AN ABSTRACT OF THE DISSERTATION OF

Matthew Benjamin Parks for the degree of Doctor of Philosophy in Botany and Plant Pathology presented on May 26, 2011.

Title: Plastome Phylogenomics in the Genus *Pinus* Using Massively Parallel Sequencing Technology.

Abstract approved:

_____

Aaron I. Liston Richard C. Cronn

This thesis summarizes work completed over the previous four years primarily focusing on chloroplast phylogenomic inquiry into the genus *Pinus* and related Pinaceae outgroups using next-generation sequencing on Illumina platforms. During the time of our work, Illumina sequence read lengths have essentially been limited to 25 to 100 base pairs, presenting challenges when trying to assemble genomic space featuring repetitive regions or regions divergent from established reference genomes. Our assemblies initially relied on previously constructed high quality plastome sequences for each of the two *Pinus* subgenera, yet we were able to show clear negative trends in assembly success as divergence from reference sequences. This was most evident in assemblies of Pinaceae outgroups, but the trend was also apparent within *Pinus* subgenera. To counter this problem, we used a combination of *de novo* and reference-guided assembly approaches, which allowed us to more effectively assemble highly divergent regions.

From a biological standpoint, our initial focus was on increasing phylogenetic resolution by using nearly complete plastome sequences from select *Pinus* and Pinaceae outgroup species. This effort indeed resulted in greatly increased phylogenetic resolution as evidenced by a nearly 60-fold increase in parsimony informative positions in our dataset as compared to previous datasets comprised of only several chloroplast loci. In addition, bootstrap support levels across the resulting phylogenetic tree were consistently high, with ≥95% bootstrap

support at 30/33 ingroup nodes in maximum likelihood analysis. A positive correlation between the length/amount of sequence data applied to our phylogeny and overall bootstrap support values was also supported, although trends indicated some nodes would likely remain recalcitrant even with the application of complete plastomes. This correlation was important to demonstrate, as it was reflective of trends seen in a meta-analysis of contemporary, infrageneric chloroplast-based phylogenies. In addition, our meta-analysis indicated that most researchers rely on relatively small regions of the chloroplast genome in these studies and obtain relatively little in resolution and support in resulting phylogenies. Clearly, the application of plastome sequences to these types of analyses has great potential for increasing our understanding of evolutionary relationships at low taxonomic levels.

An unexpected finding of this work involved two putative protein-coding regions in the chloroplast, *ycf*1 and *ycf*2, which featured strongly elevated rates of mutation, and together accounted for over half of exon parsimony informative sites although making up only 22% of exon sequence length. Of these two loci, clearly *ycf*1 was more problematic to assemble from short read data, as it featured numerous indels as well as several repetitive regions. We designed primers based on conserved regions allowing essentially complete amplification of this locus and sequenced the *ycf*1 locus (with Sanger technology) for a representative of each of the 11 *Pinus* subsections, using accessions from the previous study. Importantly, these primers were also effective across Pinaceae and should facilitate future work throughout the family.

Accessions with full *ycf*1 sequences were in turn utilized as subsectional references as we sequenced and assembled plastomes for most of the remaining *Pinus* species. To efficiently produce these sequences, we relied on a solution-based hybridization strategy developed by Richard Cronn to enrich preparations of total genomic DNA for chloroplast-specific DNA. While the phylogenetic results of a full-plastome, full-genus analysis were certainly of interest, our final focus was on the investigation of ‘noise’ in our dataset, and whether it affected phylogenetic conclusions drawn from the plastome. To determine this, we explored the removal of variable sites from our alignment and the resultant effect on topology and resolution. This allowed us to identify a window of alignment partitions in which nodal

bootstrap support remained high across the genus, yet sufficient noise was removed to identify important patterns in the positioning of three clades with historically problematic phylogenetic positioning.

Specific Chapter Copyrights

Parks, M., Liston, A. and Cronn, R. 2010. MEETING THE CHALLENGES OF NON-REFERENCED GENOME ASSEMBLY FROM SHORT-READ SEQUENCE DATA. Acta Hort. (ISHS) 859:323-332. http://www.actahort.org/books/859/859_38.htm

Parks, M., Liston, A. and R. Cronn. 2011. NEWLY DEVELOPED PRIMERS FOR COMPLETE YCF1 AMPLIFICATION IN PINUS (PINACEAE) CHLOROPLASTS WITH POSSIBLE FAMILY-WIDE UTILITY. American Journal of Botany (in press).

Plastome Phylogenomics in the Genus *Pinus* Using Massively Parallel Sequencing Technology

by
Matthew Benjamin Parks

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented May 26, 2011
Commencement June 2012

Doctor of Philosophy dissertation of Matthew Benjamin Parks presented on May 26, 2011.

APPROVED:

_____

Co-Major Professor, representing Botany and Plant Pathology

_____

Co-Major Professor, representing Botany and Plant Pathology

_____

Chairperson of the Department of Botany and Plant Pathology

_____

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

_____

Matthew Benjamin Parks, Author

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation for the support from many friends and colleagues both within the Botany and Plant Pathology department and elsewhere. I would also like to express my gratitude to my committee members Joseph Spatafora, Dee Denver and Dave Myrold for their input and advice, as well as to Chris Sullivan of the CGRB, whose expertise in computing and computing infrastructure has been invaluable.

Especial thanks are deserved for the unfailing support of my family, including my father and stepmother, Jerry and Lin Parks, my brother, Nathan Parks, and my fiancée, Ariadne Luh. This work also would not have been possible without the continual support of my advisors, Aaron Liston and Richard Cronn, whose effort, enthusiasm and constructive critique for and of my work I hope has paid dividends. Finally, this thesis is dedicated to my father and the memory of my mother, Ann Huenemann Parks, who instilled in me integrity, honesty, and a great desire for education.

CONTRIBUTION OF AUTHORS

Dr. Aaron Liston and Dr. Richard Cronn were deeply involved in all aspects of the research presented in the present thesis, including study design, the development of novel laboratory and bioinformatic procedures, data collection and analysis, and editing of manuscripts. Certainly without their input and efforts this work would not have been possible.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

LIST OF APPENDIX FIGURES

LIST OF APPENDIX TABLES

Plastome Phylogenomics in the Genus *Pinus* Using Massively Parallel Sequencing Technology

GENERAL INTRODUCTION

The genus *Pinus* is an intriguing taxon for biological inquiry on many levels. From an ecological perspective, pines are important and often dominant components of the northern boreal forests, are able to exploit challenging environments, and are often important colonizers after disturbance events (Richardson and Rundel 1998). From an evolutionary standpoint, *Pinus* is notable in that the genus has a history of over 100 million years, a relatively rich fossil record, and encompasses over one sixth of extant gymnosperms. *Pinus* thereby represents an important link between seedless vascular plants and flowering plants, and can offer important insight into past ecosystems and evolutionary events. Not surprisingly, there is a substantial history of inquiry into the genus *Pinus* and its evolutionary relationships, including early to contemporary morphological analyses (Engelmann 1880, Shaw 1914, 1924, Pilger 1926, Gaussen 1960, Little and Critchfield 1969, Van der Berg 1973, Farjon 1984, Frankis 1993, Ickert-Bond 2001, Gernandt et al. 2005, Klymiuk et al. 2011), extensive crossability studies largely performed by Elbert Little and William Critchfield (Critchfield 1963, Critchfield 1966, Little and Critchfield 1969, Critchfield 1975, 1986), and a number of molecular-based studies (Liston et al. 1999, Wang et al. 1999, Wang et al. 2000, Geada Lopez et al. 2002, Liston et al. 2003, Zhang and Li 2004, Gernandt et al. 2005, Syring et al. 2005, Eckert and Hall 2006, Gernandt et al. 2009, Palmé et al. 2009). Studies based on morphology (of both extant and extinct species) and crossability have provided valuable information in determining the relationships between *Pinus* species, yet both are limited tools in comparison to molecular data. Specifically, morphological characters are problematic in the pines due to extensive homoplasy (Gernandt et al. 2005), while crossability studies provide only a rough estimate of species relationships by relying on a single measured character and provide little information outside of the currently accepted *Pinus* subsections (Gernandt et al. 2005). Although previous *Pinus* molecular studies have sampled very limited amounts of nuclear and chloroplast genomes, theoretically these genetic compartments should be much more information-rich than other types of data, as they contain tens of thousands to billions of measurable sites (depending which genome is investigated). In addition, these sites are subject to a range of evolutionary pressures largely dependent on their position in the genome (for

example coding vs. non-coding regions, loci under positive vs. negative selection) and should more effectively mark evolutionary events. The majority of molecular phylogenetic analyses in *Pinus* have relied at least in part on loci residing in the chloroplast genome (plastome), a situation broadly reflective of plant phylogenetic analyses. The plastome is a reasonable target for plant molecular phylogenetic studies for a number of reasons. For instance, an overall moderate mutation rate (Wolfe et al. 1987, Palmer 1990) makes determination of orthology straightforward even when considering distantly related taxa. On the other hand, a haploid state and uniparental inheritance result in a decreased effective population size compared to nuclear markers (Birky 1978, Birky et al. 1983), thereby increasing the chance of capturing signal during divergence events. In addition, while the chloroplast genome technically represents a single linkage group, there is a diversity in evolutionary rates of change across the plastome (Graham and Olmstead 2000, Shaw 2005, Shaw et al. 2007) allowing the chloroplast to be applied to different levels of phylogenetic inquiry. Finally, the genome sizes of chloroplasts are much smaller and less variable than are those of plant nuclear genomes, making full plastome sequences a more tractable target for sequencing and analysis. For comparison, the average nuclear genome size in *Pinus* is around 29 billion base pairs (Grotkopp et al. 2004) and consists of perhaps 80% repetitive elements (Kriebel 1985, Kovach et al. 2010), while typical plastome size is closer to 115-120000 base pairs, about half of which accounts for the ca. 130 protein and RNA molecules encoded by the chloroplast (Wakasugi et al. 1994, Parks et al. 2009).

The most recent genus-wide molecular taxonomy of *Pinus* was completed in 2005 (Gernandt et al. 2005) and was based on just over 2800 aligned base pairs of chloroplast sequence from over 100 species of pine. This work clarified many of the broad relationships within *Pinus* (in terms of chloroplast evolutionary history), supporting the division of the genus into two subgenera, four sections and 11 subsections (Figure 1.1). Notably, most species-rich clades failed to resolve internally, resulting in extensive polytomies within subsections. Resolution in several enigmatic clades was also lacking or problematic. For example, the morphologically unique *Pinus krempfii* (subsection *Krempfianae*) showed strong support for inclusion within section *Quinquefoliae*, but its relation to subsections *Strobus* and *Gerardianae* was not able to be determined (Figure 1.1). The Southeast Asian species *Pinus merkusii*, on the other hand, was found to have relatively strong support for inclusion within subsection *Pinus* of section

Pinus, although its morphology suggests a stronger affinity for subsection *Pinaster* of the same section (Frankis 1993). Similarly, subsection *Contortae* resolved with high support as sister to the clade of subsections *Ponderosae* and *Australes* within section *Trifoliae* (Figure 1.1), yet a relatively shallow fossil record and capability to hybridize with members of subsection *Australes* support a much more recent derivation of the *Contortae* within the section (Gernandt et al. 2005).

Considering the relatively small portion of the chloroplast genome sampled prior to our work, it was reasonable to expect that more extensive sampling of the plastome could result in greater resolution of the infrageneric relationships in *Pinus*, and a more accurate understanding of the phylogenetic positions of problematic taxa like subsections *Krempfianae* and *Contortae*, and *Pinus merkusii*. However, until relatively recently the sequencing capacity necessary to efficiently produce full plastomes was substantially hindered by a combination of the high per base pair sequencing cost and low throughput per sequencing run that are associated with traditional Sanger sequencing, as well as difficulty in isolating large or numerous genomic targets in preparation for sequencing. The development and commercial availability of 'next-generation' or massively parallel sequencing technologies over the last 5-10 years, however, have very nearly turned sequence-based research on its head, such that the expectation of cheaply and quickly sequencing large or numerous genomic targets from many samples is now the norm rather than the exception (Mardis 2008, Shendure and Ji 2008, Mamanova et al. 2010, Mardis 2011). Plant systematics as a field has been somewhat slow to harness the high-throughput capacity of massively parallel sequencers. In addition to the work presented herein, currently only a relative handful of the ca. 200 published chloroplast genomes have been sequenced using next-generation technology rather than traditional Sanger sequencing-based approaches (Moore et al. 2006, Moore et al. 2007, Asif et al. 2010, Atherton et al. 2010, Tangphatsornruang et al. 2010, Yang et al. 2010, Doorduin et al. 2011, Jansen et al. 2011, Nock et al. 2011, Shulaev et al. 2011, Straub et al. 2011). Of the chloroplast genomes currently sequenced using massively parallel sequencing, most have utilized the Roche/454 pyrosequencing technology, in part because it features longer reads than other next-generation platforms. Nonetheless, the most powerful platforms to date are designed by Illumina, with sequence output on the order hundreds of millions to several billions of base pairs per sequencing run on the early Illumina GA (Mardis 2008), to hundreds of billions of base pairs

on the latest full-capacity model, the HiSeq2000 (Schweiger et al. 2011). However, all next-generation platforms to date are limited to some degree by the short length of their sequence reads. This is particularly true for the Illumina platforms, reads lengths from which essentially ranged from 25-100 base pairs during the time our work was performed. Due to this limitation, a number of effective 'short-read' assemblers and aligners have been developed in the last decade. For example, assembly programs such as Velvet (Zerbino and Birney 2008), EDENA (Hernandez et al. 2008), ABySS (Simpson et al. 2009) and Euler-SR (Chaisson and Pevzner 2008) utilize various strategies to assemble short sequence reads into longer, 'de novo' contigs. These assemblers rely on overlapping k-mers of some length shorter than the read length to assemble large contigs without the aid of a reference sequence. On the other hand, reference-guided assemblers, such as Bowtie (Langmead et al. 2009), SOAP (Li et al. 2008), BWA (Li and Durbin 2009), and YASRA (Ratan 2009) align reads to a specified reference sequence in order to assemble short reads into longer contigs. Although most reference-guided assemblers are utilized for resequencing projects in which a closely-related genome is available to be used as a reference, YASRA employs a combination of de novo and reference-guided assembly in an iterative fashion in order to more effectively assemble short read sequences across divergent regions of the reference genome.

Considering the exponential growth in sequencing throughput and its potential applications to phylogenetic pursuits, we attempted to apply massively parallel sequencing to more effectively resolve the evolutionary relationships within the genus *Pinus*. As this project was begun in the early stages of the development of massively parallel sequencers, our original challenge was simply to efficiently sequence and assemble pine plastome sequences using the Illumina GA platform and its associated 25-40 base pair reads. This proved a challenging, but not insurmountable task, and was greatly aided by the development of a multiplexing strategy that allowed multiple individual samples to be run in single lanes of the Illumina flowcell (Cronn et al. 2008). Still, initial plastome assemblies featured some poorly assembled or incomplete regions, primarily due either to short read lengths (problematic in repetitive or highly divergent regions) or the polymerase chain reaction- (PCR-) based amplification strategy we employed (problematic at primer junctions) (Cronn et al. 2008). In particular, assemblies proved difficult as divergence increased from either of our two established *Pinus* reference genomes, and were least effective as we moved outside of the genus into the broader

Pinaceae family. Nonetheless, we were still able to assemble 37 mostly complete plastome sequences using our initial strategies, as detailed in Chapter II of this thesis. For our assemblies, we employed a combination of de novo and reference-guided approaches using the de novo assemblers Velvet and Edena (Hernandez et al. 2008, Zerbino and Birney 2008) and an in-house reference-guided assembler RGA (Shen and Mockler). While not as efficient as our later approaches, this pipeline nonetheless aided in assembling accessions divergent from our references.

The remainder and large part of our work was somewhat analogous to how a traditional ecologist might investigate a novel habitat or ecosystem, as it involved a mixture of exploration, documentation, and testing. Our primary goal, as mentioned above, was to document the utility of the plastome in resolving the interspecific relationships within the genus *Pinus*. Based on alignment of our initial 37 assemblies and two established reference genomes, we found that plastome sequences indeed resulted in substantial increases in both phylogenetic resolution and bootstrap support as compared to previous efforts based on smaller portions of the chloroplast genome (Wang et al. 1999, Gernandt et al. 2005, Eckert and Hall 2006), although for some relationships, such as those resulting from rapid radiations, it appeared unlikely that even the application of complete plastome sequences would result in full resolution. Regardless, using a meta-analysis of contemporary, chloroplast-based phylogenetic analyses, we were able to show that similar gains in resolution and support likely would be found in most plant genera by applying full plastome sequences. These results are the topic of Chapter III of this thesis.

Through the course of these analyses, the putative protein-coding chloroplast loci *ycf*1 and *ycf*2 were identified as highly variable when compared to all other chloroplast protein-coding loci, and likely targets of positive selection. This is particularly the case with *ycf*1, which features not only numerous variable positions, but several stretches of repetitive units and many apparent insertion and deletion mutations. In part due to our findings, *ycf*1 has now been used to help further untangle the complex relationships of subsection *Ponderosae* (Gernandt et al. 2009) and is serving as a DNA ‘barcode’ locus for validating species identities in commercially available pine nuts potentially linked to dysgeusia (Handy et al. 2011). The highly divergent nature of *ycf*1 also led us to develop a novel set of primers that allowed

complete amplification of this large (ca. 5.5-6 kbp) locus, and subsequent resequencing and validation using Sanger-sequencing technology. This was performed for representatives of all *Pinus* subsections, and helped us to create accurate plastome references for every subsection. These primers also appear to be effective in the broader Pinaceae, so it is likely that they will have applications in genera beyond *Pinus*. The development and application of these primers is the topic of Chapter IV of this thesis.

In our final sequencing efforts, resulting in ca. 70 new plastome assemblies, increased read lengths and the replacement of PCR-based chloroplast enrichment with a solution-based hybridization approach greatly improved both the success of our assemblies and the efficiency of our sequencing, such that we were able to produce nearly complete plastome sequences for almost all remaining species of pines in a relatively short period of time. In addition, a greatly improved assembly pipeline and more complete subsectional references allowed for more accurate and complete plastome assemblies. Assembly and alignment of plastome sequences for nearly all of the world's pines species, as predicted by our initial 37 assemblies, did indeed result in greatly increased resolution and an unprecedented view of the interspecific relationships within *Pinus*. Nonetheless, as seen in our earlier analyses, some relationships still failed to resolve with high support. This was most evident in species-rich subsections (subsections *Australes* and *Quinquefoliae*, for example), where decreased support values were clearly clustered in regions with short branch lengths and putative rapid divergence events. In these cases, it is likely that the plastome may not contain sufficient signal to resolve all relationships within the genus with high confidence. In other cases, however, it is possible that the plastome contains misleading phylogenetic signal, or 'noise', that has sufficient presence to influence or even override lesser amounts of accurate signal. To investigate this possibility, we removed the most variable sites in our alignment in 100 base pair partitions, and followed the effect of this removal on topology and support in three taxa mentioned previously with historically unresolved or contentious phylogenetic positioning – subsection *Contortae*, consisting of four North American species, subsection *Krempfianae*, consisting of the morphologically distinct *Pinus krempfii*, and a clade of two closely related southeast Asian pines, *Pinus merkusii* and *Pinus latteri*. In each of these cases, previous, chloroplast-based phylogenetic positioning is either weakly supported and/or contentious due to conflicting positioning based on alternative data, such as nuclear sequence, morphology, or the fossil

record. In all three cases, there appeared to be a fairly strong signature of phylogenetic noise, although the response to the removal of noise was not the same in all three taxa. For both *P. krempfii* and the clade of *P. merkusii* / *P. latteri*, removal of highly variable sites resulted in a much more highly supported positioning that also reflected conclusions based on non-chloroplast data. At the same time, support for the positioning of subsection *Contortae* was substantially reduced as highly variable sites were removed, suggesting that 'noise' in the chloroplast genome may be a major contributor to a putatively incorrect phylogenetic positioning. The results of this study are the topic of the final research chapter of this thesis, Chapter V.

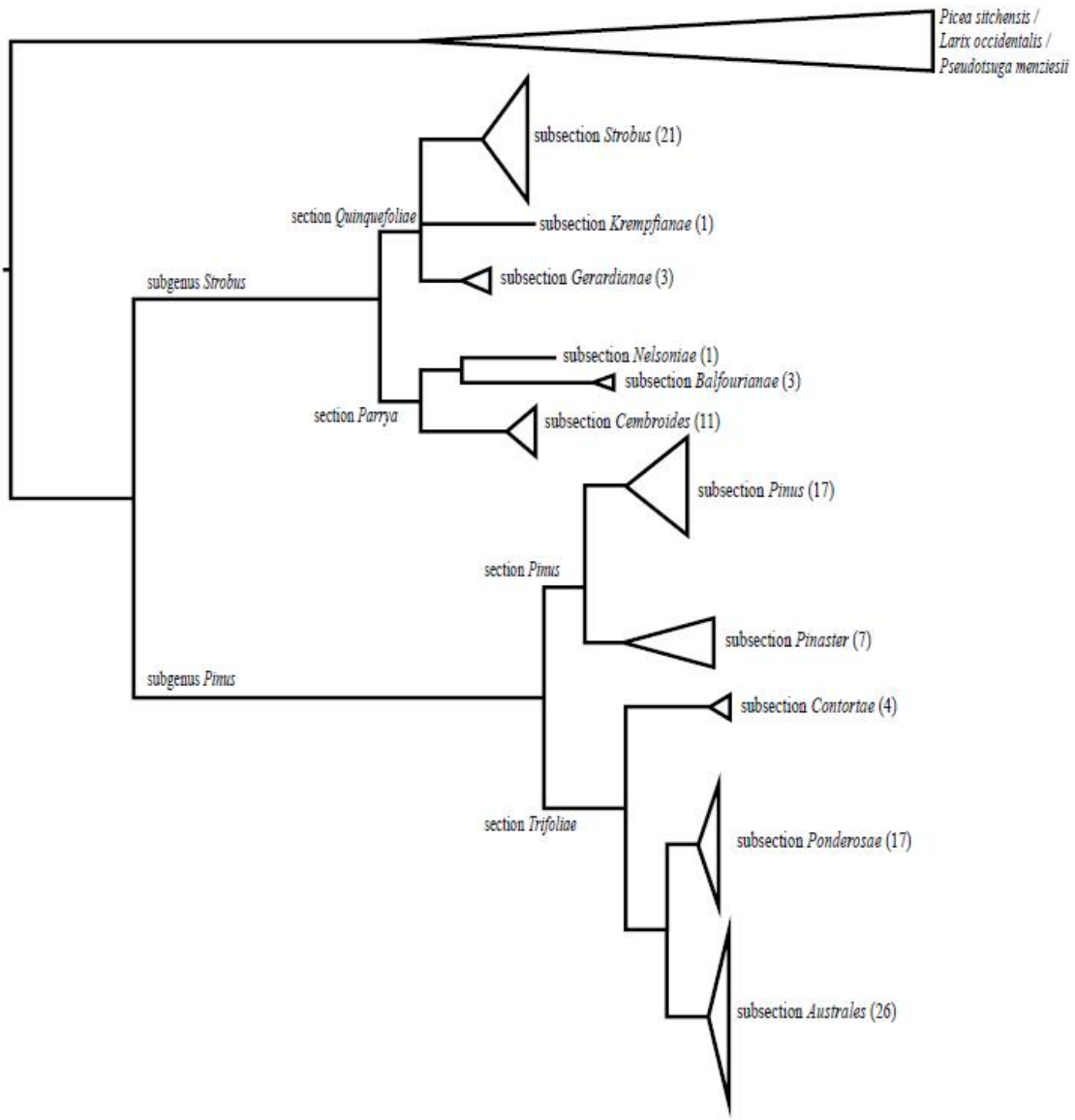


Figure 1.1. Systematic subdivisions of the genus *Pinus*. Phylogenetic positioning and branch lengths are based on chloroplast sequence data from Gernandt et al. (2005) and the present work. Numbers in parentheses indicate the number of species in each subsection as described in Gernandt et al. (2005).

