTTTT		72059	72066	8	*ycf3* intron
AT		73572	73572	2	*ycf3-psaA*
	C	82837	82837	1	*psbD-trnT*
CAATTTGTTGT		93896	93906	11	*chlL-chlN*
	T	101253	101253	1	*ycf1-rps15*
	T	102793	102793	1	*ndhI-ndhE*
A		108576	108576	1	*trnV-rps12*
	TCATA	109477	109481	5	*trnV-rps12*
	A	109734	109734	1	*trnV-rps12*
	AA	110195	110196	2	*rps12* intron
	AGAAAAAAA	115341	115349	9	*ycf2*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26506v1 | CC BY 4.0 Open Access | rec: 12 Feb 2018, publ: 12 Feb 2018

Table 3(on next page)

Distribution of SNPs in *Pinus* chloroplast genomes

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26506v1 | CC BY 4.0 Open Access | rec: 12 Feb 2018, publ: 12 Feb 2018

P. densiflora	*P. sylvestris*	Seqeunce No.	Mutation	Loci	*P. densiflora*	*P. sylvestris*	Seqeunce No.	Mutation	Loci
A	G	4190	transition	*trnK-chlB*	A	T	48380	transversion	*trnV-trnH*
A	C	5708	transversion	*chlB*	G	T	59337	transversion	*rps11*
G	T	7608	transversion	*psbK-psbI*	C	T	62762	transition	*rps3*
A	T	7613	transversion	*psbK-psbI*	G	A	69894	transition	*rps4*
A	G	8182	transition	*trnS-psaM*	G	T	79253	transversion	*trnfM-psbZ*
T	A	8183	transversion	*trnS-psaM*	G	T	80551	transversion	*psbC*
C	T	8184	transition	*trnS-psaM*	G	A	82941	transition	*psbD-trnT*
A	G	9855	transition	*trnG-trnR*	G	C	83659	transversion	*psbD-trnT*
A	G	9857	transition	*trnG-trnR*	A	T	93066	transversion	*chlL*
A	G	10538	transition	*atpA*	G	A	95123	transition	*chlN*
T	G	14210	transversion	*atpH-atpI*	C	A	96483	transversion	*ycf1*
A	C	15041	transversion	*atpI-rps2*	C	A	96893	transversion	*ycf1*
C	A	22063	transversion	*rpoC1 intron*	T	G	97302	transversion	*ycf1*
G	A	22401	transition	*rpoC1*	T	G	97763	transversion	*ycf1*
C	A	22743	transversion	*rpoC1-rpoB*	C	T	99931	transition	*ycf1*
T	A	29216	transversion	*trnE-clpP*	G	T	100184	transversion	*ycf1*
G	A	29217	transition	*trnE-clpP*	G	A	100412	transition	*ycf1*
A	T	29274	transversion	*trnE-clpP*	A	G	100435	transition	*ycf1*
A	T	29283	transversion	*trnE-clpP*	T	A	100541	transversion	*ycf1*
A	T	29288	transversion	*trnE-clpP*	A	C	100552	transversion	*ycf1*
A	G	29289	transition	*trnE-clpP*	C	A	101591	transversion	*rps15-ndhH*
T	A	30313	transversion	*clpP-rps12*	C	A	103342	transversion	*ndhE-psaC*
C	A	30314	transversion	*clpP-rps12*	T	G	108063	transversion	*rpl32-trnV*
T	C	31522	transition	*rpl20*	C	A	108688	transversion	*trnV-rps12*
T	G	33070	transversion	*psaJ-trnP*	G	T	110180	transversion	*rps12 intron*
G	T	34025	transversion	*petL*	A	C	111793	transversion	*ndhB*
A	C	34211	transversion	*petL-psbE*	T	G	111948	transversion	*ndhB-trnL*
A	G	42181	transition	*accD*	C	A	112340	transversion	*trnL-ycf2*
C	T	42668	transition	*accD-trnR*	C	A	113412	transversion	*ycf2*
A	G	42992	transition	*trnR-rbcL*	C	A	115338	transversion	*ycf2*
G	A	43272	transition	*rbcL*	T	C	115351	transition	*ycf2*
G	A	43740	transition	*rbcL*	A	T	115354	transversion	*ycf2*
C	T	43762	transition	*rbcL*	A	G	115791	transition	*ycf2*
G	C	44583	transversion	*rbcL-atpB*	A	C	116833	transversion	*ycf2*

C	T	46241	transition	*atpB*	G	T	118062	transversion	*ycf2*
G	A	47378	transition	*trnM-trnV*	T	G	118560	transversion	*ycf2*
T	A	48379	transversion	*trnV-trnH*					

Table 4(on next page)

Primers and restriction enzymes used for identification of *Pinus* species

	forward primer	reverse primer	enzyme	Loci
Pdest-cp1	GACTCTCGCCGTATGAAAGC	GCAAGGTAGAGCACCTCGTT	*Hinf*I	*rpoC1*
Pdest-cp2	GCTAAAGGCATCCCAATGAG	TTCTCCAGGGATTGGAAATG	*BsaW*I, *Sph*I	*trnM-trnV*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26506v1 | CC BY 4.0 Open Access | rec: 12 Feb 2018, publ: 12 Feb 2018