LITERATURE CITED

Asif M, Mantri S, Sharma A, Srivastava A, Trivedi I, Gupta P, Mohanty C, Sawant S, Tuli R. 2010. Complete sequence and organisation of the *Jatropha curcas* (Euphorbiaceae) chloroplast genome. Tree Genetics & Genomes, 6:941-952.

Atherton R, McComish B, Shepherd L, Berry L, Albert N, Lockhart P. 2010. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. Plant Methods, 6:22.

Birky CW, Jr. 1978. Transmission genetics of mitochondria and chloroplasts. Annu. Rev. Genet., 12:471-512.

Birky CW, Jr., Maruyama T, Fuerst P. 1983. An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts, and some results. Genetics, 103:513-527.

Chaisson MJ, Pevzner PA. 2008. Short read fragment assembly of bacterial genomes. Genome Res., 18:324-330.

Critchfield WB. 1963. Hybridization of the southern pines in California. Southern Forest Tree Improvement Committee Publications, 22:40-48.

Critchfield WB. 1966. Crossability and relationships of the closed-cone pines. Silvae Genetica, 16:89-97.

Critchfield WB. 1975. Interspecific hybridization in *Pinus*: a summary review. In: Fowler DP, Yeatman CY editors. Symposium on Interspecific and Interprovenance Hybridization in Forest Trees, p. 99-105.

Critchfield WB. 1986. Hybridization and classification of the white pines (*Pinus* section *Strobus*). Taxon, 35:647-656.

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res., 36:e122.

Doorduin L, Gravendeel B, Lammers Y, Ariyurek Y, Chin-A-Woeng T, Vrieling K. 2011. The complete chloroplast genome of 17 individuals of pest species *Jacobaea vulgaris*: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. DNA Research, 18:93-105.

Eckert AJ, Hall BD. 2006. Phylogeny, historical biogeography, and patterns of diversification for *Pinus* (Pinaceae): phylogenetic tests of fossil-based hypotheses. Mol. Phylogenet. Evol., 40:166-182.

Engelmann G. 1880. Revision of the genus *Pinus*, and description of *Pinus elliottii*. Transactions of the Saint Louis Acaemy of Science, 4:161-189.

Farjon A. 1984. Pines: drawings and descriptions of the genus. Leiden, W. Backhuys.

Frankis MP. 1993. Morphology and affinities of *Pinus brutia*. International Symposium on *Pinus brutia* Ten. Ankara, Ministry of Forestry, p. 11-18.

Gaussen H. 1960. Les gymnospermes actuelles et fossiles. Fassicule VI. Les Coniferales. Chapter 11. Generalites, Genre *Pinus*. *Travaux du Toulous Universite Laboratoire Forestier*, p. 1-272.

Geada Lopez G, Kamiya K, Harada K. 2002. Phylogenetic relationships of diploxylon pines (subgenus *Pinus*) based on plastid sequence data. Int. J. Plant Sci., 163:737-747.

Gernandt DS, Hernández-León S, Salgado-Hernández E, Rosa JAPdl. 2009. Phylogenetic relationships of *Pinus* subsection *Ponderosae* inferred from rapidly evolving cpDNA regions. Syst. Bot., 34:481-491.

Gernandt DS, Lopez G, Garcia SO, Liston A. 2005. Phylogeny and classification of *Pinus*. Taxon, 54:29-42.

Graham SW, Olmstead RG. 2000. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. Am. J. Bot., 87:1712-1730.

Grotkopp E, Rejmánek M, Sanderson MJ, Rost TL, Soltis P. 2004. Evolution of genome size in pines (*Pinus*) and its life-history correlates: supertree analyses. Evolution, 58:1705-1729.

Handy SM, Parks M, Rader JI, Diachenko GW, Callahan J, Liston A, Deeds JR. 2011. Genetic identification of pine nuts obtained from consumers experiencing dysgeusia. Manuscript in preparation.

Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J. 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res., 18:802-809.

Ickert-Bond S. 2001. Reexamination of wood anatomical features in *Pinus krempfii* (Pinaceae). IAWA Journal, 22:355-365.

Jansen RK, Saski C, Lee S-B, Hansen AK, Daniell H. 2011. Complete plastid genome sequences of three rosids (*Castanea*, *Prunus*, *Theobroma*): evidence for at least two independent transfers of rpl22 to the nucleus. Mol. Biol. Evol., 28:835-847.

Klymiuk AA, Stockey RA, Rothwell GW. 2011. The first organismal concept for an extinct species of Pinaceae. Int. J. Plant Sci., 172:294-313.

Kovach A, Wegrzyn J, Parra G, Holt C, Bruening G, Loopstra C, Hartigan J, Yandell M, Langley C, Korf I, *et al.* 2010. The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics, 11:420.

Kriebel HB. 1985. DNA sequence components of the *Pinus strobus* nuclear genome. Can. J. For. Res., 15:1-4.

Langmead B, Trapnell C, Pop M, Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol., 10:R25.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25:1754-1760.

Li R, Li Y, Kristiansen K, Wang J. 2008. SOAP: short oligonucleotide alignment program. Bioinformatics, 24:713-714.

Liston A, Gernandt DS, Vining TF, Campbell CS, Pinero D. 2003. Molecular phylogeny of Pinaceae and *Pinus*. Acta Hort.:107-114.

Liston A, Robinson WA, Piñero D, Alvarez-Buylla ER. 1999. Phylogenetics of *Pinus* (Pinaceae) based on nuclear ribosomal DNA internal transcribed spacer region sequences. Mol. Phylogenet. Evol., 11:95-109.

Little EL, Critchfield WB. 1969. Subdivision of the genus *Pinus* (pines). In: U.S. Department of Agriculture FS editor. Washington, D.C., p. 1-51.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. 2010. Target-enrichment strategies for next-generation sequencing. Nat Meth, 7:111-118.

Mardis ER. 2008. Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics, 9:387-402.

Mardis ER. 2011. A decade's perspective on DNA sequencing technology. Nature, 470:198-203.

Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci USA, 104:19363.

Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Folta KM, Soltis DE. 2006. Rapid and accurate pyrosequencing of angiosperm plastid genomes. BMC Plant Biol., 6:17.

Nock CJ, Waters DLE, Edwards MA, Bowen SG, Rice N, Cordeiro GM, Henry RJ. 2011. Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnology Journal, 9:328-333.

Palmé AE, Pyhäjärvi T, Wachowiak W, Savolainen O. 2009. Selection on nuclear genes in a *Pinus* phylogeny. Mol. Biol. Evol., 26:893-905.

Palmer JD. 1990. Contrasting modes and tempos of genome evolution in land plant organelles. Trends Genet., 6:115-120.

Parks M, Cronn R, Liston A. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biology, 7:84.

Pilger R. 1926. Pinus. In: Engler A, Prantl K editors. Die Naturliche Pflanzenfamilien. Liepzig, Wilhelm Engelmann, p. 331-342.

Ratan A. 2009. Assembly algorithms for next-generation sequence data. *Dissertation*, The Pennsylvania State University.

Richardson DM, Rundel PW. 1998. Ecology and biogeography of *Pinus*: an introduction. In: Richardson DM editor. Ecology and Biogeography of *Pinus*. Cambridge, Cambridge University Press.

Schweiger M, Kerick M, Timmermann B, Isau M. 2011. The power of NGS technologies to delineate the genome organization in cancer: from mutations to structural variations and epigenetic alterations. Cancer and Metastasis Reviews, 30:199-210.

Shaw GR. 1914. The genus *Pinus*. Forage Village, The Murray Printing Co.

Shaw GR. 1924. Notes on the genus *Pinus*. Journal of the Arnold Arboretum 5:225-227.

Shaw J, Lickey EB, Schilling EE, Small RL. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. Am. J. Bot., 94:275.

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

Shen R, Mockler TC. RGA - a reference-guided assembler. http://rga.cgrb.oregonstate.edu/rga_about.html. Manuscript in preparation.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nat Biotech, 26:1135-1145.

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, *et al*. 2011. The genome of woodland strawberry (*Fragaria vesca*). Nat Genet, 43:109-116.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol İ. 2009. ABySS: A parallel assembler for short read sequence data. Genome Res., 19:1117-1123.

Straub SC, Fishbein M, Livshultz T, Foster Z, Parks M, Weitemier K, Cronn RC, Liston A. 2011. Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. BMC Genomics, 12:211.

Syring J, Willyard A, Cronn R, Liston A. 2005. Evolutionary relationships among *Pinus* (Pinaceae) subsections inferred from multiple low-copy nuclear loci. Am. J. Bot., 92:2086-2100.

Tangphatsornruang S, Sangsrakru D, Chanprasert J, Uthaipaisanwong P, Yoocha T, Jomchai N, Tragoonrung S. 2010. The chloroplast genome sequence of mungbean (*Vigna radiata*)

determined by high-throughput pyrosequencing: structural organization and phylogenetic relationships. DNA Research, 17:11-22.

Van der Berg J. 1973. Holzer der niederrheinischen braunkohlenformation 2. Holzer der braunkohlengruben "Maria Theresia" zu Herzogenrath, "Zukunft West" zu Eschweiler und "Victor" Zulpich mitte zu Zulpich. Nebst einer systematisch-anatomischen bearbeitung der gattung *Pinus* L. Review of Palaeobotany and Palynology, 15:73-275.

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*. Proc Natl Acad Sci USA, 91:9794-9798.

Wang XR, Szmidt AE, Nguyên HN. 2000. The phylogenetic position of the endemic flat-needle pine *Pinus krempfii* (Pinaceae) from Vietnam, based on PCR-RFLP analysis of chloroplast DNA. Plant Syst. Evol., 220:21-36.

Wang XR, Tsumura Y, Yoshimaru H, Nagasaka K, Szmidt AE. 1999. Phylogenetic relationships of Eurasian pines (*Pinus*, Pinaceae) based on chloroplast *rbc*L, *mat*K, *rpl*20-*rps*18 spacer, and *trn*V intron sequences. Am. J. Bot., 86:1742-1753.

Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Acad. Nat. Sci. Phila., 84:9054-9058.

Yang M, Zhang X, Liu G, Yin Y, Chen K, Yun Q, Zhao D, Al-Mssallem IS, Yu J. 2010. The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). PLoS ONE, 5:e12762.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res., 18:821-829.

Zhang ZY, Li DZ. 2004. Molecular phylogeny of section *Parrya* of *Pinus* (Pinaceae) based on chloroplast *mat*K gene sequence data. Acta Bot. Sin., 46:171-179.

Meeting the Challenges of Non-Referenced Genome Assembly from Short-Read Sequence Data

Matthew Parks, Aaron Liston and Richard Cronn

Acta Horticulturae (ISHS)
P.O. Box 500
3001 Leuven 1
Belgium
Acta Hort. 859: 323-332.

ABSTRACT

Massively parallel sequencing technologies (MPST) offer unprecedented opportunities for novel sequencing projects. MPST, while offering tremendous sequencing capacity, are typically most effective in resequencing projects (as opposed to the sequencing of novel genomes) due to the fact that sequence is returned in relatively short reads. Nonetheless, there is great interest in applying MPST to genome sequencing in non-model organisms. We have developed a bioinformatics pipeline to assemble short read sequence data into nearly complete chloroplast genomes using a combination of *de novo* and reference-guided assembly, while decreasing reliance on a reference genome. Initially, short read sequences are assembled into larger contigs using *de novo* assembly. *De novo* contigs are then aligned to the corresponding reference genome of the most closely related taxon available and merged to form a consensus sequence. The consensus sequence and reference are in turn 'merged' such that aligned *de novo* sequence remains unaffected while missing sequence is filled in using the reference sequence. This chimeric reference is then utilized in reference-guided assembly to align the original short read data, resulting in a draft plastome. Using two established *Pinus* reference plastomes, our method has been effective in the assembly of 33 chloroplast genomes within the genus *Pinus*, and results with four species representing other genera of Pinaceae suggest the method will be of general use in land plants, particularly once limitations of PCR-based chloroplast enrichment are overcome.

INTRODUCTION

High throughput sequencing technologies have increased DNA and RNA sequencing capacity by orders of magnitude within the last decade. For example, the first human genomic sequence, finished in the early part of this decade, took over a decade to produce at an estimated total cost of between 0.3 - 3 billion US dollars (Lander et al. 2001, Venter et al. 2001, Collins et al. 2004, Bentley et al. 2008). In contrast, resequencing of human genomes today can be completed in a matter of weeks, with the cost measured in tens to hundreds of thousands of dollars (Bentley et al. 2008, Wang et al. 2008, Ahn et al. 2009, Mardis et al. 2009), and trends suggest the rate of progress in sequencing capacity is still increasing (Gupta 2009). Currently at the forefront of sequencing efforts are massively parallel sequencing technologies (MPST). MPST platforms generate millions of short reads (currently 30-400 bp depending on the platform used (Simon et al. 2009)) in parallel during sequencing, which are then typically mapped back onto a previously sequenced reference genome to determine genomic sequence of sampled organismal or cellular lineages (Holt and Jones 2008, Ley et al. 2008, Wang et al. 2008, Mardis et al. 2009). Even considering the tremendous sequencing capacity of these technologies, challenges remain in their application to a broad range of sequencing projects. For example, sequence capacity measured in Gbp is clearly excessive for the sequencing of small genomes (such as bacterial, organellar or viral genomes). Thus far, this challenge has been approached through the development of multiplex strategies in which multiple accessions indexed by short barcode tags are sequenced simultaneously (Porreca et al. 2007, Craig et al. 2008, Cronn et al. 2008). Another, and perhaps more daunting challenge, lies in utilizing MPST to sequence the genomes of organisms lacking a closely related reference genome. Clearly, with more distantly related species, the likelihood for sequence divergence and genomic structural rearrangements increases. Utilizing short read sequence data in such cases can quickly become problematic, as divergence and rearrangement make it difficult or impossible to map short reads onto the genomes of distantly related references (Whiteford et al. 2005, Pop and Salzberg 2008). To counter this second problem, we have developed a short read assembly pipeline which transforms raw sequence data into genomic sequence in four basic steps: 1) *de novo* assembly of short read data into larger contigs, 2) alignment of *de novo* contigs to the most-closely related reference genome available, 3) formation of a chimeric reference using aligned *de novo* contigs, with gaps filled in by the reference genome, and 4) alignment of short read data to the chimeric assembly to form final

genomic contigs. While several commercial "all-in-one" software packages are currently available to serve a similar purpose, these tend to be fairly expensive (typically several thousands of US dollars). In contrast, our assembly pipeline can function entirely with open source software.

To date, we have used our pipeline to assemble 33 chloroplast genomes within the genus *Pinus* (32 of which lacked a same-species reference), as well as four chloroplast genomes of non-pine members of Pinaceae. All genomic sequencing was performed in multiplex (typically 4-6x) on the Illumina IG genome analyzer, and resulting genomes were estimated to average 92% complete.

MATERIALS AND METHODS

Sequence preparation

Amplification and sequence preparation followed Cronn et al. (Cronn et al. 2008).

Processing of raw sequence data

Microread sequence and quality files were converted from raw sequence output, sorted, binned and had their tags removed using custom perl scripts available at http://www.science.oregonstate.edu/~knausb/genomics_scripts/knaus_scripts.html.

Chloroplast genome assembly

Assembly from microreads to chloroplast genomes is described in detail elsewhere (Whittall et al. 2009). In brief, microreads from an accession were assembled into larger contigs using de novo assemblers, and aligned to the most closely related reference chloroplast genome available (Fig. 2.1A). A chimeric reference sequence was then created by merging aligned contigs with the reference genome, such that aligned de novo sequence persisted and reference sequence was used in areas missing de novo coverage (Fig. 2.1A). The accession's microreads were then aligned against this chimeric reference using a reference-guided assembler to form the contigs of the draft genome (Fig. 2.1B). These contigs were then checked for quality and manually edited also as previously described (Whittall et al. 2009).

RESULTS

Assemblies overall (including outgroups) averaged 92% complete (Fig. 2.2). Assemblies in subgenus *Strobus* averaged 117 kb, with an estimated 8.8% missing data (compared to *P. koraiensis* reference). Subgenus *Pinus* assemblies averaged just less than 120 kb (6% estimated missing data, compared to *P. thunbergii* reference). Outgroup assemblies averaged just over 119 kb (10.4% average estimated missing data compared to *P. thunbergii* reference). De novo assemblies ranged from 64% to over 97% of estimated plastome lengths (avg.=89.2% ± 7.0% standard deviation), while finished assemblies were slightly higher (92.2% ± 5.9% sd) as noted above (Table 2.1 for details). Our alignment of all assemblies was 132,715 bp in length, with slightly less than half (62,298 bp) from exons encoding 71 conserved protein coding genes (20,638 amino acids), 36 tRNAs and 4 rRNAs. A high degree of co-linearity is inferred for these genomes due to: 1) the absence of major rearrangements within *de novo* contigs, and 2) the overall success of the PCR-based sequence isolation strategy (indicating conservation of the order of anchor genes containing primer sites). Nonetheless, several known structural rearrangements, including a tandem duplication of *psb*A in *P. contorta* (Lidholm and Gustafsson 1991) and the apparent loss of duplicate copies of *psa*M and *rps*4 in *P. koraiensis*, could not be confirmed. Two loci, *ycf*1 and *ycf*2, stood out as highly variable regions among exons, accounting for 22% of all exon sequence but nearly 52% of exon variable sites. From our assemblies, these two loci also exhibit numerous indels in *Pinus*, although these are difficult to validate completely based on short-read assembly. Because of their variability, assembly success for protein-coding exons was determined both with and without these loci (see below).

Uncorrected p-distances between finished assemblies and their closest established references ranged greater than two orders of magnitude, from 0.000645 (76 total differences to reference in *P. thunbergii*) to 0.079221 (7615 total differences to reference in *Abies firma*) (Table 2.1). Assembly success generally was correlated weakly with divergence from reference and sequencing effort (here defined as microread count), although significant correlations were found in some cases (Table 2.2, Figure 2.3). Assembly success and divergence from reference were correlated negatively in subgenus *Pinus* and outgroup accessions; this correlation was positive in subgenus *Strobus* and when all pines were considered together (Fig. 2.3A, Table

2.2). Assembly success was correlated positively with sequencing effort (here defined as microread count) in both subgenus *Pinus* and *Strobus*, but negatively correlated in outgroup accessions (Fig. 2.3B, Table 2.2). Significant correlation (i.e., 95% confidence interval for slope does not include zero) was found between assembly success and sequencing effort in subgenus *Pinus* and when all pines were considered together (positive correlation), and between assembly success and divergence in subgenus *Strobus* (Table 2.2).

Noncoding regions (aligned positions excluding exons, introns and RNA loci) contained the highest proportions of variable sites when all accessions were considered together, or when subgenus *Pinus*, subgenus *Strobus* and non-pine assemblies were considered separately (Table 2.3). Exonic (protein coding) sequence, while approximately the same overall length as total noncoding sequence, contained ca. 60-70% as many of the variable sites by comparison; this proportion decreased further with the exclusion of *ycf*1 and *ycf*2. Both noncoding and exonic regions appear to have assembled with similar success (Table 2.3). In contrast, RNA loci had the lowest proportion of variable sites and also the lowest estimated assembly success (Table 2.3). Similarly, intron sequence was less variable than exonic sequence and noncoding sequence (but not exonic sequence without *ycf*1 and *ycf*2), and had a lower assembly success.

DISCUSSION

We have presented an effective and efficient method for assembling small, non-referenced genomes from short-read sequence data. Relying primarily on open source software, we assembled a total of 37 chloroplast genomes (36 non-referenced) of approximately 118kb length to an average of 92% completion. The process described herein is an iterative process. Initially, preliminary genomic contigs are created through *de novo* assembly. These contigs are then refined through alignment to a closest reference, formation of a chimeric reference, and subsequent re-alignment of short read sequences (reference-guided assembly) to the chimeric reference. Key to this process is the formation of the chimeric reference prior to reference-guided assembly, consisting of aligned *de novo* contigs with reference sequence utilized in place of missing data. In theory, this allows for more accurate final assemblies for several reasons, including: 1) clear identification of indels within aligned *de novo* contigs, 2) potential identification of structural rearrangements through *de novo* assemblies, and 3) improved reference-guided assembly due to higher sequence identity between the chimeric

reference and short read genomic sequences as opposed to simply relying on the closest reference without manipulation. It is worth noting that we utilized what we considered to be the most up to date and applicable software at the time of our assemblies. However, since that time a considerable amount of effort has been put into developing and refining short read assembly software, such that for each step in our assembly process there are now several options. For example, the open source aligner Mummer (http://mummer.sourceforge.net/) could be used to in place of the commercially available aligner CodonCode; alternative reference-guided alignment programs, such as Maq (http://maq.sourceforge.net/maq-man.shtml), Bowtie (http://bowtie-bio.sourceforge.net/index.shtml, Langmead et al. 2009) and Yasra (Miller and Ratan, unpublished), could be used in place of RGA. While the bulk of our assemblies had fairly closely related references (within the same subgenus), it is noteworthy that 13 of our *Pinus* assemblies reside in different sections than the complete reference used in their assembly. This represents an estimated divergence period of 19-47 million years in the case of subgenus *Strobus* accessions and 14-31 million years in the case of subgenus *Pinus* accessions (Willyard et al. 2007, Gernandt et al. 2008). Further, the divergence period between the pines and our non-pine representatives is estimated at 87 to 193 million years (Willyard et al. 2007, Gernandt et al. 2008), thereby representing a several-fold greater divergence period yet. It is not surprising, then, that assembly success was somewhat negatively impacted by increasing phylogenetic distance from the reference (although the opposite trend was seen in subgenus *Strobus*). Nonetheless, these trends were not particularly strong, and our worst assembly was estimated at 75% complete (*P. cembra*). This provides validation for the effectiveness of our strategy in assembling genomes fairly divergent from the nearest available reference. Further support is found in the similar success rates of assembling coding and non-coding regions, in that one would expect to see decreased assembly success in more poorly conserved noncoding regions if our assembly strategy was lacking.

Considering that divergence plays a limited role in assembly success, it is then reasonable to ask what the most difficult obstacles are in assembling non-referenced genomic sequences. As reported by Cronn et al. (Cronn et al. 2008), assembly gaps are consistently found directly adjacent to primer sites used in PCR amplifications with our sequencing strategy. In addition, sequence repeats (such as microsatellites) may also be difficult or impossible to bridge with

short read data (Cronn et al. 2008). In our assemblies, primer regions are likely a more significant problem, as they accounted for a substantial portion of missing sequence and precluded the assembly of any contig greater than the largest amplicon (just over 4 kb). In addition, a PCR-based method is more prone to failure in capturing genomic structural rearrangements, as amplification failures will occur when rearrangements span more than one amplicon. This could result in an amplicon being scored as missing or failed without any indication of a rearrangement. Regions with problematic amplification due to technical difficulties, primer divergence, or rearrangements can also eliminate substantial regions of the genomic assembly. For example, missing amplicons are part of the reason for the lower assembly success in rRNA regions, particularly in subgenus *Strobus* (likely due to technical problems with amplification and primer divergence, data not shown).

The limitations of PCR-based approaches noted above may be overcome through hybridization-based strategies (Gnirke et al. 2009, Herman et al. 2009), as these methods promise both more even and more thorough coverage of targeted regions. However, these methods have yet to be proven widely applicable. Alternatively, paired-end sequencing of whole genomic extractions may allow the simultaneous capture of large portions of chloroplast, nuclear, and mitochondrial genomes from a single organism (Meyers et al, this volume), particularly if nuclear genomes are relatively small (<1.5 Gbp).

Sequencing effort also clearly plays a role in the overall success and accuracy of assemblies, although there may be a point of diminishing returns and the phylogenetic distance to reference may still play a significant role. For example, with the exception of *P. ponderosa* (which was sequenced over several sequencing runs and prepared with variable methodology), our greatest sequencing effort was in *P. thunbergii* (4.54 million reads, >1.8x sequencing effort of any other assembly). Nonetheless, this assembly was essentially identical in its completion to those of *P. taeda* and *P. pinaster*, which had 56% and 38% of the sequencing effort, respectively. On the other hand, missing sequence in these accessions is mostly associated with primer locations, which are impossible to recover with our strategy. Notably, other studies (Hillier et al. 2008, Whittall et al. 2009) have also demonstrated improved SNP discrimination with increasing coverage depth. For these reasons, it may be a good strategy in future projects to overestimate necessary sequencing effort rather than trying to maximize

taxon density through low coverage levels, depending on the specific aims of the project. Alternatively, when assembling numerous non-referenced genomic sequences a reasonable strategy might be to initially dedicate a larger proportion of sequencing effort to the assembly of one or several representative reference genomes. This could in turn be followed by higher levels of multiplexing (relatively less sequencing effort) for subsequent accessions/taxa.

In the near future, it is reasonable to expect that increasing sequence capacity, as well as concurrent improvements in both targeted sequencing and short read assembly strategies will make *de novo* assembly of small genomes an increasingly simple process. Illumina predicts that their sequencing capacity will approach 100 Gbp per run by the end of 2009. Theoretically, this is sufficient capacity to sequence over 600 average-sized chloroplast genomes or 5500 average sized animal mitochondria to a depth of 100x in a single sequencing run. In order to efficiently utilize this capacity, however, it is clearly incumbent upon those involved in the sequencing of small genomes to maintain a similar pace of development in the areas of sample preparation and downstream assembly.

ACKNOWLEDGEMENTS

We thank Mariah Parker-deFeniks and Sarah Sundholm for lab assistance, Uranbileg Daalkhaijav, Zachary Foster and Brian Knaus for computing assistance, Linda Raubeson for providing a chloroplast isolation of *Larix occidentalis*, David Gernandt for Sanger sequencing, and Christopher Campbell and Justen Whittall for DNA samples. We also thank Mark Dasenko, Scott Givan, Chris Sullivan and Steve Drake of the OSU Center for Genome Research and Biocomputing. This work was supported by National Science Foundation grants (ATOL-0629508 and DEB-0317103 to A.L. and R.C.), the Oregon State University College of Science and the US Forest Service Pacific Northwest Research Station. New sequences have been deposited in GenBank with accessions numbers FJ899555-FJ899583.

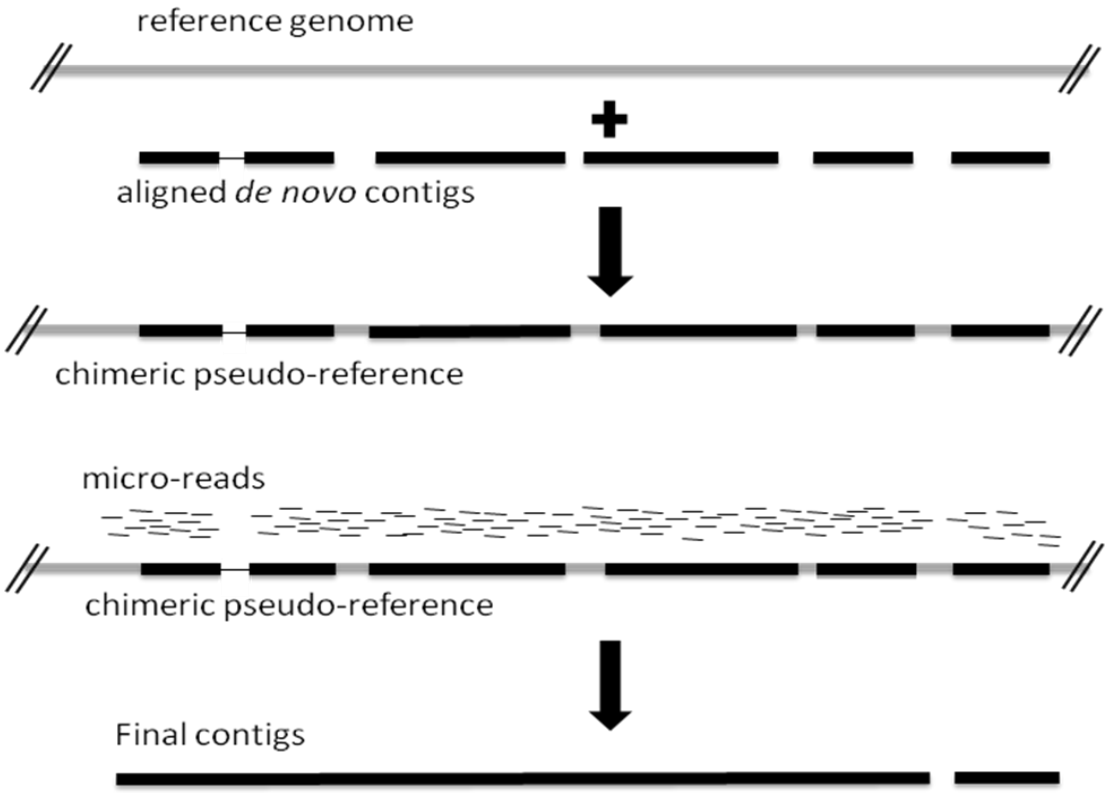


Figure 2.1. Schematic diagram of assembly process from short read data. A) Alignment of de novo contigs to reference genome and merging to form chimeric reference. B) Alignment of microreads to chimeric reference to form genomic contigs.

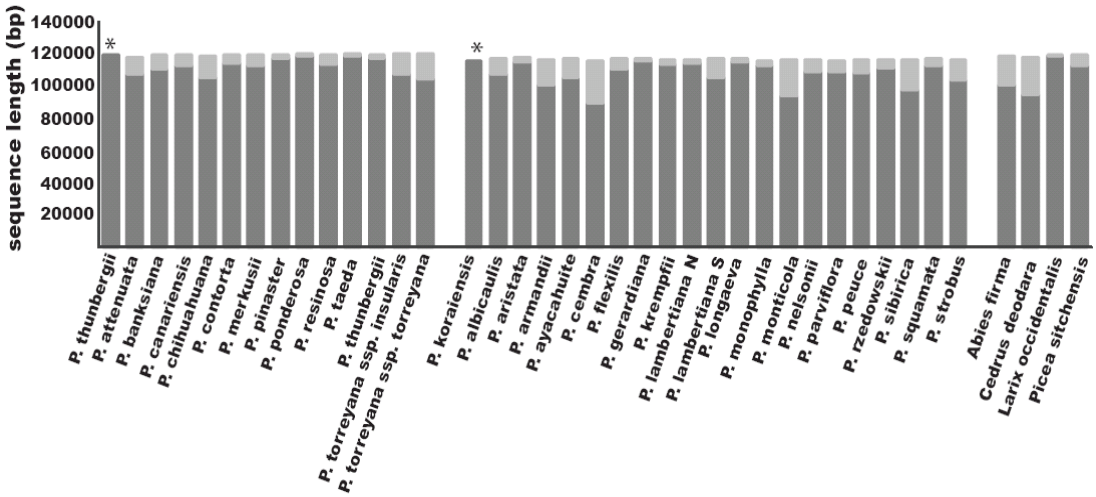


Figure 2.2. Assembly success for *Pinus* and outgroup assemblies. Dark shading indicates amount of chloroplast genomic sequence successfully assembled; light shading indicates estimated unassembled sequence. * indicate reference genomes.

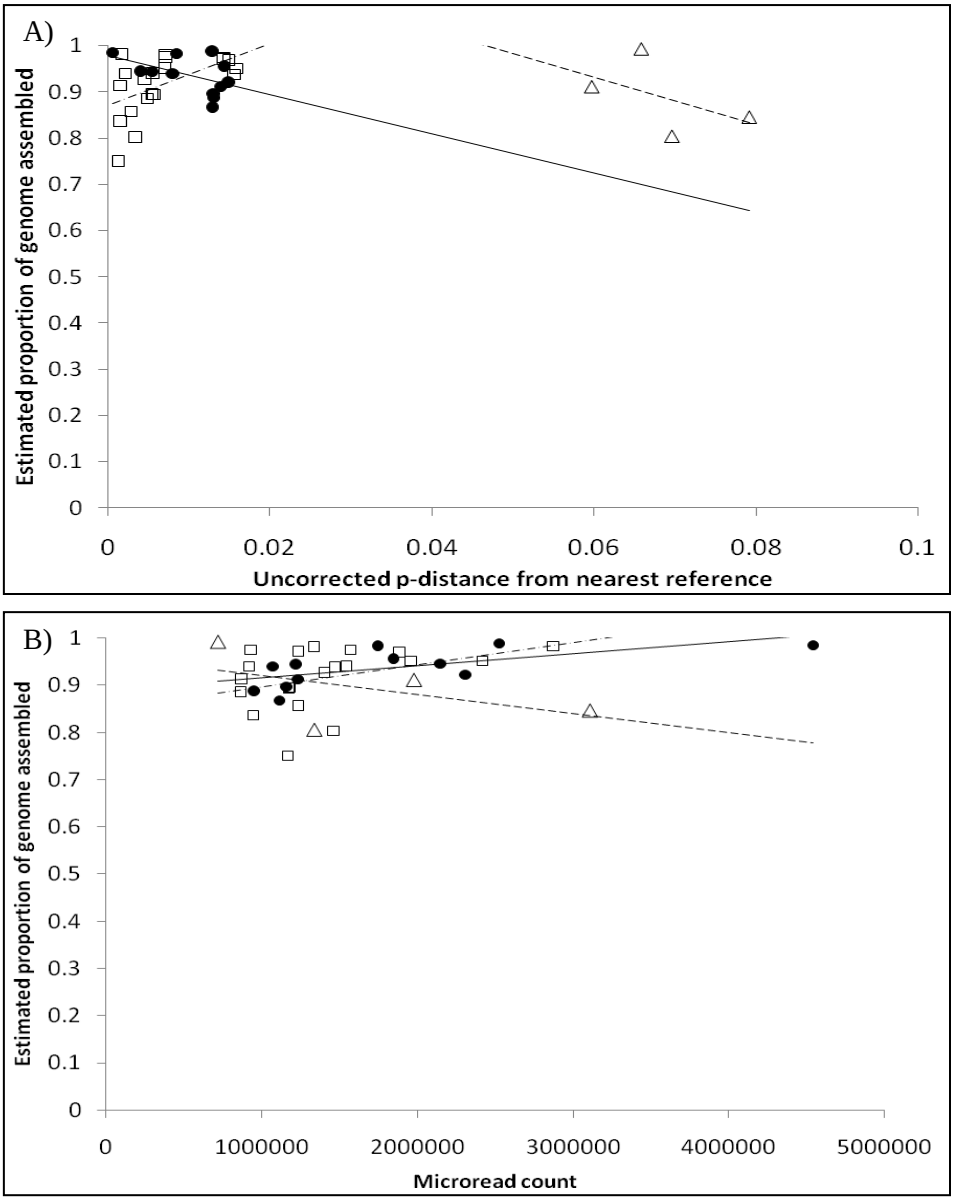


Figure 2.3. Potential factors contributing to assembly success in ingroup accession assemblies. A) Relationship of assembly success and divergence from nearest complete reference used in assemblies. B) Relationship between assembly success and sequencing effort (i.e., number of microreads). In each chart, relationships are shown for subgenus *Pinus* (solid circles and lines), subgenus *Strobus* (squares and dash-dot lines) and outgroup (triangles and dashed lines) accessions. Results for *P. ponderosa* were not included in these estimations as the sequencing effort for this accession was substantially higher (10-20x the number of reads from several sequencing runs) than that for other accessions. Regression lines for analyses of all pines not shown.

Table 2.1. Summary of genome assemblies and pairwise distances between assembled accessions and their original reference.

Reference	Accession	Estimated plastome length, bp[1]	Aligned de novo length, bp[2]	Determined genome length, bp[3]	Uncorrected p-distance from reference	# of differences to reference
Pinus thunbergii	*Abies firma*	119207	97641	100921	0.079221	7615
	Cedrus deodara	118072	92231	95083	0.069649	6337
	Larix occidentalis	119680	114509	118797	0.065859	7411
	Picea sitchensis	120176	106899	109548	0.059725	6243
	Pinus attenuata	118229	102684	107865	0.013953	1494
	P. banksiana	120166	106027	110792	0.014866	1624
	P. canariensis	120158	114492	112900	0.008007	895
	P. chihuahuana	119202	110034	114793	0.013127	1373
	P. contorta	120011	76276	105833	0.014363	1628
	P. merkusii	119665	107634	113005	0.005494	616
	P. resinosa	120179	114145	117914	0.004079	460
	P. pinaster	119904	109302	119246	0.008528	995
	P. ponderosa	120289	108301	113659	0.012249	1444
	P. taeda	120422	116074	119057	0.012867	1512
	P. thunbergii	119717	114086	117936	0.000645	76
	P. torreyana ssp. *torreyana*	120401	108722	104432	0.012955	1335
	P. torreyana ssp. *insularis*	120412	108402	107978	0.013025	1388
Pinus koraiensis	*P. albicaulis*	117266	106403	107163	0.001573	168
	P. aristata	118226	111542	114628	0.014976	1683
	P. armandii	117141	99769	100399	0.002927	293
	P. ayacahuite	117424	103665	104985	0.005723	596
	P. cembra	115825	81630	86922	0.001317	114
	P. flexilis	117346	110273	110415	0.005543	607
	P. gerardiana	117615	114769	115466	0.007133	814
	P. krempfii	116598	110554	113730	0.007139	803
	P. lambertiana S	117515	104050	105203	0.005472	571
	P. lambertiana N	116449	107317	114390	0.001759	200
	P. longaeva	117726	111303	114798	0.014308	1611
	P. monophylla	116104	107095	112804	0.014291	1582
	P. monticola	116841	94630	93800	0.003391	316
	P. nelsonii	116616	106160	109434	0.015587	1671
	P. parviflora	115986	106600	109043	0.002189	238
	P. peuce	116697	106938	108157	0.004556	489
	P. rzedowskii	116802	106225	111128	0.015918	1727
	P. sibirica	116593	96630	97547	0.001572	153
	P. squamata	117848	109936	112199	0.007035	780
	P. strobus	116854	100714	103545	0.004956	511

1 Determined based on full alignment.
2 Total length of aligned de novo contigs.
3 Total of all positions in length with determined base call.

Table 2.2. Statistical summaries of relationships shown in Figure 2.3.

Group	Regression	Slope of linear regression line	95% confidence interval for slope	correlation (R^2)
subgenus *Pinus*	assembly success / p-distance	-4.22	-9.38, 0.94	0.249
	assembly success / microread count	2.55 ($x10^{-8}$)	0.45 ($x10^{-8}$), 4.66 ($x10^{-8}$)	0.422
subgenus *Strobus*	assembly success / p-distance	6.70	1.56, 11.83	0.294
	assembly success / microread count	4.70 ($x10^{-8}$)	-0.88 ($x10^{-8}$), 10.27 ($x10^{-8}$)	0.148
all pines	assembly success / p-distance	3.42	-0.41, 7.25	0.100
	assembly success / microread count	3.34 ($x10^{-8}$)	0.85, 5.82	0.201
outgroups	assembly success / p-distance	-5.12	-31.33, 21.09	0.261
	assembly success / microread count	-4.00 ($x10^{-8}$)[1]	-0.88 ($x10^{-8}$), 10.27 ($x10^{-8}$)	0.249

1 evidence of non-normality in data set

Table 2.3. Summaries of alignment length, variable sites and assembly success for different partitions of aligned assemblies.

Type of region	Accessions included	Alignment length (bp)	# variable sites in alignment	Percentage of variable sites in alignment	Average percent sequence completion
non-coding	all (n=37)	58967	13209	22.4	93.5
	subgenus *Pinus* (n=13)			4.1	
	subgenus *Strobus*	51342	2081	5.1	96.3
	(n=20)	49497	2526		92.9
	non-pines only (n=4)	55134	7495	13.6	87.1
exons	all	57694	7924	13.7	93.4
	subgenus *Pinus*	55464	1726	3.1	93.9
	subgenus *Strobus*	56479	1797	3.2	93.5
	non-pines only	56508	3844	6.8	90.8
exons, no *ycf*1 or *ycf*2	all	48919	4500	9.2	94.2
	subgenus *Pinus*	48769	815	1.7	94.2
	subgenus *Strobus*	48726	974	2.0	94.7
	non-pines only	48776	2377	4.9	91.3
introns	all	13053	1780	13.6	87.7
	subgenus *Pinus*	12570	265	2.1	87.7
	subgenus *Strobus*	12627	387	3.1	87.2
	non-pines only	12784	1065	8.3	90.1
rRNA	all	4524	99	2.2	78.4
	subgenus *Pinus*	4518	18	0.4	79.0
	subgenus *Strobus*	4515	16	0.4	74.6
	non-pines only	4523	47	1.0	95.3
tRNA	all	1356	47	3.5	83.4
	subgenus *Pinus*	1356	5	0.4	90.6
	subgenus *Strobus*	1356	13	1.0	81.3
	non-pines only	1356	28	2.0	69.9

LITERATURE CITED

Ahn S-M, Kim T-H, Lee S, Kim D, Ghang H, Kim D-S, Kim B-C, Kim S-Y, Kim W-Y, Kim C, *et al.* 2009. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res., 19:1622-1629.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, *et al.* 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 456:53.

Collins FS, Lander ES, Rogers J, Waterston RH, Conso I. 2004. Finishing the euchromatic sequence of the human genome. Nature, 431:931-945.

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. Nature Meth., 5:887-893.

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res., 36:e122.

Gernandt DS, Magallon S, Geada Lopez G, Zeron Flores O, Willyard A, Liston A. 2008. Use of simultaneous analyses to guide fossil-based calibrations of Pinaceae phylogeny. Int. J. Plant Sci., 169:1086-1099.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotech., 27:182-189.

Gupta PK. 2009. Single-molecule DNA sequencing technologies for future genomics research. Trends in Biotechnology, 26:602-611.

Herman DS, Hovingh GK, Iartchouk O, Rehm HL, Kucherlapati R, Seidman JG, Seidman CE. 2009. Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. Nature Meth., 6:507-510.

Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, *et al.* 2008. Whole-genome sequencing and variant discovery in *C. elegans*. Nature Meth., 5:183-188.

Holt RA, Jones SJM. 2008. The new paradigm of flow cell sequencing. Genome Res., 18:839.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W. 2001. Initial sequencing and analysis of the human genome. Nature, 409:860-921.

Langmead B, Trapnell C, Pop M, Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol., 10:R25.

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature, 456:66-72.

Lidholm J, Gustafsson P. 1991. The chloroplast genome of the gymnosperm *Pinus contorta*: a physical map and a complete collection of overlapping clones. Curr. Genet., 20:161-166.

Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K, Koboldt DC, Fulton RS, Delehaunty KD, McGrath SD, *et al.* 2009. Recurring mutations found by sequencing an acute myeloid leukemia genome. New England Journal of Medicine, 361:1058-1066.

Pop M, Salzberg SL. 2008. Bioinformatics challenges of new sequencing technology. Trends Genet., 24:142-149.

Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, *et al.* 2007. Multiplex amplification of large sets of human exons. Nature Meth., 4:931-936.

Simon SA, Zhai J, Nandety RS, McCormick KP, Zeng J, Mejia D, Meyers BC. 2009. Short-read sequencing technologies for transcriptional analyses. Annual Review of Plant Biology, 60:305-333.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA. 2001. The sequence of the human genome, 291:1304-1351.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J. 2008. The diploid genome sequence of an Asian individual. Nature, 456:60-65.

Whiteford N, Haslam N, Weber G, Prugel-Bennett A, Essex JW, Roach PL, Bradley M, Neylon C. 2005. An analysis of the feasibility of short read sequencing. Nucleic Acids Res., 33:e171.

Whittall JB, Syring J, Parks M, Buenrostro J, Dick C, Liston A, Cronn R. 2009. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. Mol. Ecol.:in press.

Willyard A, Syring J, Gernandt DS, Liston A, Cronn R. 2007. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. Mol. Biol. Evol., 24:90-101.

Increasing Phylogenetic Resolution at Low taxonomic Levels Using Massively Parallel Sequencing of Chloroplast Genomes

Matthew Parks, Richard Cronn and Aaron Liston

BMC Biology
BioMed Central Ltd.
Floor 6
236 Gray's Inn Road
London WC1X 8HB
United Kingdom
BMC Biology 7:84.

ABSTRACT

Molecular evolutionary studies share the common goal of elucidating historical relationships, and the common challenge of adequately sampling taxa and characters. Particularly at low taxonomic levels, recent divergence, rapid radiations, and conservative genome evolution yield limited sequence variation, and dense taxon sampling is often desirable. Recent advances in massively parallel sequencing (MPS) make it possible to rapidly obtain large amounts of sequence data, and multiplexing makes extensive sampling of megabase sequences feasible. Is it possible to efficiently apply MPS to increase phylogenetic resolution at low taxonomic levels?

We reconstruct the infrageneric phylogeny of *Pinus* from 37 nearly-complete chloroplast genomes (avg. 109 kilobases each of an approximately 120 kilobase genome) generated using multiplexed MPS. 30/33 ingroup nodes resolved with ≥95% bootstrap support; this is a substantial improvement relative to prior studies, and shows MPS-based strategies can produce sufficient high quality sequence to reach support levels originally proposed for the phylogenetic bootstrap. Resampling simulations show that at least the entire plastome is necessary to fully resolve *Pinus*, particularly in rapidly radiating clades. Meta-analysis of 99 published infrageneric phylogenies shows that whole plastome analysis should provide similar gains across a range of plant genera. A disproportionate amount of phylogenetic information resides in two loci (*ycf*1, *ycf*2), highlighting their unusual evolutionary properties.

Plastome sequencing is now an efficient option for increasing phylogenetic resolution at lower taxonomic levels in plant phylogenetic and population genetic analyses. With continuing improvements in sequencing capacity, the strategies herein should revolutionize efforts requiring dense taxon and character sampling, such as phylogeographic analyses and species-level DNA barcoding.

INTRODUCTION

Molecular phylogenetic and phylogeographic analyses are typically limited by DNA sequencing costs, and this forces investigators to choose between dense taxon sampling with a small number of maximally informative loci, or genome-scale sampling across a sparse taxon sample (Delsuc et al. 2005, Philippe 2005, Jansen et al. 2007, Moore et al. 2007). Balancing these choices is particularly difficult in studies focused on recently diverged taxa or ancient rapid radiations, as taxon sampling needs to be sufficiently large to define the magnitude of intraspecific variation and the phylogenetic depth of shared alleles (Liston et al. 2007, Whitfield and Lockhart 2007). Similarly, broad genome sampling is necessary to offset the low level of genetic divergence among individuals of recent co-ancestry and to overcome low phylogenetic signal to noise ratios characteristic of rapid radiations (Whitfield and Lockhart 2007). Next generation DNA sequencing is poised to bring the benefits of affordable genome-scale data collection to such studies at low taxonomic levels (genera, species, and populations). Massively parallel sequencing (MPS) has increased per instrument sequence output several orders of magnitude relative to Sanger sequencing, with a proportional reduction in per-nucleotide sequencing costs (Hudson 2008, Mardis 2008). In principle this could allow the rapid sequencing of large numbers of entire organellar genomes (chloroplast or mitochondria) or nuclear loci, and result in greatly increased phylogenetic resolution (Cronn et al. 2008). To date, comparatively few plant or animal evolutionary genetic analyses have utilized MPS (Moore et al. 2006, Gilbert et al. 2008, Ossowski et al. 2008), due to associated costs and the technical challenge of assembling large contiguous sequences from micro-reads. These barriers have been largely eliminated through four innovations: development of strategies for targeted isolation of large genomic regions (Porreca et al. 2007, Cronn et al. 2008, Gnirke et al. 2009, Herman et al. 2009); harnessing the capacity of these platforms to sequence targeted regions in multiplex (Porreca et al. 2007, Craig et al. 2008, Cronn et al. 2008); streamlining sample preparation and improving throughput (Quail et al. 2008); and developing accurate *de novo* assemblers that reduce reliance upon a predefined reference sequence (Hernandez et al. 2008, Zerbino and Birney 2008).

In this paper we demonstrate the feasibility and effectiveness of MPS-based chloroplast phylogenomics for one-third of the world's pine species (*Pinus*), a lineage with numerous unresolved relationships based on previous cpDNA-based studies (Wang et al. 1999, Gernandt et al. 2005, Eckert and Hall 2006). We also highlight the broad applicability of our approach to other plant taxa, and remark on the potential applications to similar mitochondrial-based studies in animals and plant DNA barcoding. Using multiplex MPS approaches, we sequenced nearly-complete chloroplast genomes (120 kilobases (kb) each total length) from 32 species in *Pinus* and four relatives in Pinaceae. Our sampling of *Pinus* includes both subgenera (subg. *Pinus*, 14 accessions; subg. *Strobus*, 21 accessions) and species exemplars chosen from all 11 taxonomic subsections (Gernandt et al. 2005) to evenly cover the phylogenetic diversity of the genus. Taxon density is highest for a chosen subsection (subsect. *Strobus*) as representative of a species-rich clade lacking phylogenetic resolution in previous studies (Wang et al. 1999, Gernandt et al. 2005, Liston et al. 2007, Syring et al. 2007). Three species are also represented by two chloroplast genomes each (*P. lambertiana, P. thunbergii, P. torreyana*).

METHODS AND MATERIALS

DNA Extraction, Amplification and Sequencing

DNA extraction, amplification and sequencing are described in and followed Cronn et al. (Cronn et al. 2008), with 4 base pair (bp) tags, replacing the original 3 bp tags (Table 3.1). For one sample, *P. ponderosa*, additional reads from three non-multiplexed lanes of genomic DNA were also included.

Sequence Assembly and Genome Alignments

Sequence assembly and alignment are described in and followed Whittall et al. (Whittall et al. 2009). An analysis of interspecific recombination was conducted using RDP v. 3.27 (Martin et al. 2005b). Rather than using the full genomic alignment, which was too memory-intensive, concatenated nucleotide sequences for 71 exons common to all accessions were used (reflective of order on the plastome). Subgenera were investigated separately as members of opposing subgenera appear incapable of hybridization (Price et al. 1998). Each subgenus was checked for recombination events using standard settings for several recombination-detection strategies, including: Recombination Detection Program (RDP) (Martin and Rybicki 2000), GeneConv (Padidam et al. 1999), Chimaera (Posada and Crandall 1998), MaxChi (Smith 1992),

BootScan (Martin et al. 2005a), and SiScan (Gibbs et al. 2000). A total of 24 putative recombination events were identified. On close investigation, all events involved one or more of the following: misalignment, autapomorphic noise coupled with missing data, and amplification of pseudogenes. In cases of misalignment, alignments were corrected prior to subsequent phylogenetic analyses. In cases of amplification of pseudogenes, the entire amplicon for the accession involved was turned to N's. Inspection of the alignment also revealed that some amplicons in some accessions had failed to amplify, or amplified apparently paralogous loci (evidenced by substantially higher divergence). These regions were masked in affected accessions. The locus *mat*K was determined to be a putative paralog in several accessions, and in four (*P. armandii*, *P. lambertiana* S, *P. albicaulis*, and *P. ayacahuite*) it was replaced with Sanger sequence (Liston et al. 2007). We also replaced 2180 bp of poor quality sequence of the locus *ycf*1 in *P. ponderosa* with Sanger sequence. In all accessions amplified by PCR, the regions adjacent to primer sites typically had low coverage, while primers had very high coverage, thus primer-flanking regions (where problematic) and the primers were also excluded. It was also determined through Sanger sequencing that a 600 bp region of the previously published *P. koraiensis* plastome (positions 48808-49634 in GenBank AY228468) is apparently erroneous. This region was removed and reference guided analysis was rerun for this amplicon.

Aligned sequences were annotated using Dual Organellar Genome Annotator (DOGMA) (Wyman et al. 2004) with manual adjustments to match gene predictions from GenBank and the Chloroplast Genome Database (http://chloroplast.cbio.psu.edu/). Exons were evaluated for reading frame and translations, and validity of exon mutations was judged based on presence in de novo sequence, effect on the resulting polypeptide sequence, and sequence coverage depth.

Data Deposition

Illumina sequencing reads and quality scores have been deposited in the NCBI SRA database as accession SRA009802. New sequences have been deposited in GenBank as accessions FJ899555-FJ899583.

Phylogenetic Analyses

Sequence data was analyzed using all genome positions and concatenated nucleotide sequence from 71 exons common to all pine accessions; both partitions were analyzed

with and without the loci *ycf*1 and *ycf*2. A relatively short (~630 bp) repetitive stretch of the locus *ycf*1 of subgenus *Strobus* accessions was masked in all analyses due to alignment ambiguity. The loci *ycf*1 and *ycf*2 (ca. 14 kb combined) were also analyzed individually and together.

Maximum Likelihood (ML) phylogenetic analyses were performed through the Cipres Web Portal (http://www.phylo.org/portal/Home.do) using RAxML bootstrapping with the general model of nucleotide evolution (GTR+G) (Stamatakis 2008) and automatically determined numbers of bootstrap replicates. Bayesian inference (BI) analyses were performed using MrBayes v. 3.1.2 (Ronquist and Huelsenbeck 2003) using the GTR+G+I model, which was selected using MrModelTest v. 2.3 (Nylander 2004) under both Aikake Information Criterion and Hierarchical Likelihood Ratio Test frameworks. Each analysis consisted of two runs with four chains each (three hot and one cold chain), run for 1000000 generations with trees sampled every 100 generations. The first 25% percent of trees from all runs were discarded as burn-in. Unweighted maximum parsimony (MP) analyses of data partitions were conducted in PAUP* v. 4.0b10 (Swofford 2000) by heuristic search with 10 replicates of random sequence addition, tree bisection and reconnection branch swapping and a maxtrees limit of 1000. Non-parametric bootstrap analysis was conducted under the same conditions for 1000 replicates to determine branch support.

Topological differences between the full alignment topology and each of the three other largest data partitions (full alignment without *ycf*1 and *ycf*2, and exon nucleotides both with and without *ycf*1 and *ycf*2) were tested for significance using the Shimodaira-Hasegawa test (Shimodaira and Hasegawa 1999) with resampling estimated log-likelihood (RELL) bootstrapping (1000 replicates) under the GTR+G model of evolution. To further determine which topological differences were most influential, tests were repeated with the positions of topology-variable accessions alternately modified to match the full alignment topology. In total, the full alignment data set was compared to nine different topologies.

Exon indels and stop codon shifts were mapped onto the topology determined by ML analysis of the full alignment by parsimony mapping using Mesquite v. 2.6 (Maddison and Maddison, http://mesquiteproject.org). Tests of selection for exons were performed in MEGA v. 4.0 (Tamura et al. 2007) using the codon-based Z-test for

selection, with pairwise deletion and the Nei-Gojobori (p-distance) model; variance of the differences were computed using the bootstrap method with 500 replicates.

Estimation of Divergence Times for Poorly Resolved Nodes

Divergence times for four nodes with topological uncertainty (*P. albicaulis* - *P. lambertiana* N - *P. parviflora*, *P. sibirica* - *P. cembra* - *P. koraiensis*, *P. krempfii* – section *Quinquefoliae* of subgenus *Strobus*) were estimated according to Pollard et al. (Pollard et al. 2006). Chloroplast mutation rate was estimated by averaging maximum and minimum mutation rates for Pinaceae chloroplast genomes from two previous studies (Willyard et al. 2007, Gernandt et al. 2008) and assuming a generation time of 50 years (Bouille and Bousquet 2005). Two estimates were calculated for each node using either low (10000) or high (100000) effective population size (Syring et al. 2007).

Effect of Character Number on Phylogenetic Resolution

(1) Empirical data from *Pinus* genomes

Variable-size random subsamples of the full alignment were tested under the parsimony criteria using PAUP* v. 4.0b10 (the faststep option was used for all but the two smallest partitions due to time considerations). Eleven partition sizes were tested (2.5, 5, 10, 20, 30, 40, 50, 60, 80, 100 and 120 kb) in five replicates each, with resolution measured as the percentage of ingroup nodes produced with ≥95% jackknife support. Relationships between partition size and ingroup resolution were estimated using least squares regressions, and 95% confidence limits for individual points were estimated based on linear regression using SAS JMP 7.0.1 (S.A.S. Institute, Inc., http://www.jmp.com/). Our full alignment, exon nucleotides and *ycf*1/*ycf*2 partitions were analyzed under the same parsimony criteria for comparison, as were the alignments of (Wang et al. 1999, Gernandt et al. 2005, Eckert and Hall 2006). Accessions from Gernandt et al. and Eckert et al. (Gernandt et al. 2005, Eckert and Hall 2006) were pruned to include only taxa common to our sampling; the original analysis of Wang et al. (Wang et al. 1999) was used since this data matrix was not available for alternative phylogenetic analyses.

(2) Meta-Analysis of Published Studies

We evaluated 99 phylogenetic analyses from 86 studies published between 2006-2008 in Systematic Botany, Systematic Biology, American Journal of Botany, Taxon,

Molecular Phylogenetics and Evolution, and Annals of the Missouri Botanical Garden (see additional data file 2). Analyses were selected based on: 1) the presented phylogeny was based solely on chloroplast DNA sequence; 2) the analysis included ≥10 species from a monophyletic genus; 3) there were more inter- than intra-specific taxa analyzed within the genus; 4) parsimony-based bootstrap or jackknife values were presented. Ingroup branches with bootstrap support ≥95%, the number of ingroup taxa and the aligned base pairs used in the analysis were recorded for each case. The authors' taxonomic interpretations were accepted in instances of taxonomic uncertainty. Conspecific clades were treated as one taxon unless clearly differentiated from one another, and internal bootstrap values were disregarded. The number of branches with bootstrap support ≥95% was regressed both on the number of aligned base pairs and the number of taxa (both log-transformed to meet assumptions of normality and equal variances).

RESULTS

Genomic Assemblies and Alignment

Assemblies in subgenus *Strobus* averaged 117 kbp, with an estimated 8.8% missing data (compared to *P. koraiensis* reference); subg. *Pinus* assemblies averaged just less than 120 kbp (6% estimated missing data, compared to *P. thunbergii* reference). Outgroup assemblies averaged just over 119 kbp (10.4% average estimated missing data compared to *P. thunbergii* reference). Median coverage depth for determined positions was variable but typically high (range 21-156×) (Table 3.1, also see Appendix Figure 3.1). Full alignment of all assemblies was 132,715 bp in length, including 62,298 bp from exons encoding 71 conserved protein coding genes (20,638 amino acids), 36 tRNAs and 4 rRNAs. A high degree of co-linearity is inferred for these genomes due to the absence of major rearrangements within *de novo* contigs, and by the overall success of the polymerase chain reaction (PCR)-based sequence isolation strategy (indicating conservation of the order of anchor genes containing primer sites). However, minor structural changes (a tandem duplication in two species (Lidholm and Gustafsson 1991) and the apparent loss of duplicate copies of *psa*M and *rps*4 in *P. koraiensis*) could not be confirmed. No evidence of interspecific recombination was detected, consistent with the rarity of recombination in plant plastomes (Palmer 1985).

The aligned matrix contained 7,761 parsimony informative ingroup substitutions (4,286 non-coding positions and 3,475 coding positions) (Table 3.2). Over one-half of parsimony informative sites (55.0%) in protein coding regions resided in *ycf*1 and *ycf*2, two large genes of uncertain function (Drescher et al. 2000), that accounted for 22% of all exon sequence (Fig. 3.1A, 3.1B). No other exons in the pine plastome exhibit such a disproportionate number of parsimony informative sites (Fig. 3.1C). These loci have an elevated nonsynonymous substitution rate (Table 3.3) and appear to have a substantial number of indels in *Pinus*, although it was not possible in many cases to confidently score indels in these loci due to the inherent limitations of reference-guided assembly of short reads in length variable regions. Start codon position, overall length and stop codon positions were nonetheless largely preserved in these loci across the genus. In addition to substitutions in exons, 48 ingroup exon indels and 23 ingroup stop codon shifts were identified in 26 loci.

Phylogenetic Resolution in Non-Random and Randomized Data Partitions

Full alignment partitions yielded a higher proportion of highly supported nodes, with 88-91% (29-30/33) of ingroup nodes resolved with bootstrap support ≥95% in likelihood analysis. The four largest data partitions tested (full alignment and concatenated exon nucleotides, both with and without *ycf*1 and *ycf*2) yielded results that were topologically identical with the exception of four taxa (*P. albicaulis*, *P. krempfii*, *P. lambertiana* N, *P. parviflora*) (Figs. 3.2 and 3.3). In addition, support for the branching order of *P. cembra*, *P. koraiensis* and *P. sibirica* was low in full alignment partitions. Topological differences were found to be significant according to Shimodaira-Hasegawa comparisons of the full alignment topology to two of the other major partitions (full alignment and exon nucleotides without *ycf*1 and *ycf*2). Trends in significance were most strongly influenced by the two alternative positions of *P. krempfii* (Fig. 3.2 vs. Fig. 3.3A, C; Table 3.4). With the exception of *P. krempfii*, areas of topological uncertainty reside in a single clade that historically has lacked internal resolution (subsection *Strobus*) (Wang et al. 1999, Gernandt et al. 2005, Eckert and Hall 2006). Coalescent estimations suggest that these poorly resolved subsection *Strobus* haplotypes diverged in rapid succession relative to the age of their shared nodes (0.009 to 0.44 coalescent units, or ca. 90,000 – 450,000 years) (Table 3.5). A putative chloroplast capture event in *P. lambertiana* previously documented (Liston et al. 2007) was also supported with whole-plastome results. Substantial resolution was achieved in analyses of *ycf*1 and *ycf*2 data partitions, however we observed several

topological differences from the full alignment with high support (primarily involving the species discussed above) (Fig. 3.4).

Of the 71 exon coding indels and stop codon shifts identified, 35 mapped unambiguously to monophyletic groups (i.e., no accessions in a group were missing data for that event) (Figs. 3.5 and 3.6). All of these groups had strong support in nucleotide-based phylogenetic analyses (100% likelihood and parsimony bootstrap support). The remainder of these events were primarily either putatively monophyletic (missing data in one or more members of a clade) or showed strong evidence of homoplasy (Figs. 3.5 and 3.6).

In parsimony analyses of variable-sized jackknife samples of our full alignment, nodal support showed a strong positive correlation with the length of the nucleotide matrix (proportion nodes $\geq$95% = -1.0808 + 0.38497×$\log_{10}$[matrix size, bp]; r^2=0.915, P<0.0001) (Fig. 3.7A). Resolution of full alignment and exon nucleotide partitions was indistinguishable from random jackknife samples of comparable size, indicating similar phylogenetic content of these partitions and corresponding similar-sized random genomic subsamples. Partitions consisting of *ycf*1 and *ycf*2 – in particular *ycf1*, and *ycf1* and *ycf2* combined – showed significantly higher resolution than the genome-wide average (Fig. 3.7A). The concatenated partition *ycf1 + ycf2* (13.1 kb; 77.4% nodes ≥95% bootstrap support) yielded only slightly less phylogenetic resolution than all exons combined (62.3 kb; 80.6% nodes ≥95% bootstrap support) in parsimony analysis.

Comparisons to Previous *Pinus* Phylogenies

Previous chloroplast DNA-based estimates of infrageneric relationships in *Pinus* (Wang et al. 1999, Gernandt et al. 2005, Eckert and Hall 2006) sampled the same species and/or lineages as our study, and inferred relationships using 2.82 to 3.57 kb of chloroplast DNA. Results of these studies are largely consistent with our results, although highly supported nodes (≥95%) accounted for only 13 to 23% of the total ingroup nodes (23% to 42% if (Gernandt et al. 2005, Eckert and Hall 2006) adjusted to match our species composition). The empirical results of these studies fell within or close to the 95% prediction intervals established from our jackknife resampling response from our full genome alignment (Fig. 3.7A), indicating that the loci used in

prior studies (primarily *rbc*L and *mat*K) are similarly informative as a comparable sample of random nucleotides from the chloroplast genome.

Meta-Analysis of Published Infrageneric Studies

From our sampling, infrageneric analyses in plants published from 2006-2008 were typically based on 2574 aligned bp (95% bootstrap confidence interval: 2292, 2864) of sequence data, evaluated 31.7 ingroup species (95% bootstrap confidence interval: 20.2, 43.2), and resolved 22.6% of nodes at ≥95% bootstrap support (95% bootstrap confidence interval: 18.6, 26.5). Regression analysis shows that the proportion of highly resolved nodes in these studies is significantly and positively correlated with matrix length ($F_{1,96} = 18.032$; r^2=0.149; $P < 0.0001$) but not the number of included taxa ($F_{1,97} = 0.546$; $r^2 = 0.006$; $P = 0.461$), although there was a negative trend in the latter (Fig. 3.7B, 3.7C). Our current sample size is typical in the number of taxa sampled, but both matrix length (132.7 kbp) and the proportion of highly bootstrap-supported nodes (84.8% parsimony, 90.3% maximum likelihood (ML)) were substantially higher.

DISCUSSION

Our results highlight that whole plastome sequencing is now a feasible and effective option for inferring phylogenies at low taxonomic levels. Compared to previous chloroplast-based phylogenetic analyses in *Pinus*, our data matrix contained approximately 60 times more phylogenetically informative characters resulting in an approximately two- to four-fold increase in the proportion of highly resolved nodes (after adjusting results of previous studies to match our species composition) (Fig. 3.8, Table 3.2). An important question arising from these comparisons is whether the difference in resolution is entirely attributable to the increase in nucleotides, or whether the genomic partitions sequenced in prior studies were less informative on average than the rest of the genome. In fact, the resolution provided by loci used in previous *Pinus* studies is indistinguishable from or slightly greater than that of comparably sized random genomic subsamples from our full alignment. Combined with the strong correlation between resolution and the size of random genomic subsample, this suggests that the increase in resolution in this study is primarily due to the increase in matrix length. This is further supported by a significant relationship between resolution and matrix length in a broad sampling of chloroplast-based

infrageneric phylogenies. Based on these results, we predict that whole-plastome analysis will yield similar gains in phylogenetic resolution not only in the genus *Pinus* but for most land plant genera. On the other hand, it is apparent that even the entire chloroplast genome may be insufficient to fully resolve the most rapidly radiating lineages. In this regard, our results are reflective of previous analyses of ancient rapid radiations wherein nodal resolution does not scale proportionately to the length of sequence analyzed (Fishbein et al. 2001, Wortley et al. 2005). Notably, the position of *P. krempfii* was significantly different between the four largest data partitions (Table 3.4), even though this species does not appear to be associated with a rapid radiation (Table 3.5). This result is not completely unexpected, as this species has previously been difficult to place phylogenetically (Wang et al. 2000, Syring et al. 2005). An unequivocal resolution of this species will likely require the inclusion of multiple nuclear loci (Syring et al. 2005).

When considering recent divergence, the disproportionately high mutation rate in *ycf*1 (and *ycf*2, to a lesser extent) demonstrated here is of importance, and mirrors findings in other plant taxa (Chung et al. 2007, Neubig et al. 2009) and recently in *Pinus* subsection *Ponderosae* (Gernandt et al. 2009). These loci should be informative for phylogenetic studies in recently-diverged clades or in population-level studies in a range of plant species. Discretion is advised, however, as *ycf*1 (and possibly *ycf*2) appears to be a target of positive selection at least in *Pinus* and may reflect adaptive episodes rather than neutral genealogies. In likelihood analyses of *ycf*1 and *ycf*2, we observed several topological differences from the full alignment at the subsectional level, further demonstrating that caution must be taken in drawing phylogenetic conclusions from these two loci. Although we were able to confidently score small structural changes (indels and stop codon shifts) for all other exons, it was not possible to score indels for *ycf*1 and *ycf*2 due to the apparent high rate of indel formation in these loci. In all other loci examined, small structural changes only delineated clades with concurrent high support from nucleotide-based analyses (both in present study and (Wang et al. 1999, Gernandt et al. 2005, Eckert and Hall 2006)), and thus are likely to be of limited use in species or population level discrimination. It is not clear whether this will also be the case in *ycf*1 and *ycf*2.

It is reasonable to ask whether increased resolution is worth the effort of assembling whole plastomes. Considering the conservative nature of bootstrap measures (Hillis

and Bull 1993, Suzuki et al. 2002, Alfaro et al. 2003, Douady et al. 2003), systematists often accept bootstrap values of ≥70% as reliable indicators of accurate topology (Hillis and Bull 1993). Simulation studies (Alfaro et al. 2003), however, have demonstrated greatly increased accuracy (~42×) with bootstrap values ≥95% versus ≥70%, and the initial formulation of the phylogenetic bootstrap used ≥95% as the threshold for topological significance (Felsenstein 1985). Our results similarly support using a 95% bootstrap support cutoff for conclusive evidence as in both areas of topological differences, more than one clade received bootstrap support ≥70% by analysis of alternate data partitions. It is probable that conflicting topologies with ≥70% but <95% bootstrap support accurately reflect data partitions yet may not represent the plastome phylogeny, and here the use of entire organelle genomes makes it possible to adopt more conservative criteria of nodal support. There are further biological reasons why an organellar phylogeny (essentially a single-gene estimate) may not accurately represent the organismal phylogeny; these include interspecific hybridization, incomplete lineage sorting, and stochastic properties of the coalescent process. Nonetheless, phylogenetic reconstruction based on complete organellar sequences may facilitate the detection of such phenomena, by reducing errors and uncertainty due to insufficient sampling of DNA sequence.

In conclusion, plastome sequencing is now a reasonable option for increasing resolution in phylogenetic studies at low taxonomic levels and will continue to become an increasingly simple process. As sequencers evolve to even higher capacity and multiplexing becomes routine in the near future, this will allow more extensive taxon and genomic sampling in phylogenetic studies at all taxonomic levels. It is estimated that sequencing capacity on next generation platforms will approach 100 gigabase pairs per sequencing run by the end of 2009. For perspective, this is sufficient sequence capacity to produce all 100 genus-level data sets used in our meta-analysis (including ours) at greater than 100× coverage depth in a single sequencing run. Based on the estimates of Cronn et al. (Cronn et al. 2008), this sequencing capacity would also allow the simultaneous sequencing of several thousands of animal mitochondria, which could greatly benefit low-level taxonomic or population-based studies in animals that currently tend to rely on relatively short sequences from many individuals (Patenaude et al. 2007). It is also clear that these improvements could enable other pursuits that are currently hindered by limited sequencing capacity, such as identification of plants by diagnostic DNA sequences (DNA barcoding). The

recently agreed upon two locus chloroplast barcode for plants claims only 72% "unique identification to species level" (Hollingsworth et al. 2009). Based on results herein, whole plastome sequences have the potential to be more highly discriminating and efficient plant DNA barcodes; in fact, the possibility of plastome- and mitome-scale barcodes has been raised previously (Erickson et al. 2008). Results in this area (as well as in phylogenetic and phylogeographic analyses) will be impacted particularly if advances in target isolation and enrichment (Porreca et al. 2007, Gnirke et al. 2009, Herman et al. 2009) and streamlining sample preparation (Quail et al. 2008) prove globally effective.

Acknowledgements

We thank Mariah Parker-deFeniks and Sarah Sundholm for lab assistance, Uranbileg Daalkhaijav, Zachary Foster and Brian Knaus for computing assistance, Linda Raubeson for providing a chloroplast isolation of *Larix occidentalis*, Christopher Campbell and Justen Whittall for DNA samples, David Gernandt, Chris Pires, Jonathan Wendel and Mark Fishbein for editorial comments, and Steffi Ickert-Bond for timely questions. We also thank Mark Dasenko, Scott Givan, Chris Sullivan and Steve Drake of the OSU Center for Genome Research and Biocomputing. This work was supported by National Science Foundation grants (ATOL-0629508 and DEB-0317103 to A.L. and R.C.), the Oregon State University College of Science Venture Fund and the US Forest Service Pacific Northwest Research Station.

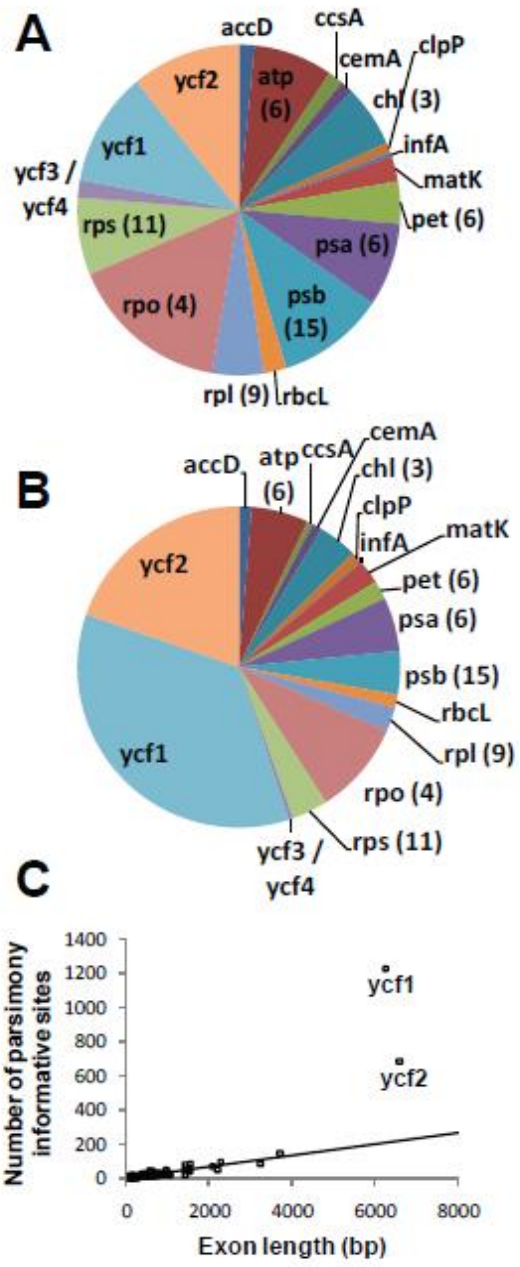


Figure 3.1. Length and information content of 71 exons common to *Pinus* accessions sampled in this study. A) Exon contributions to length as proportion of total exome length. B) Exon contributions to parsimony informative sites as proportion of total exome parsimony informative sites. C) Distribution of exons in relation to length and parsimony informative sites. In A) and B) most exons are shown by functional group (i.e., atp(), psb(); number of corresponding loci indicated in parentheses) for visualization purposes. In C) all exons were treated individually (N=71). Trendline in C) based on all exons with exception of *ycf*1 and *ycf*2 to emphasize their departure from trend in other exons.

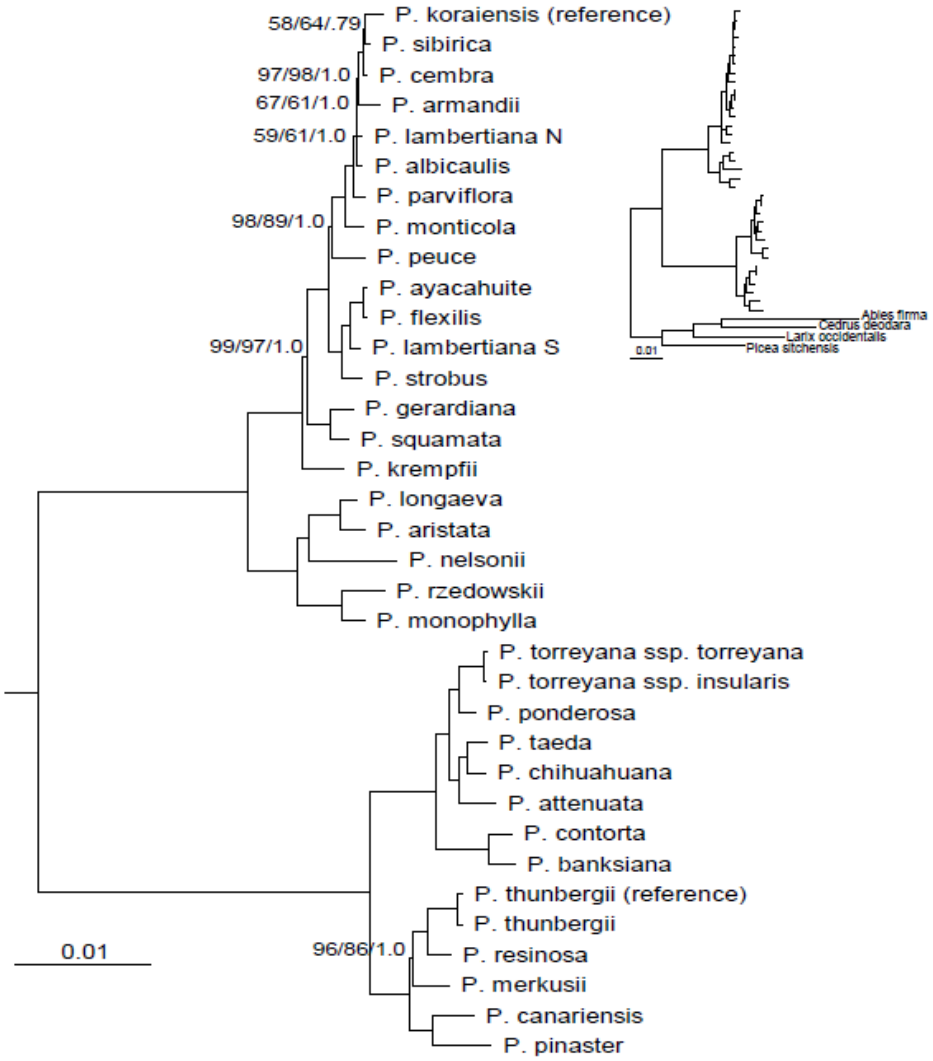


Figure 3.2. Phylogenetic relationships of 35 pines and four outgroups as determined from full plastome sequences. Support values are only shown for nodes with bootstrap / posterior probability values less than 100% / 1.0, and are shown as ML bootstrap / MP bootstrap / BI posterior probability. Branch lengths calculated through RAxML analysis, and correspond to scale bar (in units of changes / nucleotide position). Inset shows topology of outgroups relative to ingroup accessions.

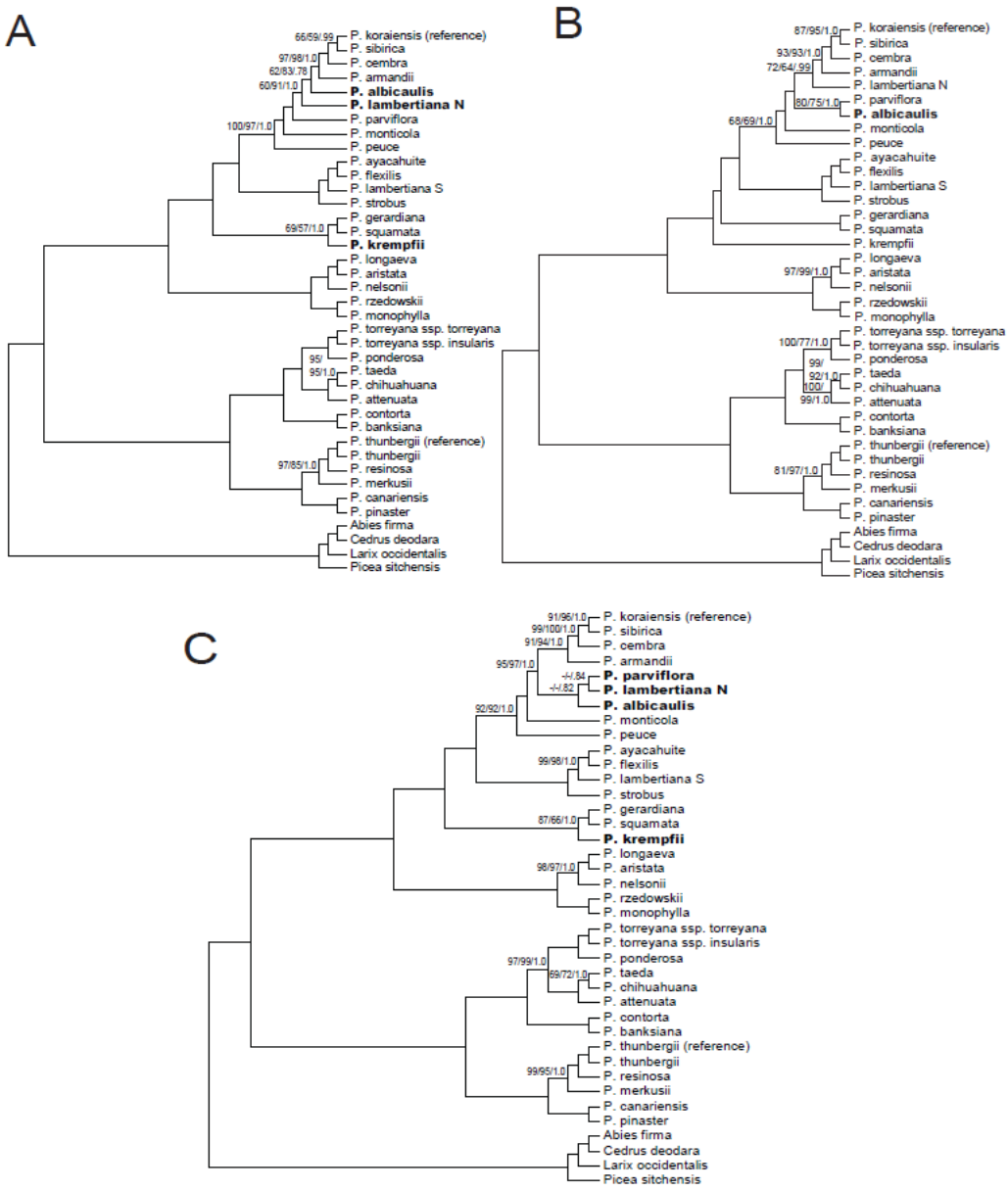


Figure 3.3. Phylogenetic relationships of 35 pines and four outgroups as determined from different data partitions. A) Full alignment without *ycf*1 and *ycf*2. B) Exon nucleotide sequences. C) Exon nucleotide sequences without *ycf*1 and *ycf*2. Support values are only shown for nodes with bootstrap / posterior probability values less than 100% / 1.0, and are shown as ML bootstrap / MP bootstrap / BI posterior probability. Branch lengths correspond to scale bar (in units of changes / nucleotide position, ML analysis). Dashes indicate <50% bootstrap support or <.50 posterior probability.

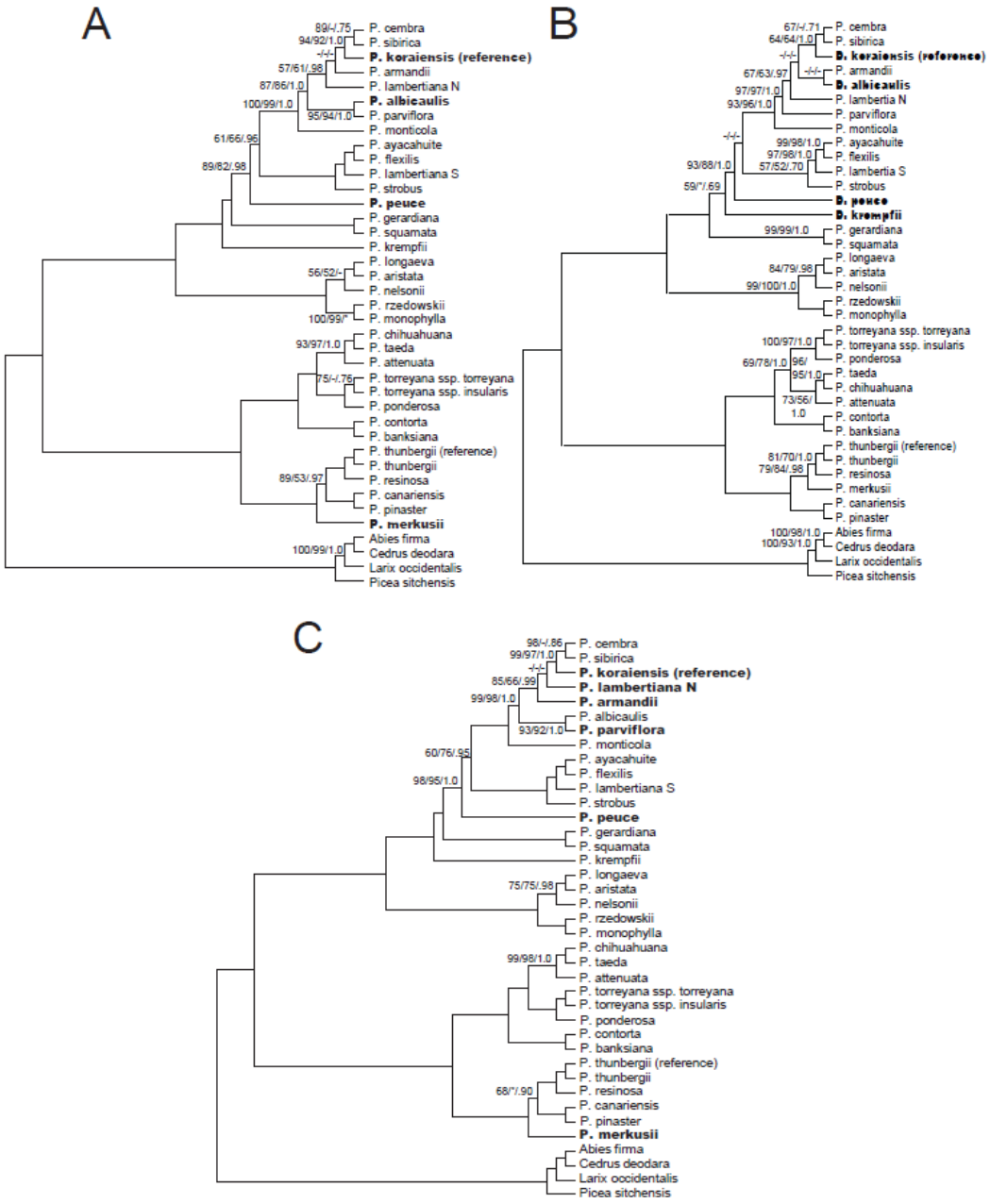


Figure 3.4. Phylogenetic relationships of 35 pines and four outgroups as determined from *ycf1* and *ycf*2 partitions. A) *ycf*1 only. B) *ycf*2 only. C) *ycf*1 and *ycf*2 combined. Support values are only shown for nodes with bootstrap / posterior probability values less than 100% / 1.0, and are shown as ML bootstrap / MP bootstrap / BI posterior probability. Branch lengths correspond to scale bar (in units of changes / nucleotide position, ML analysis). Dashes indicate <50% bootstrap support or <.50 posterior probability, * indicate topological difference between either parsimony or Bayesian analyses and ML. Accessions whose position differs from that in full alignment analysis indicated in **bold**.

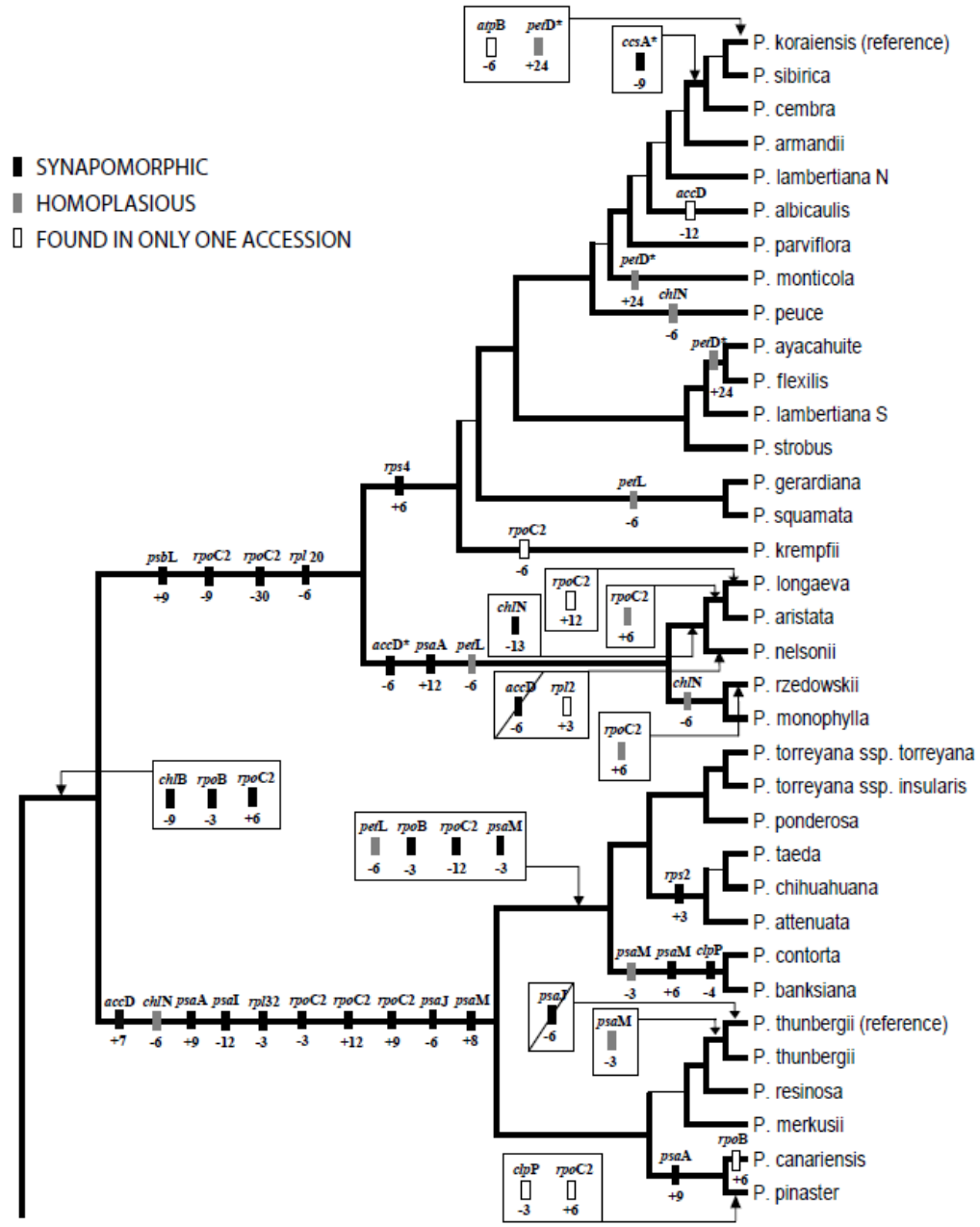


Figure 3.5. Phylogenetic distribution of exon coding indel mutations in sampled *Pinus* accessions. Exon names given above boxes, size of indel (bp) and polarity ("+" = insertion, "-" = deletion) given below boxes. Polarity of events determined by comparison to most distant outgroups. Due to the apparent high rate of indel formation in *ycf*1 and *ycf*2, these loci were not able to be confidently scored for indels and are not included in this diagram. Events for only the first copy of *psa*M are reported. Branching order of tree corresponds to RAxML analysis of complete alignment. Diagonal lines represent putative reversals of indel events. * indicates missing data for one or more accessions of clade. Thin internal branches correspond to ML bootstrap support <95% or topological difference in four largest data partitions (full alignment and exon nucleotides, with and without *ycf*1 and *ycf*2).

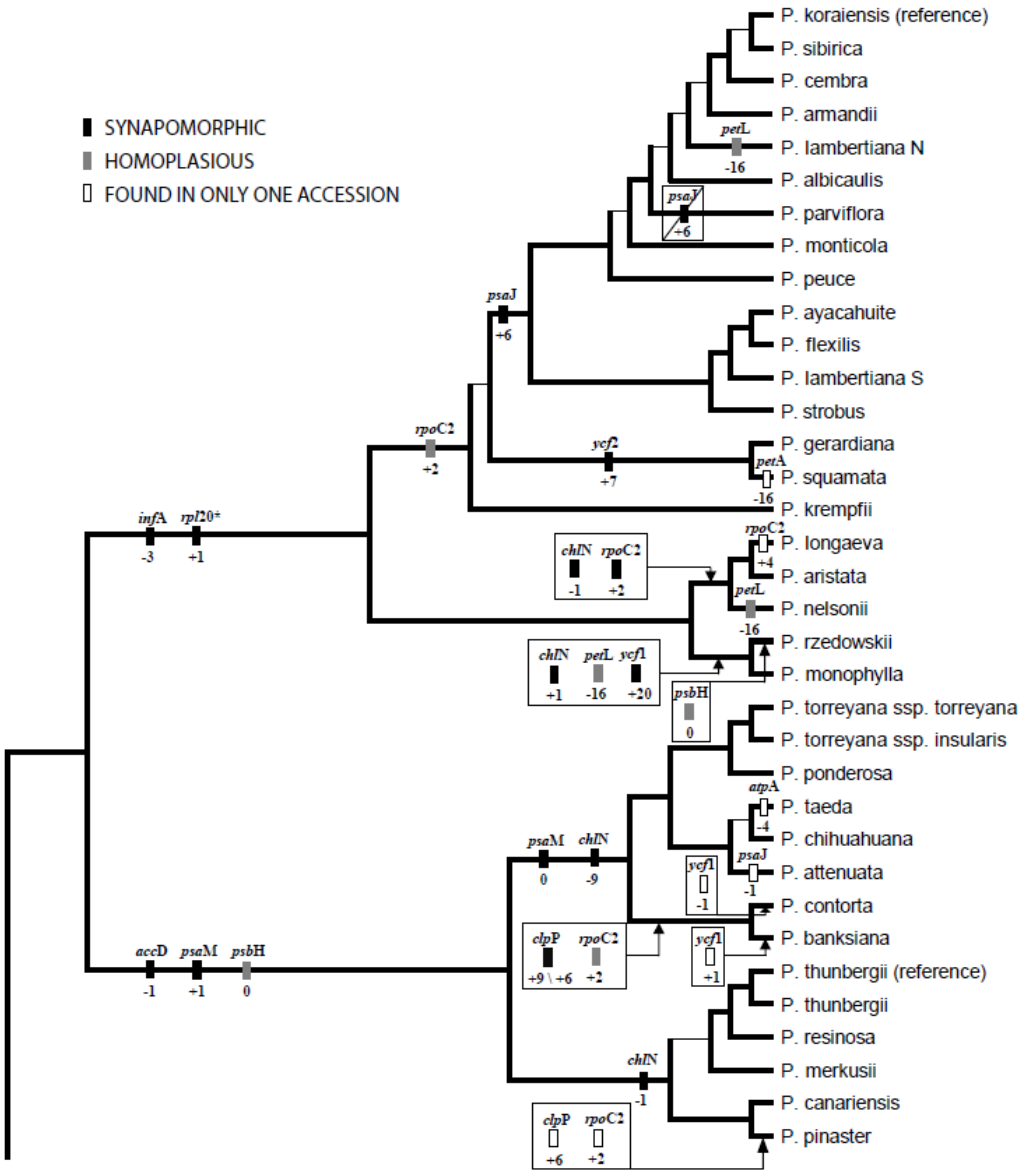


Figure 3.6. Phylogenetic distribution of stop codon mutations in sampled *Pinus* accessions. Exon names given above boxes, amino acid shift relative to stop codon position in outgroups given below boxes. Polarity of events determined by comparison to most distant outgroups; "+" signifies extension of coding region due to stop codon mutation, "-" signifies shortening. The value of zero for the *psb*H- and *psa*M-associated events corresponds to events that alter the original stop codon without altering the total number of codons in the locus. Events for only the first copy of *psa*M are reported. Diagonal line represents a putative reversal in *psa*J of *P. parviflora*. Branching order of tree corresponds to RAxML analysis of complete alignment. * indicates missing data for one or more accessions of clade. Thin internal branches correspond to ML bootstrap support <95% or topological difference in four largest data partitions (full alignment and exon nucleotides, with and without *ycf*1 and *ycf*2).

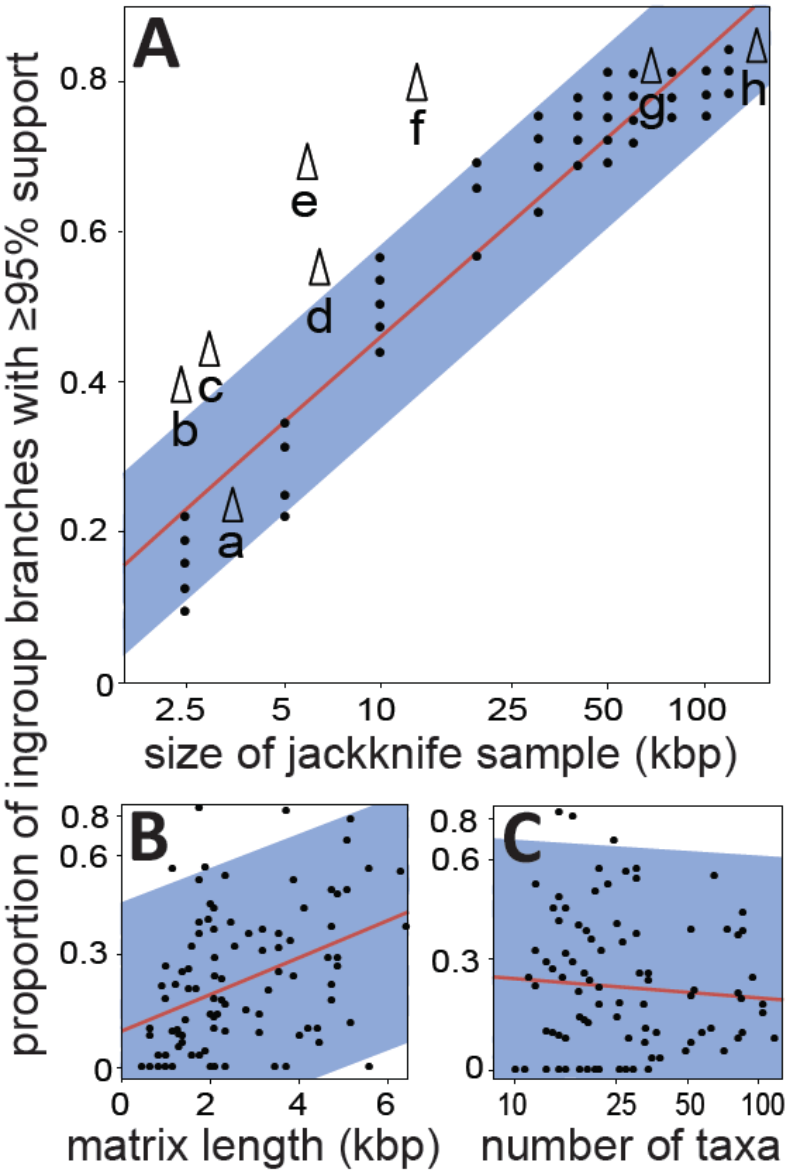


Figure 3.7. Relationships between matrix size and resolution in current study and meta-analysis of published studies. A) Parsimony resolution of jackknifed partitions (●) of full alignment of current study. Labelled data points (Δ) represent resolution of the following: a - Wang et al. (Wang et al. 1999), b - Gernandt et al. (Gernandt et al. 2005), c - Eckert and Hall (Eckert and Hall 2006), d - *ycf*2, e - *ycf*1, f - combined *ycf*1 and *ycf*2, g - exon nucleotides, h - complete alignment. B) Relationship between matrix length and phylogenetic resolution in published studies (N=99). C) Relationship between number of taxa and phylogenetic resolution in published studies (N=99). Regression lines are shown in red; 95% confidence intervals shown in blue. X-axes of A, B and C and Y-axes of B and C are in log scale.

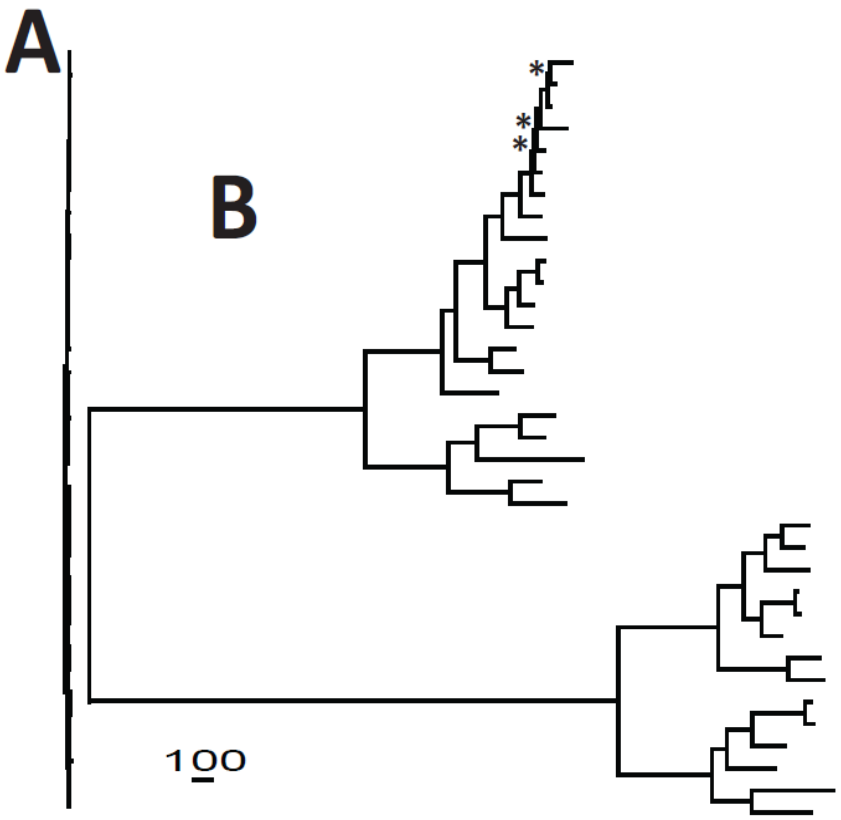


Figure 3.8. Comparative phylogenetic resolution of *Pinus* species used in this study. Resolution from A) two chloroplast loci (Gernandt et al. 2005) and B) our complete alignment. Distance bar corresponds to 100 nucleotide changes, and is scaled for either tree. * indicate branches with <95% (likelihood) bootstrap support in B) (likelihood and parsimony topologies were completely congruent).

Table 3.1. Multiplex tags and read count for sampled accession. "/" indicates accession was multiplex sequenced in two sequencing runs. Median coverage is reported for determined positions (≥2× coverage depth) in reference-guided analysis.

Accession	Multiplex Tag	Number of Reads	Read Length (bp, without tag)	Median coverage
Abies firma	AGCT	3110857	36	116
Cedrus deodara	CCCT	1338443	36	74
Larix occidentalis	GGT	719060	33	30
Picea sitchensis	ATT / AATT	1268688 / 710117	33 / 37	80
Pinus albicaulis	AGCT	869509	36	54
P. aristata	ACGT	1884108	36	100
P. armandii	AGCT	1233280	36	109
P. attenuata	ACGT	1230397	36	64
P. ayacahuite	CCCT	1173420	36	96
P. banksiana	AGCT	2307302	36	65
P. canariensis	CCCT	1069293	36	95
P. cembra	CTGT	1166707	36	40
P. contorta	CCT	1423631 / 423905	33 / 37	65
P. chihuahuana	CTGT	950336	36	21
P. flexilis	GGGT	1545509	36	136
P. gerardiana	GGT	1336725	33	98
P. krempfii	AAT	1569301	33	112
P. lambertiana N	ATT	1426598 / 1443555	33 / 37	99
P. lambertiana S	CCCT	1180289	36	113
P. longaeva	CCT	930078	33	89
P. merkusii	ATT	632411 / 585832	33 / 37	37
P. monophylla	GGT	1233556	33	145
P. monticola	CTGT	1460934	36	75
P. nelsonii	AAT	1139491 / 329838	33 / 37	81
P. parviflora	CCCT	920102	36	45
P. peuce	TACT	1402996	36	98
P. pinaster	GGT	1745043	33	77
P. ponderosa	CCT	16859450	33	44
P. resinosa	GGGT	2145134	36	48
P. rzedowskii	TACT	2419507	36	156
P. sibirica	CTGT	947216	36	60
P. squamata	TACT	1956311	36	97
P. strobus	GGGT	864197	36	42
P. taeda	CGT	1305703 / 1219158	33 / 37	90
P. thunbergii	AAT	1850050 / 2690553	33 / 37	104
P. torreyana ssp. *torreyana*	CTGT	1114111	36	76
P. torreyana ssp. *insularis*	ACGT	1157851	36	88

Table 3.2. Summary of variable and parsimony informative sites in data partitions. Data from Gernandt et al. (2005) and Eckert and Hall (2006) pruned to include only ingroup species and outgroup genera common to our study. (PI = parsimony informative.)

		Pines only		**Pines and outgroups**	
Treatment	**Aligned length**	**Variable positions (% of total)**	**PI positions (% of total)**	**Variable positions (% of total)**	**PI positions (% of total)**
All Nucleotides	132085	11179 (8.5)	7761 (5.9)	22834 (17.3)	11534 (8.7)
All Nucleotides without *ycf*1, *ycf*2	118935	8755 (7.4)	5852 (4.9)	18978 (16.0)	9038 (7.6)
Exon Nucleotides	62298	4716 (7.6)	3475 (5.6)	8346 (13.4)	4867 (7.8)
Exon Nucleotides without *ycf*1, *ycf*2	49044	2291 (4.7)	1566 (3.2)	4489 (9.2)	2381 (4.9)
*ycf*1	6355	1514 (23.8)	1227 (19.3)	2165 (34.1)	1507 (23.7)
*ycf*2	6794	910 (13.4)	682 (10.0)	1686 (24.8)	987 (14.5)
*ycf*1+*ycf*2	13149	2424 (18.4)	1909 (14.5)	3851 (29.3)	2494 (19.0)
Wang et al. (1999)	3513	196 (5.6)	127 (3.6)	482 (13.5)	243 (6.8)
Gernandt et al. (2005)	2817	197 (7.0)	128 (4.5)	345 (12.2)	167 (5.9)
Eckert and Hall (2006)	3288	217 (6.6)	123 (3.7)	411 (12.5)	206 (6.3)

Table 3.3. Codon-based Z-test for selection results for exon sequences. Results shown are overall average of all ingroup pairwise comparisons, with significance at $P \leq 0.05$ indicated in **bold**.

exon	P value H_A: dN > dS	P value H_A: dN < dS	test statistic	exon	P value H_A: dN > dS	P value H_A: dN < dS	test statistic
*acc*D	1	0.2013	0.8400	*psb*K	0.3925	1	0.2735
*atp*A	1	**0.0146**	2.2071	*psb*L	0.0922	1	1.3350
*atp*B	1	**0.0007**	3.2809	*psb*M	**0.0125**	1	2.2697
*atp*E	0.0632	1	1.5390	*psb*N	1	0.1632	0.9854
*atp*F	0.0888	1	1.3559	*psb*T	1	0.1193	1.1842
*atp*H	1	**0.0210**	2.0561	*psb*Z	1	0.0783	1.4253
*atp*I	1	0.0622	1.5477	*rbc*L	1	**0.0000**	4.5278
*ccs*A	1	0.1785	0.9248	*rpl*2	1	**0.0031**	2.7867
*cem*A	1	0.2453	0.6915	*rpl*14	1	**0.0234**	2.0097
*chl*B	1	**0.0002**	3.6305	*rpl*16	1	**0.0463**	1.6957
*chl*L	1	**0.0039**	2.7022	*rpl*20	1	**0.0359**	1.8161
*chl*N	1	**0.0000**	5.9654	*rpl*22	1	**0.0057**	2.5720
*clp*P	0.4634	1	0.0920	*rpl*23	1	0.2150	0.7919
*inf*A	1	0.1554	1.0177	*rpl*32	1	0.1692	0.9613
*mat*K	1	0.1628	0.9871	*rpl*33	1	0.0695	1.4893
*pet*A	1	**0.0140**	2.2233	*rpl*36	1	0.1550	1.0194
*pet*B	1	**0.0022**	2.9021	*rpo*A	1	0.0691	1.4928
*pet*D	1	0.1025	1.2742	*rpo*B	1	**0.0000**	4.2298
*pet*G	1	0.0697	1.4881	*rpo*C1	1	**0.0103**	2.3448
*pet*L	0.0791	1	1.4197	*rpo*C2	1	**0.0017**	2.9858
*pet*N	1	0.1594	0.9990	*rps*2	1	0.0583	1.5804
*psa*A	1	**0.0000**	5.5339	*rps*3	1	**0.0019**	2.9447
*psa*B	1	**0.0000**	5.3084	*rps*4	1	**0.0062**	2.5373
*psa*C	1	0.1711	0.9537	*rps*7	**0.0130**	1	2.2541
*psa*I	**0.0482**	1	1.6756	*rps*8	1	0.3590	0.3619
*psa*J	1	0.4104	0.2270	*rps*11	1	0.0638	1.5339
*psa*M	0.4967	1	0.0084	*rps*12	1	0.1016	1.2795
*psb*A	1	**0.0004**	3.4212	*rps*14	1	0.0984	1.2977
*psb*B	1	**0.0003**	3.5747	*rps*15	1	**0.0070**	2.4949
*psb*C	1	**0.0002**	3.6848	*rps*18	1	0.1515	1.0343
*psb*D	1	**0.0045**	2.6582	*rps*19	1	0.0863	1.3722
*psb*E	1	0.0642	1.5310	*ycf*1	**0.0000**	1	4.0848
*psb*F	0.0587	1	1.5769	*ycf*2	**0.0156**	1	2.1793
*psb*H	**0.0124**	1	2.2732	*ycf*3	1	0.0813	1.4051
*psb*I	1	0.1810	0.9151	*ycf*4	1	0.0531	1.6274
*psb*J	0.0916	1	1.3389				

Table 3.4. Shimodaira-Hasegawa test results. Results of significance testing for topology comparisons of the full alignment (Fig. 3.2) versus the three other largest data partitions (Fig. 3.3). For each set of comparisons, the first row represents comparison of unmodified maximum likelihood topologies. In the second and third rows the positions of *P. krempfii* and *P. albicaulis – P. lambertiana* N – *P. parviflora* were modified as indicated. Topologies that differ within a comparison are indicated in **bold**. Significant topological differences at $P < 0.05$ are indicated with an asterisk.

***P. krempfii* topologies**	***P. albicaulis, P. lambertiana* N, *P. parviflora* topologies**	**P- value**
Fig. 3.2 vs. 3.3A	**3.2 vs. 3.3A**	0.011*
Fig. 3.2 vs. 3.2	**3.2 vs. 3.3A**	0.153
Fig. 3.2 vs. 3.3A	3.2 vs. 3.2	0.024*
Fig. 3.2 vs. 3.3B	**3.2 vs. 3.3B**	0.351
Fig. 3.2 vs. 3.3A	**3.2 vs. 3.3B**	0.063
Fig. 3.2 vs. 3.3A	3.2 vs. 3.2	0.063
Fig. 3.2 vs. 3.3C	**3.2 vs. 3.3C**	0.005*
Fig. 3.2 vs. 3.2	**3.2 vs. 3.3C**	0.050
Fig. 3.2 vs. 3.3C	3.2 vs. 3.2	0.024*

Table 3.5. Estimated divergence times of poorly resolved nodes
All divergence time estimates assume a chloroplast mutation rate of 3.26×10^{-10} substitutions / site / year. Coalescent units reported are based on either high (100000) or low (10000) effective population (N_e) sizes. Maximum likelihood (ML) branch lengths are shown as substitutions/site. Estimated divergence times are presented in years (top), generations (middle) and coalescent units for high/low N_e (bottom).

Node	ML branch length (substitutions/site)	Estimated divergence time
P. krempfii - section *Quinquefoliae*	0.000370	1126539 22531 0.113/1.13
P. parviflora - *P. albicaulis*	0.000144	442057 8841 0.044/0.44
P. albicaulis - *P. lambertiana* N	0.000030	92095 1842 0.009/0.09
P. cembra - *P. koraiensis* / *sibirica*	0.000085	260936 5219 0.026/0.26

LITERATURE CITED

Alfaro ME, Zoller S, Lutzoni F. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov Chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol., 20:255-266.

Bouille M, Bousquet J. 2005. Trans-species shared polymorphisms at orthologous nuclear gene loci among distant species in the conifer *Picea* (Pinaceae): implications for the long-term maintenance of genetic diversity in trees. Am. J. Bot., 92:63-73.

Chung SM, Gordon VS, Staub JE. 2007. Sequencing cucumber (*Cucumis sativus* L.) chloroplast genomes identifies differences between chilling-tolerant and-susceptible cucumber lines. Genome, 50:215-225.

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. Nature Meth., 5:887-893.

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res., 36:e122.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nature Rev. Genet., 6:361-375.

Douady CJ, Delsuc F, Boucher Y, Doolittle WF, Douzery EJ. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol., 20:248-254.

Drescher A, Ruf S, Calsa T, Carrer H, Bock R. 2000. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J., 22:97-104.

Eckert AJ, Hall BD. 2006. Phylogeny, historical biogeography, and patterns of diversification for *Pinus* (Pinaceae): phylogenetic tests of fossil-based hypotheses. Mol. Phylogenet. Evol., 40:166-182.

Erickson DL, Spouge J, Resch A, Weigt LA, Kress JW. 2008. DNA barcoding in land plants: developing standards to quantify and maximize success. Taxon, 57:1304-1316.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 39:783-791.

Fishbein M, Hibsch-Jetter C, Oltis DES, Hufford L. 2001. Phylogeny of Saxifragales (angiosperms, eudicots): analysis of a rapid, ancient radiation. Syst. Biol., 50:817-847.

Gernandt DS, Hernández-León S, Salgado-Hernández E, Rosa JAPdl. 2009. Phylogenetic relationships of *Pinus* subsection *Ponderosae* inferred from rapidly evolving cpDNA regions. Syst. Bot., 34:481-491.

Gernandt DS, Lopez G, Garcia SO, Liston A. 2005. Phylogeny and classification of *Pinus*. Taxon, 54:29-42.

Gernandt DS, Magallon S, Geada Lopez G, Zeron Flores O, Willyard A, Liston A. 2008. Use of simultaneous analyses to guide fossil-based calibrations of Pinaceae phylogeny. Int. J. Plant Sci., 169:1086-1099.

Gibbs MJ, Armstrong JS, Gibbs AJ. 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences, 16:573-582.

Gilbert MTP, Drautz DI, Lesk AM, Ho SYW, Qi J, Ratan A, Hsu CH, Sher A, Dalen L, Gotherstrom A. 2008. Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes. Proc Natl Acad Sci USA, 105:8327.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotech., 27:182-189.

Herman DS, Hovingh GK, Iartchouk O, Rehm HL, Kucherlapati R, Seidman JG, Seidman CE. 2009. Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. Nature Meth., 6:507-510.

Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J. 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res., 18:802-809.

Hillis DM, Bull JJ. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol., 42:182-192.

Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, Bank Mvd, Chase MW, Cowan RS, Erickson DL, Fazekas AJ, *et al.* 2009. A DNA barcode for land plants. Proc Natl Acad Sci USA, 106:12794-12797.

Hudson ME. 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. Molecular Ecology Resources, 8:3-17.

Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack J, Muller KF, Guisinger-Bellian M, Haberle RC, Hansen AK. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA, 104:19369.

Lidholm J, Gustafsson P. 1991. The chloroplast genome of the gymnosperm *Pinus contorta*: a physical map and a complete collection of overlapping clones. Curr. Genet., 20:161-166.

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*. Mol. Ecol., 16:3926-3937.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet., 24:133-141.

Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences, 16:562-563.

Martin DP, Posada D, Crandall KA, Williamson C. 2005a. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Research & Human Retroviruses, 21:98-102.

Martin DP, Williamson C, Posada D. 2005b. RDP2: recombination detection and analysis from sequence alignments, 21:260-262.

Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci USA, 104:19363.

Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Folta KM, Soltis DE. 2006. Rapid and accurate pyrosequencing of angiosperm plastid genomes. BMC Plant Biol., 6:17.

Neubig KM, Whitten WM, Carlsward BS, Blanco MA, Endara L, Williams NH, Moore M. 2009. Phylogenetic utility of *ycf*1 in orchids: a plastid gene more variable than *mat*K. Plant Syst. Evol., 277:75-84.

Nylander JAA. 2004. MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. 2008. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res., 18:2024.

Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new geminiviruses by frequent recombination. Virology, 265:218-225.

Palmer JD. 1985. Comparative organization of chloroplast genomes. Annu. Rev. Genet., 19:325-354.

Patenaude NJ, Portway VA, Schaeff CM, Bannister JL, Best PB, Payne RS, Rowntree VJ, Rivarola M, Baker CS. 2007. Mitochondrial DNA diversity and population structure among southern right whales (*Eubalaena australis*). J. Hered., 98:147-157.

Philippe H, Frederic, D., Henner, B., and Lartillot, N. 2005. Phylogenomics. Annu. Rev. Ecol. Evol. Syst., 36:541-C-542.

Pollard DA, Iyer VN, Moses AM, Eisen MB, McAllister BF. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genet, 2:e173.

Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, *et al.* 2007. Multiplex amplification of large sets of human exons. Nature Meth., 4:931-936.

Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics, 14:817-818.

Price RA, Liston A, Strauss SH. 1998. Phylogeny and Systematics of *Pinus*. In: Richardson DM editor. Ecology and Biogeography of Pinus. Cambridge, Cambridge University Press, p. 49-68.

Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. 2008. A large genome center's improvements to the Illumina sequencing system. Nature Meth., 5:1005-1010.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, 19:1572-1574.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol, 16:1114-1116.

Smith JM. 1992. Analyzing the mosaic structure of genes. J. Mol. Evol., 34:126-129.

Stamatakis A. 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol., 57:758-771.

Suzuki Y, Glazko GV, Nei M. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc Natl Acad Sci USA, 99:16138-16143.

Swofford DL. 2000. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland, Massachusetts, Sinauer Associates.

Syring J, Farrell K, Businsky R, Cronn R, Liston A. 2007. Widespread genealogical nonmonophyly in species of *Pinus* subgenus *Strobus*. Syst. Biol., 56:163-181.

Syring J, Willyard A, Cronn R, Liston A. 2005. Evolutionary relationships among *Pinus* (Pinaceae) subsections inferred from multiple low-copy nuclear loci. Am. J. Bot., 92:2086-2100.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol., 24:1596-1599.

Wang XR, Szmidt AE, Nguyên HN. 2000. The phylogenetic position of the endemic flat-needle pine *Pinus krempfii* (Pinaceae) from Vietnam, based on PCR-RFLP analysis of chloroplast DNA. Plant Syst. Evol., 220:21-36.

Wang XR, Tsumura Y, Yoshimaru H, Nagasaka K, Szmidt AE. 1999. Phylogenetic relationships of Eurasian pines (*Pinus*, Pinaceae) based on chloroplast *rbc*L, *mat*K, *rpl*20-*rps*18 spacer, and *trn*V intron sequences. Am. J. Bot., 86:1742-1753.

Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol., 22:258-265.

Whittall JB, Syring J, Parks M, Buenrostro J, Dick C, Liston A, Cronn R. 2009. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. Mol. Ecol.:in press.

Willyard A, Syring J, Gernandt DS, Liston A, Cronn R. 2007. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. Mol. Biol. Evol., 24:90-101.

Wortley AH, Rudall PJ, Harris DJ, Scotland RW. 2005. How Much Data are Needed to Resolve a Difficult Phylogeny? Case Study in Lamiales. Syst. Biol., 54:697-709.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics, 20:3252-3255.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res., 18:821-829.

Newly Developed Primers for Complete *ycf*1 Amplification in *Pinus* (Pinaceae) Chloroplasts with Possible Family-Wide Utility

Matthew Parks, Aaron Liston and Richard Cronn

American Journal of Botany
Botanical Society of America
P.O. Box 299
St. Louis, Missouri
63166-0299
In press.

ABSTRACT

Primers were designed to amplify the highly variable locus *ycf*1 from all 11 subsections of *Pinus* to facilitate plastome assemblies based on short sequence reads as well as future phylogenetic and population genetic analyses. Primer design was based on alignment of 33 *Pinus* and four Pinaceae plastomes with mostly incomplete *ycf*1 sequences. Sanger sequencing of 12 *Pinus* accessions resulted in open reading frames ranging in size from 5.2 to 6.1 kbp. Highest sequence diversity was identified in two regions totaling 5.5 kbp aligned length which can be targeted in all pine subsections with three primer combinations. Preliminary results suggest the primers described also amplify homologous targets in the broader Pinaceae. The successful design and implementation of PCR primers spanning the large, variable locus *ycf*1 in *Pinus* represents the development of a valuable tool in pine genetic studies, and should facilitate studies throughout Pinaceae.

INTRODUCTION

'Next-generation' sequencing technologies, which feature nucleotide sequence output measured in millions to billions of base pairs, are revolutionizing DNA- and RNA-based research. However, the short read length characteristic of these technologies (currently tens to several hundreds of base pairs) often makes it difficult or impossible to accurately assemble and characterize regions highly divergent from available reference sequences. For phylogenetic or population genetic pursuits, such regions are often of great utility. In the genus *Pinus*, the locus *ycf*1 was identified as highly variable based on nearly complete plastome sequences assembled from Illumina short-read sequence data (Parks et al. 2009), and subsequently used to study the phylogeny of the species-rich subsection *Ponderosae* (Gernandt et al. 2009). Nonetheless, a fuller accounting of this locus is desirable, as the *ycf*1 sequences assembled by Parks et al. (2009) and Cronn et al. (2008) contained a substantial portion of undetermined positions (largely due to difficulties in assembly of short reads), and the analyses of Gernandt et al. (2009) used only ca. 20% of total *ycf*1 sequence length.

We designed 14 primers to allow amplification of the entire *ycf*1 locus from all 11 of the currently recognized *Pinus* subsections (Gernandt et al. 2005). Our sequencing efforts focused on 10 of the 11 subsections, as a complete *ycf*1 sequence is already available for subsection *Pinus* (GenBank record NC_001631.1, Wakasugi et al. 1994). In addition, although a complete *ycf*1 sequence is also available for subsection Strobus (*Pinus koraiensis*, GenBank record NC_004677.2), we sequenced two additional members of this subsection in order to verify a repetitive region reported in *P. koraiensis*. Amplicons were subsequently sequenced using Sanger technology and combined to assemble complete or nearly complete reading frames of *ycf*1 for each accession. Potential applicability of these primers to the broader Pinaceae was also investigated by alignment of *ycf*1 sequences to four non-*Pinus* members of Pinaceae and PCR amplifications using two primer pairs.

METHODS AND RESULTS

Total genomic DNA was extracted from frozen leaf or mega-gametophyte tissues of 12 accessions representing 10 of the 11 *Pinus* subsections sensu Gernandt et al. (2005) (Appendix 1) using the standard FastDNA extraction protocol (MP Biomedicals, Ohio, USA).

Fourteen primers (Table 4.1, Figure 4.1) were designed to cover the *ycf*1 locus based on conserved regions in an alignment of 33 *Pinus* species and four non-pine outgroups from the Pinaceae (Parks et al. 2009) (GenBank FJ899555-FJ899583, EU998739-EU998746). PCR amplifications using various combinations of these primers were performed with either Phusion DNA polymerase and Phusion buffer HF or Taq polymerase and Thermopol buffer (New England Biolabs, Massachusetts, USA). Generally, PCR reactions were carried out as: 30 sec 98° C (one cycle), 8 sec 98° C, 30 sec 55-59° C, 30 sec/kb 72° C (25-30 cycles), 5min 72° C (1 cycle), final hold temperature of 4° C. For problematic amplifications, several strategies were pursued, including: 1) varying annealing temperatures (max 60° C, min 52° C); 2) use of alternative buffer or additives (for example, Phusion GC buffer with 0.25 µl 100% DMSO / 50 µl reaction volume or addition of 0.1 µl BSA (10mg / ml) / 50 µl reaction volume when using Taq polymerase); 3) pairing primers herein with alternative primers designed to target the *Pinus* plastome in the vicinity of *ycf*1 (Cronn et al. 2008). Amplification success was determined by gel electrophoresis using 1% agarose gels stained with GelRed (ca. 1:30,000 v/v) (Phenix Research Products, North Carolina, USA) and run for 30 min at 110 V. Successful amplifications (i.e., strong, single bands) were submitted for Sanger sequencing to the University of Washington High-Throughput Genomics Unit (www.htseq.org) using the above described PCR primers for sequencing.

Quality-trimming of sequence reads was performed by the UW High-Throughput Genomics Unit based on a quality score cutoff of Q20 (corresponding to 99% confidence in a base call), which is a commonly used cutoff for initial quality filtering (Ewing and Green 1998, Richterich 1998). Reads were trimmed from their 5´ and 3´ ends until windows of 50 consecutive positions contained fewer than 10 base calls <Q20. At this point the remaining sequence, including all positions in the current window, was considered to have passed quality filtering, although some remaining positions were subject to manual masking in subsequent steps (see below). The resulting filtered sequences were manually aligned to their reference genome in BioEdit v7.0.5 (Hall 1999); in all cases, the original assembly of Cronn et al. (2008) or Parks et al. (2009) served as the reference for an accession's assembly. Assembly problems were identified by translation of the *ycf*1 reading frame as well as comparison of overlapping aligned reads where possible. Putative internal stop-codons, frame-shift mutations and discordant base calls between overlapping aligned reads or aligned reads and their reference were investigated through examination of sequence chromatograms and/or

resequencing. In some cases positions within the quality-filtered reads were masked by hand based on a combination of low quality score (<Q20) and discordance with overlapping reads or their previously published assemblies. Sequence polymorphism of final assemblies (measured as Watterson's θ) was evaluated using the program DnaSP v.5.10.00 (Librado and Rozas 2009).

Alignment of sequence reads allowed complete or nearly complete assembly of the *ycf*1 locus from 12 *Pinus* accessions representing all of the targeted *Pinus* subsections (Table 4.2). Assembled lengths of *ycf*1 ranged from ca. 5.2 to 6.2 kbp, and averaged over 500 bp longer in subgenus *Strobus* than in subgenus *Pinus* (Table 4.2). Areas of highest sequence diversity in aligned *Pinus ycf*1 were found between positions 500-3500 and 5500-8000 of the 8.3 kbp alignment. These regions are mostly bounded by the primers A to B (E to G in subgenus *Strobus*), and C to D, respectively (Figure 4.1).

Alignment of primer sequences to four non-pine members of Pinaceae (*Picea sitchensis*, *Larix occidentalis*, *Cedrus deodara*, *Abies firma*, GenBank identifiers included above) suggested that most of the primers should be effective throughout the family. For primers 1-4 and A-G, each outgroup accession on average differed at 1.18 ± 1.23 (SD) positions in the primer sequence, while the average distance of these differences from the primer's 3′ end was 10.21 ± 4.98 bp (SD). Using the described PCR strategies, all four non-pine species tested also successfully PCR-amplified with each of two primer combinations (ycf1.1/ycf1.3 and ycf1.2/ycf1.4), resulting in amplicons of expected size.

CONCLUSIONS

Based on our results, the primers described should be useful tools in studies employing *ycf*1 as a phylogenetic or population genetic marker in species of *Pinus*, and likely throughout Pinaceae. While full sequencing of this locus may require additional primers, a substantial portion of the most variable regions of this locus can be targeted with a small number of primers. In addition, the complete to nearly complete subsectional *ycf*1 sequences generated should aid in reference-guided assembly from next-generation sequencing data in future projects targeting Pinaceae chloroplast genomes or the *ycf*1 locus specifically.

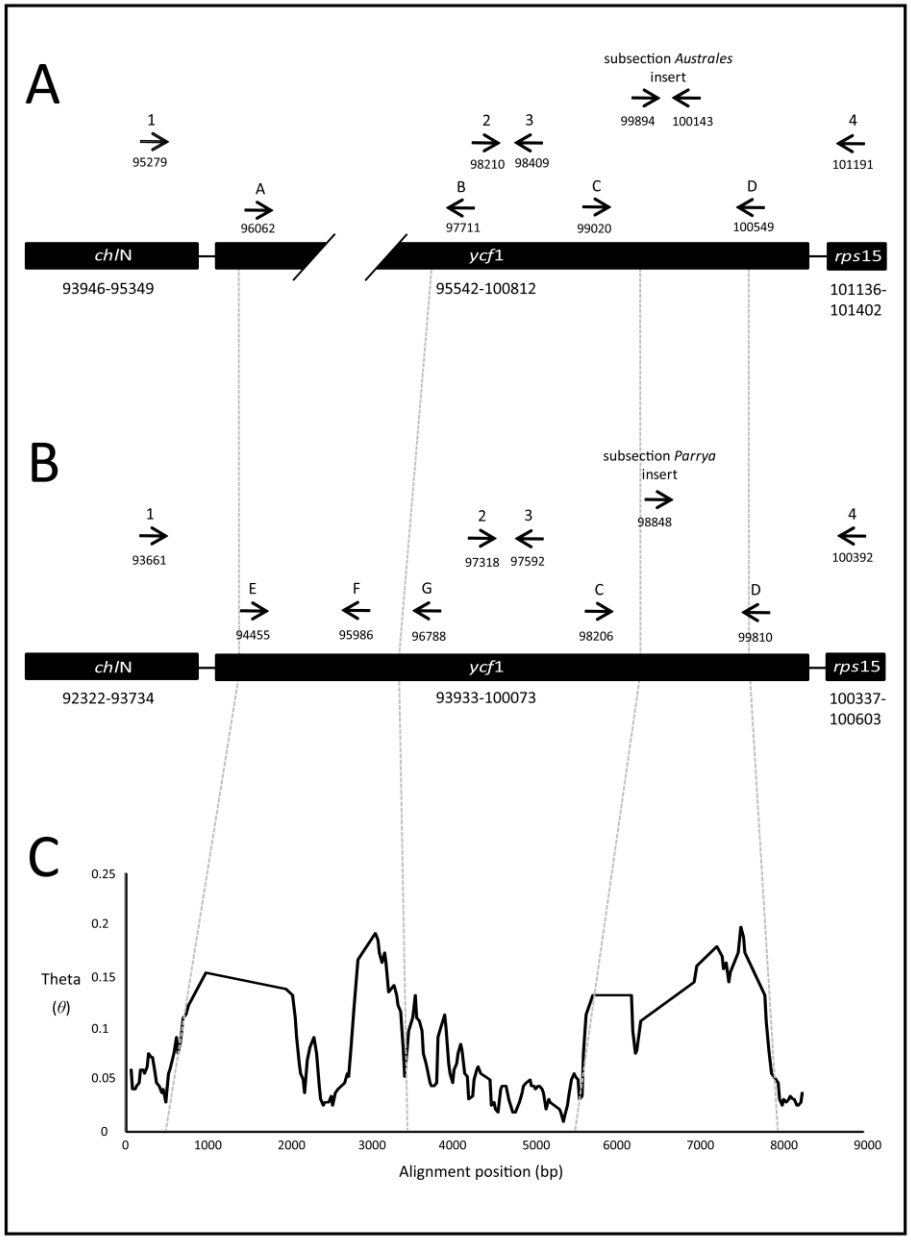


Figure 4.1. Map of primer locations used in ycf1 amplifications. A) Map of primer locations used in *ycf*1 amplifications in subgenus *Pinus*. Coordinates correspond to locations in the *Pinus thunbergii* chloroplast genome (Wakasugi et al. 1994); B) Map of primer locations used in *ycf*1 amplifications in subgenus *Strobus*. Coordinates correspond to locations in the *Pinus koraiensis* chloroplast genome (GenBank NC_004677.2). In both A) and B), primer coordinates represent 5′ end of the primer; gap in *ycf*1 of A) corresponds to a repetitive region of ca. 900 bp aligned length found in subgenus *Strobus* but not present in subgenus *Pinus*. C) Graph of Watterson's theta (θ) as measured in 100bp windows along the aligned length of *ycf*1 for *Pinus* accessions sequenced in this study and *Pinus thunbergii* (Wakasugi et al. 1994) and *Pinus koraiensis* (GenBank NC_004677.2). Light dotted lines indicate approximate locations of regions of high θ values in relation to primer locations in parts A) and B).

Table 4.1. Information for primers used in *ycf*1 amplifications and sequencing. Primers from Cronn et al. (2008) were used mainly as alternative primers for difficult to amplify accessions/regions.

Region	**Primer Name**	**Source**	**Sequence (5´ to 3´)**
*chl*N	ycf1.1	This paper	TAGATAACTTGGATCGGACCAC
*ycf*1	ycf1.2	This paper	TTCCTTTTCGTTTGAAGCCTT
*ycf*1	ycf1.3	This paper	TCTTATTCCTGTAGATCCCATCAAT
*rps*15	ycf1.4	This paper	GATCCTCTCTGTTTATCGGGAA
*ycf*1	ycf1.A	This paper	TGGGCGGTCATATTCTATT
*ycf*1	ycf1.B	This paper	TTAAGTTCCGACGATAATCTG
*ycf*1	ycf1.C	This paper	AAGATTTTGAAATTCGTCCTG
*ycf*1	ycf1.D	This paper	TACGACGTTTTGGAAGC
*ycf*1	ycf1.E	This paper	GGCGGTCATATTCTATTCAT
*ycf*1	ycf1.F	This paper	TGCCAATGCTCAGAGATA
*ycf*1	ycf1.G	This paper	CTCGGCATGATAACGTTT
*ycf*1	subs. *Australes* insert F	This paper	GAAGGAAACAACAAATGTTCTAG
*ycf*1	subs. *Australes* insert R	This paper	CATAACCCTGCAAATATTCG
*ycf*1	subs. *Parrya* insert F	This paper	GATCCGGATTAGATTTAAAATTCTGG
*trn*N-GUU	28F	(Cronn et al., 2008)	TTAACAGCCGACCGCTCTAC
*ycf*1	28R	(Cronn et al., 2008)	GTAGAGCGGTCGGCTGTTA
*ycf*1	29F	(Cronn et al., 2008)	TCCCGTATTAACAAGACTGGTG
*ycf*1	29R	(Cronn et al., 2008)	CCAGTCTTGTTAATACGGGATTT
*ycf*1	30F	(Cronn et al., 2008)	TTGGATCACGAAAAACCACA
*psa*C	30R	(Cronn et al., 2008)	TGTGGTTTTTCGTGATCCAA

Table 4.2. *ycf*1 sequencing and assembly success for accessions representing *Pinus* subsections. (P) and (S) indicate subgenus *Pinus* and *Strobus*, respectively.

Subsection	**Species**	**Estimated length of *ycf*1 (bp)**	**Estimated number of undetermined positions**
Australes (P)	*Pinus taeda*	5259	350
Contortae (P)	*P. contorta*	5547	0
Pinaster (P)	*P. canariensis*	5448	48
Pinaster (P)	*P. pinaster*	5472	48
Ponderosae (P)	*P. ponderosa*	5784	1
	average length (SD)	5502 (190)	
Balfourianae (S)	*P. aristata*	5772	2
Cembroides (S)	*P. monophylla*	6054	0
Gerardianae (S)	*P. gerardiana*	5781	497
Krempfianae (S)	*P. krempfii*	6222	0
Nelsoniae (S)	*P. nelsonii*	6036	0
Quinquefoliae (S)	*P. flexilis*	5901	0
Quinquefoliae (S)	*P. lambertiana*	6114	0
	average length (SD)	5983 (170)	

LITERATURE CITED

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res., 36:e122.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using Phred.II. Error probabilities. Genome Res., 8:186-194.

Gernandt DS, Hernández-León S, Salgado-Hernández E, Rosa JAPdl. 2009. Phylogenetic relationships of *Pinus* subsection *Ponderosae* inferred from rapidly evolving cpDNA regions. Syst. Bot., 34:481-491.

Gernandt DS, Lopez G, Garcia SO, Liston A. 2005. Phylogeny and classification of *Pinus*. Taxon, 54:29-42.

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

Librado P, Rozas J. 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics, 25:1451-1452.

Parks M, Cronn R, Liston A. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biology, 7:84.

Richterich P. 1998. Estimation of errors in "raw" DNA sequences: a validation study. Genome Res., 8:251-259.

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*. Proc Natl Acad Sci USA, 91:9794-9798.

Separating the Wheat from the Chaff: Mitigating the Effects of Noise in a Plastome Phylogenomic Data Set

Matthew Parks, Richard Cronn and Aaron Liston

In preparation for submission.

ABSTRACT

Through next-generation sequencing, the amount of sequence data potentially available for phylogenetic analyses has increased exponentially in recent years. Simultaneously, the risk of incorporating 'noisy' data with misleading phylogenetic signal has also increased, and may disproportionately influence the topology of weakly supported nodes and lineages with rapid radiations and/or elevated rates of evolution. We investigated the influence of phylogenetic noise in large data sets by applying two fundamental strategies, variable site removal and long-branch exclusion, to the phylogenetic analysis of a full plastome alignment of 107 species of *Pinus* and six Pinaceae outgroups. While high overall phylogenetic resolution resulted from inclusion of all data, three historically recalcitrant nodes remained conflicted. Close investigation of these nodes revealed dramatically different responses to data removal. Whereas topological resolution and bootstrap support for two clades peaked with removal of highly variable sites, the third clade resolved most strongly when all sites were included. Similar trends were observed using long-branch exclusion, but patterns were neither as strong nor as clear. When compared to previous phylogenetic analyses of nuclear loci and morphological data, the most highly supported topologies seen in our plastome analysis are consistent for the two clades gaining support from noise removal and long-branch exclusion, but inconsistent for the clade with highest support from the full data set. These results suggest that removal of noise in phylogenomic datasets can result not only in increased resolution for poorly supported nodes, but serve as a tool for identifying highly supported, but likely incorrect topologies. In addition, removal of variable sites appears to be more effective than long-branch exclusion for reducing the impact of noise in our data set.

INTRODUCTION

The potential influence of phylogenetic 'noise', i.e. random or misleading signal, in molecular phylogenetic studies has been recognized for over 30 years (Felsenstein 1978, Fitch 1979, Fitch 1984, Hendy and Penny 1989). Similarly, various strategies to identify and/or mitigate noise in datasets have been formulated, including measuring skewness in the distribution of phylogenetic trees (Huelsenbeck 1991, Hillis and Huelsenbeck 1992), quantifying incongruence between data partitions (Farris et al. 1995, but see, for example, Dowton and Austin 2002, Hipp et al. 2004), likelihood mapping (Strimmer and von Haeseler 1997), increasing taxon sampling (Pollock et al. 2002, Soltis et al. 2004), and profiling loci based on phylogenetic information content (Townsend 2007, Klopfstein et al. 2010, Townsend and Leuenberger 2011), among others. While the specific details of these strategies differ, ultimately the goal of each is to reduce the impact of noise on phylogenetic hypotheses by identifying and/or reducing the influence of misleading or 'noisy' positions and long-branch attraction artefacts (Bergsten 2005). Nonetheless, as next-generation technologies continue to bring about orders-of-magnitude increases in DNA sequence output and usher in an era of phylogenomics, the challenges associated with phylogenetic noise could yet temper gains in phylogenetic resolution resulting from increased taxon and sequence sampling (Delsuc et al. 2005, Philippe et al. 2005, Degnan and Rosenberg 2006, Jeffroy et al. 2006, Kubatko and Degnan 2007, Philippe et al. 2011). Although genomic-scale data sets are relatively novel, it is clear that inherent noise still impacts phylogenetic resolution, in particular in clades that have experienced rapid divergence or radiation events, as well as lineages with elevated rates of evolution and/or long periods of genetic isolation (i.e., "long branches") (Phillips et al. 2004, Soltis et al. 2004, Stefanovic et al. 2004, Brinkmann et al. 2005, Degnan and Rosenberg 2006, Rokas and Carroll 2006, Kubatko and Degnan 2007).

While perhaps the greater drive behind phylogenomics has been the elucidation and incorporation of nuclear genome sequences into phylogenetic analyses, organellar genomes still represent tractable and informative systems for phylogenetic analyses. In plants in particular, nuclear genomes typically present daunting challenges from both a sequencing and analytical standpoint due to issues such as overall genome size, difficulty in determining orthology and the presence of homeologous alleles, all of which are often partly or wholly attributable to polyploidization (Wendel 2000, Adams and Wendel 2005). Even in the absence of polyploidization, plant nuclear genomes can be difficult to interrogate due to their size. For

example, nuclear genome sizes in *Pinus*, a diploid genus, range at least from 22.1 to 36.9 ρg, corresponding to ca. 7-12X the size of a human genome (Grotkopp et al. 2004). Likewise, plant mitochondria remain challenging genomic targets. Although the first plant mitochondrial genome was sequenced relatively shortly after the first plant chloroplast genome (Shinozaki et al. 1986, Oda et al. 1992), to date relatively few plant mitochondrial genome sequences have been fully sequenced as they are much more variable in size than their plastid counterparts, mostly due to the relatively transient incorporation and shuffling of both nuclear and chloroplast nomad sequences (Fauron et al. 2004), as well as apparently species-specific sequences of unknown function (Kubo and Newton 2008). In addition, nucleotide mutation rates in plant mitochondria are typically considerably slower than those of both plastids and plant nuclear genomes (Wolfe et al. 1987, Palmer 1990, but see Palmer et al. 2000). As a result, structural mutations may be more informative than single nucleotide polymorphisms for phylogenetic analyses, yet these types of mutations are currently difficult to detect with the relatively short read length common to next generation sequencing platforms (Alkan et al. 2011). Compared to plant nuclear and mitochondrial genomes, chloroplast genomes are reasonable targets for phylogenomic analyses for several reasons, including sufficient size and complexity (typically 120-160 kbp in length, containing ca. 130 genes) to mark most evolutionary events, yet levels of conservation that allow relatively easy determination of orthology even across broad time scales. In addition, a moderate mutation rate (Palmer 1990) and a haploid state result in relatively smaller effective population sizes and increased likelihood of fixing phylogenetic signal during speciation events and divergence (Birky 1978).

It is no surprise then, due to both technical and biological considerations, that chloroplast sequences are still the most commonly used markers in plant phylogenetic studies. Nonetheless, many studies tend to rely on relatively small portions of the chloroplast genome, and relatively few studies have applied plastome-scale sequences to phylogenetic questions (Goremykin et al. 2005, Leebens-Mack et al. 2005, Cai et al. 2006, Jansen et al. 2006, Jansen et al. 2007, Moore et al. 2007, Parks et al. 2009, Lin et al. 2010). This is particularly true at low taxonomic levels (Parks et al. 2009), while the majority of plastome-level phylogenetic analyses have focused on clarifying relationships at familial and ordinal levels. Considering the potential impact of phylogenetic noise in phylogenomic analyses (Delsuc et al. 2005, Philippe et al. 2005, Degnan and Rosenberg 2006, Jeffroy et al. 2006, Kubatko and Degnan 2007, Goremykin et al. 2010, Philippe et al. 2011), it seems appropriate to explore the effect

of noise on plastome-scale datasets (Goremykin et al. 2009), particularly as they become widespread in plant phylogenetic analyses in the near future and more commonly applied to investigations at low taxonomic levels. Further, although representing a single linkage group, mutation rate varies between different regions of the plastome (Shaw 2005, Shaw et al. 2007, Parks et al. 2009), and so the potential for misleading signal certainly exists when using full plastomes to delineate evolutionary events over varying time-scales.

The genus *Pinus*, consisting of ca. 110 species distributed primarily throughout the northern hemisphere, is an excellent system in which to investigate the impact of phylogenetic noise as it contains evolutionary patterns ranging from deep divergence events to apparent rapid and relatively shallow radiations. In addition, the moderate size of the genus facilitates extensive taxon sampling. *Pinus* is represented by a relatively well-documented fossil record reaching back over 100 million years (Millar 1998, Klymiuk et al. 2011) and has been the focus of a large body of phylogenetic work, including studies based in morphology (Little and Critchfield 1969, Frankis 1993, Ortiz Garcia 1999, Gernandt et al. 2005, Gernandt et al. 2008), crossability (Critchfield 1966, Little and Critchfield 1969, Critchfield 1975, 1986) and molecular data, including restriction fragment analyses (Strauss and Doerksen 1990, Krupkin et al. 1996) and both nuclear (Liston et al. 1999, Liston et al. 2003, Syring et al. 2005, Palmé et al. 2009) and chloroplast sequence data (Wang et al. 1999, Wang et al. 2000, Geada Lopez et al. 2002, Zhang and Li 2004, Gernandt et al. 2005, Eckert and Hall 2006, Gernandt et al. 2008, Parks et al. 2009). The most recent full taxonomic treatment of *Pinus* (Gernandt et al. 2005) recovered a well-supported systematic framework consisting of two subgenera (*Pinus* and *Strobus*), four sections (sections *Pinus* and *Trifoliae* in subgenus *Pinus*, sections *Parrya* and *Quinquefoliae* in subgenus *Strobus*) and 11 subsections (Figure 5.1) that is widely accepted today. However, while nearly complete plastome sequences for a subset of pine species support this framework and result in increased resolution across much of the genus (Parks et al. 2009), there remain some taxa with poor resolution and/or incongruence between chloroplast-based and nuclear- or morphology-based analyses. In particular, subsections *Contortae* and *Krempfianae*, as well as a clade of the two closely related species *Pinus merkusii* and *P. latteri* each demonstrate these conflicts. In the present study, we investigated whether poor or conflicting resolution in these clades was due to the influence of phylogenetic noise using two fundamental and complementary strategies: removal of highly variable alignment positions and long-branch exclusion. These strategies were applied to the

phylogenetic analysis of a full-plastome alignment which included most of the world's pine species and several Pinaceae outgroups. While responses to noise removal differed between these clades, each case provided insight into both the general patterns of response to noise removal in a phylogenomic dataset as well as specific characteristics of *Pinus* evolutionary history.

METHODS AND MATERIALS

Accessions Used in Study.

A total of 113 accessions were included in the alignment and subsequent analyses described below, including 37 *Pinus* and Pinaceae accessions reported by Cronn et al. (2008) and included in Parks et al. (2011) (GenBank FJ899555-FJ899583, EU998739-998746, NC_001631.1 and NC_004677.2) and the plastome sequence of *Cathaya argyrophylla* reported by Lin et al. (2010) (GenBank AB547400.1) (Appendix Table 5.1). The 70 novel plastome accessions included in analyses were sequenced and assembled as described in the following two sections.

Genomic DNA Extraction, Chloroplast Enrichment and Sequencing.

For plastome accessions novel to this study, total genomic DNA was extracted from frozen leaf or mega-gametophyte tissues using the FastDNA extraction protocol (MP Biomedicals, Ohio, USA). In several cases (*Pinus chiapensis*, *P. cembroides*, *Pinus dabeshanensis*, *P. discolor*, *P. douglasiana*, *P. edulis*, *P. hwangshanensis*, *P. massoniana*, *P. pumila*, and *P. sabiniana*), genomic DNA yield was insufficient for sequence preparation, so extracts were amplified by whole genome amplification with random hexamer priming (Pan et al. 2008). Genomic libraries were prepared following the Illumina protocol (Illumina 2007), with fragmentation performed using a BioRuptor Sonicator (Diagenode, Inc., Denville, NJ, USA) (setting 'high' for 5-30 one minute cycles). Adapters ligated to genomic fragments carried unique 4 bp 'barcodes' at their 3′ ends for multiplex sequencing as described in Cronn et al. (2008). Agarose gel size-selected (300-700 bp), adapter-ligated libraries were enriched through 12-18 cycles of PCR using Phusion DNA polymerase and HF Buffer (New England Biolabs, Ipswich, MA, USA) and standard Illumina paired-end primers, and quantified using a Nanodrop 1000 (ThermoFisher Scientific, Wilmington, DE, USA).

Solution-based enrichment of the chloroplast portion of genomic libraries followed the general methods of Gnirke et al. (2009) and is described in detail in Cronn et al. (Manuscript in preparation). In brief, enrichments were performed as follows. Chloroplast probe pools were synthesized by first PCR-amplifying the entire plastome of a member of *Pinus* subgenus *Pinus* (*Pinus thunbergii*), using the methods described in Cronn et al. (2008). In addition, plastome regions unique to *Pinus* subgenus *Strobus* were amplified from *P. koraiensis* to account for regions not present in the *Pinus* subgenus *Pinus* plastome. PCR products were quantified using a Nanodrop 1000 and pooled in an equimolar mix. Pooled amplicons were blunted-ended and subsequently ligated into 'concaterpillars' (Quick Blunting Kit and Quick Ligation Kit, New England Biolabs, Ipswich, MA, USA) and cleaned with Agencourt AMPure beads (Beckman-Coulter Genomics, Danvers, MA, USA). 'Concaterpillar' probe pools were denatured into single-stranded product using 0.4 N KOH, and then amplified and biotinylated in a single incubation of 18 hours at 30° C in the presence of 5′-end biotinylated random hexamers, 0.4 mM biotin-14-dCTP stock, 1 mM dNTPs and ϕ29 DNA polymerase. After cleaning by ethanol precipitation, this procedure typically yielded pools consisting of 10-25 µg of large (tens of kbp in length) biotinylated chloroplast probe. Hybridization reactions were carried out in 40 µl volumes and contained 0.5 µg probe and 0.5-1 µg of either a single enriched genomic library or equimolar-pooled 4-plex genomic libraries; Denhardts solution (Invitrogen, Inc., Carlsbad, CA, USA) and lambda DNA (New England Biolabs, Ipswich, MA, USA) were used as blocking agents to minimize binding of non-target DNA to probes. Reactions were heated to 95° C for 10 minutes, and subsequently incubated at 65° C for 64-72 hours. After incubation, hybridization products were captured using MagnaSphere streptavidin-coated paramagnetic beads (Promega, Inc., Madison, WI, USA) suspended in Sodium-Tris-EDTA buffer after equilibration with 2X Casein blocking buffer. Capture reactions were incubated for 30 minutes at room temperature, after which the streptavidin-probe-target DNA complexes were captured through magnetization and then washed four times at 65° C in the presence of 0.1% SDS and 1X, 1X, 0.5X and 0.1X SSC for 15, 10, 10 and 10 minutes, respectively. Enriched hybrids were eluted from the paramagnetic beads in 50 µl dH2O at 80° C for 10 minutes and PCR-amplified over 12-18 cycles using Phusion-Flash PCR Master Mix (New England Biolabs, Ipswich, MA, USA) and standard Illumina paired end primers. After PCR enrichment, libraries were cleaned and subsequently quantified using

the Nanodrop 1000, and size-confirmed using either gel electrophoresis or the Agilent 2100 BioAnalyzer (Agilent, Santa Clara, CA, USA).

The molarity of enriched libraries was estimated by their concentration and average fragment size, after which the libraries were submitted for sequencing singly or in barcode-specified multiplex pools ranging in size from four to 16 accessions. Most individual samples or multiplex pools were submitted to the Oregon State University Center for Gene Research and Biocomputing (OSU CGRB) (http://www.cgrb.oregonstate.edu/) for sequencing on the Illumina GAII sequencer, although several individual samples were submitted to the FAS Center for Systems Biology at Harvard University (http://sysbio.harvard.edu/csb/) for sequencing on an Illumina GAIIx sequencer. Libraries were loaded at a concentration of 5-7 ρM and sequenced in 60 or 80 bp single-end sequencing reactions. Cluster formation, primer hybridization and sequencing reactions followed Illumina protocols (Illumina 2007). Image analysis, base-calling and error estimation were performed using the Illumina GA Pipeline version 1.5.

Plastome Assembly from Microreads.

To initially determine enrichment of read pools, all reads containing Illumina adapter sequence were removed from read pools and the remaining reads were sorted by barcode using two Perl scripts, sort_fastq.pl and bcsort_fastq_se.pl (available at http://brianknaus.com). The proportion of reads representing the chloroplast was checked using the program BLAT (Kent 2002) with default settings and a reference of either *Pinus thunbergii* or *P. koraiensis* for accessions in subgenus *Pinus* or *Strobus*, respectively.

Reference-guided assembly of microreads was facilitated using a pipeline of five scripts called "alignreads", as described in (Straub et al. 2011). In this series, assembly of microreads into contigs is performed by YASRA (Ratan 2009), which assembles contiguous sequences (contigs) by iteratively aligning sequence reads to a reference genome using the lastz alignment algorithm (Harris 2007). The alignment of assembled contigs is then refined using NUCmer and Delta-Filter of the MUMmer 3.0 suite (Kurtz et al. 2004), and the resulting alignment information is paired with the original contigs and read depth information from YASRA, to be converted into an aligned consensus sequence using sumqual.py and qualtofa.py. The latter allows user-specified masking of contig positions based on read depth

and base call proportion. Both sumqual.py and qualtofa.py are available for download at http://milkweedgenome.org; YASRA and MUMmer are available online at http://www.bx.psu.edu/miller_lab/ and http://mummer.sourceforge.net/, respectively.

For assembly of the novel plastome sequences reported in this paper, subsectional references reported in Parks et al. (2011) were used (Appendix Table 5.1). The alignment of assembled contigs was checked and adjusted manually in BioEdit 7.0.9 (Hall 1999). Aligned contig positions matching the reference were masked if fewer than five overlapping reads and less than 80% of all reads overlapping to form the contig at that position agreed with the reference; aligned positions called as SNPs were similarly masked, but required a minimum coverage depth of 20 aligned reads and 80% call proportion.

Alignment and Quality Screening of Assemblies

Plastome assemblies were aligned in MAFFT v.6.240 (Katoh et al. 2005), using gap opening and extension penalties of 2.0 and 0.1, respectively. Alignments were subsequently manually adjusted and annotated in BioEdit 7.0.9. The assemblies of exonic regions were checked and adjusted as necessary by translation to identify potential misassemblies, as represented by internal stop codons and/or frameshift mutations.

Novel plastome sequences were quality-screened at this point by level of completion and relative similarity to the subsectional reference used in their assembly. Specifically, assemblies were discarded from further analyses if they were estimated to be less than 80% complete, or if the pairwise distance to their subsectional assembly reference was greater than two times the standard deviation of all pairwise distances between assembled members of their subsection and the subsectional reference. The latter measure was taken to diminish the potential effect of noise resulting from poor assemblies, for example resulting from low coverage or capture of divergent paralogous copies of chloroplast regions residing in the nuclear or mitochondrial genome. In addition, several assemblies were discarded due to poor overall assembly quality as evidenced by highly divergent exon/protein sequences and divergence from Sanger-sequenced plastome regions of the same species. Previously published *Pinus* plastome sequences were used only if they exceeded 80% estimated sequence completion.

Phylogenetic Analysis of Full Plastome Alignment

Phylogenetic analyses of the complete alignment were completed through the Cipres Science Gateway (http://www.phylo.org/) using RAxML-HPC2 (Stamatakis 2008) and MrBayes (Ronquist and Huelsenbeck 2003), both on the available teragrid. Likelihood analysis in this case was performed under the GTRGAMMA model, with the number of bootstrap replicates automatically determined under the recommended autoMRE option. Bayesian analyses were performed under the same model of evolution. Each analysis consisted of two runs with four chains each (three hot and one cold chain), run for 10,000,000 generations with trees sampled every 1000 generations, and the first 25% of trees discarded as burn-in. Stationarity was evaluated by graphing –lnL of trees across all generations and by requiring the standard deviation of the two runs to be less than 0.05. All trees were combined from both runs past the point of stationarity to determine topology and support through the majority rule consensus tree using PAUP* v.4.0b10 (Swofford 2000). Parsimony analysis was performed with PAUP* v.4.0b10, under heuristic search with ten repetitions of random sequence addition, tree bisection and reconnection branch swapping and 100 bootstrap replicates

Evaluation of the Impact of Variable Site Removal

Variable sites in the full plastome alignment were identified and ranked using the script sorter.pl (Goremykin et al. 2010), which quantifies the observed variability (OV) of each position in an alignment as:

$OV = sum(1...k)\{d_{ij}\}/k$

where:

k = the number of all possible pairwise comparisons between accessions in an alignment, excluding accessions with a gap at the position considered

d_{ij} = the score of character variability (0 for match, 1 for mismatch) in each of k pairwise comparisons of accessions in the alignment

Variable sites were then serially removed from the alignment in 100 site partitions using the script sorter.pl, resulting in two series of data partitions. The first series (FA) consisted of the

full alignment minus the most variable 100, 200, 300,...,25000 sites, while the second series (VS) consisted of the most 100, 200, 300,...,25000 variable sites.

Phylogenetic analyses on all FA and VS data partitions were run through OSU CGRB GENOME Cloud computing resources (http://bioinfo.cgrb.oregonstate.edu) using RAxML-VI-HPC v.2.2.3 (Stamatakis 2008) primarily under the GTRGAMMA model (some of the larger partitions were run under the GTRCAT model due to time constraints), with 100 bootstrap replicates. The resulting highest likelihood tree from each FA partition and its corresponding VS partition were then compared using the branch score metric (BSM) (Kuhner and Felsenstein 1994) and partition metric (PM) (Robinson and Foulds 1981) as implemented in the treedist executable of Phylip v.3.69 (Felsenstein 2005).

Topology and bootstrap support in relation to BSM and PM values were investigated in depth for three taxa with historically poorly resolved phylogenetic positions: 1) subsection *Contortae*, consisting of *Pinus contorta*, *P. banksiana*, *P. clausa* and *P. virginiana*, 2) the monotypic subsection *Krempfianae*, consisting of the morphologically distinctive flat-needled *P. krempfii*, and 3) the southeast Asian clade consisting of *P. merkusii* and *P. latteri* (Figure 5.1). For these analyses, bootstrap values for the nodes immediately ancestral to all three taxa were recorded for each FA partition, as these nodes represented the resolution between disputed alternative placements of each taxon (Figure 5.1). In addition, bootstrap values supporting the monophyly of subsection *Contortae* and the *P. merkusii*/*P. latteri* clades were recorded for each FA partition.

Evaluation of the Impact of Long-Branch Exclusion

As a general rule, the *Pinus* phylogeny contains relatively long branches (substantial divergence) separating the two subgenera and four sections, but relatively short branches (low divergence) within subsections (Gernandt et al. 2005, Parks et al. 2009). As a result, to remove long branches it is necessary in most cases to remove entire clades at the subsectional level or higher. Because of this and due to the conflicting topologies of interest residing at the subsectional level, long branches were excluded in the following manners: 1) all six Pinaceae outgroups were removed prior to phylogenetic analyses, 2) only the subgenus of interest was included in the analyses, and 3) only the section of interest and one member of the neighboring section were included in analyses. For the most exclusive strategy, *Pinus*

thunbergii (NC_001631.1), *P. monophylla* (EU998745.4) and *P. ponderosa* (FJ899555.2) were used as outgroups for sections *Trifoliae*, *Quinquefoliae* and *Pinus*, respectively. Maximum likelihood phylogenetic analyses were performed as described above for the full alignment for each strategy of long-branch exclusion on each of three partition sizes of interest (full alignment, FA.136665, FA.133065, as discussed in Results).

Impact of Noise-Removal Strategies on Saturation

To gain further insight into the impact of variable site removal and long-branch exclusion on saturation in our data matrix (i.e., the history of multiple nucleotide state changes at individual sites), pairwise genetic distances between all accessions were determined in MEGA4 (Tamura et al. 2007) both without correction and with application of a Jukes-Cantor correction. The correlation of these values was determined by linear regression for each of three partition sizes of interest (full alignment, FA.136665, FA.133065, as discussed in Results) and for each strategy of long-branch exclusion. The slope of the regression line was taken as indicative of the level of saturation present in the dataset, such that higher values for corrected pairwise distances relative to uncorrected distances correspond to higher levels of saturation (Jeffroy et al. 2006, Rodríguez-Ezpeleta et al. 2007).

RESULTS

Sequence Assembly and Alignment.

After quality/chastity filtering through the Illumina GA Pipeline v. 1.5 and removal of adapter sequences, read pools for successfully assembled plastome sequences averaged 1.77 ± 0.76 million reads per accession, while chloroplast reads accounted for 56.83 ± 13.85% of these reads on average (Appendix Table 5.1). Novel assembled plastome sequences averaged 117157 ± 3634 bp in length, and were estimated to be 98.1 ± 2.5% complete on average after masking (Appendix Table 5.1). The alignment of all successfully assembled plastome sequences, including 107 *Pinus* accessions and six Pinaceae outgroups, resulted in 141265 aligned sites.

Variable Sites

Variable sites were identified in nearly all coding and noncoding regions of the plastome, although they were unequally distributed between and among exons, introns and noncoding

regions (Table 5.1, Figure 5.2). Highest average per site OV was found in noncoding regions, followed by protein-coding exons, introns, and finally RNA-coding exons (Table 5.1). With removal of *ycf*1or *ycf*1 and *ycf*2 positions, average per site OV for protein-coding exons fell below that of intronic regions (Table 5.1).

Phylogenetic Analysis of the Full Alignment

Our full alignment contained 42468 alignment patterns, and resulted in highly supported and almost completely congruent topologies in likelihood, Bayesian and parsimony analyses (Supplementary Figures 5.1 and 5.2). All major clades at the subgenus, sectional and subsectional levels as reported by Gernandt et al. (2005) were recovered with 95-100% bootstrap support. Across the topology, average maximum likelihood bootstrap support for 105 ingroup nodes was 89.7% (standard deviation = 18.5%). Only two minor topological conflicts were found between methods. In subsection *Australes*, *Pinus caribaea* was placed sister to a clade of *P. cubensis* and *P. occidentalis* with low support in Bayesian analysis (<0.6 posterior probability), while ML and parsimony analyses recovered *P. caribaea* sister to *P. palustris*, again with low support (≤50% bootstrap support). In section *Quinquefoliae*, parsimony analysis recovered *P. morrisonicola* in a weakly supported clade with *P. armandii* (55% bootstrap support), while both Bayesian and ML methods recovered these species in a grade with variable support (43% bootstrap support, 0.97 posterior probability). Section *Trifoliae* was recovered as subsection *Contortae* + (subsection *Australes* + subsection *Ponderosae*) with high support (100% bootstrap / 1.0 posterior probability) for the monophyly and position of subsection *Contortae*. Section *Quinquefoliae* was recovered as subsection *Strobus* + (*P. krempfii* + subsection *Gerardianae*) with weak to strong support (58-73% bootstrap / 1.0 posterior probability) for the position of *P. krempfii*. Section *Pinus* was recovered as subsection *Pinus* + (*P. merkusii*/*P. latteri* + subsection *Pinaster*) with weak to moderate for the position of *P. merkusii*/*P. latteri* (50-71% bootstrap / 0.52 posterior probability) but strong support for the monophyly of these two species (100% bootstrap, 1.0 posterior probability).

Impact of Variable Site Removal

Bootstrap support values showed clear trends throughout the FA partitions, with overall values consistently high (average value > 85%, median value ≥ 98%) until the most variable 8.3 kbp

had been removed (FA.133065) (Figure 5.3). Overall bootstrap values steadily decreased from this point until the most variable 18 kbp had been removed (FA ca. 123 kbp in size), at which point values levelled off at very low values (average value < 17%, median value < 10%). BSM values initially declined rapidly, but then rose again before decreasing rapidly a final time starting with the removal of the most variable 4.6 kbp (FA.136665) (Figure 5.4). After this point, BSM values remained consistently low. PM values experienced an initial rapid decline before levelling off after the removal of the most variable 2.2 kbp (FA ca. 139.1 kbp in size) (Figure 5.4). PM values remained constant and relatively low until increasing again beginning with the removal of the most variable 8.3 kbp (FA.133065). The lowest PM values occurred between removal of the most variable 7.4 and 8.2 kbp (FA ca. 133.9-133.1 kbp in size).

Monophyly of subsection *Contortae* was highly supported until removal of 15.3 kbp of the most variable sites (FA ca. 126 kbp in size), while support for the phylogenetic position of the *Contortae* decreased fairly steadily after removal of only 4.2 kbp (FA ca. 137.2 kbp in size) (Figure 5.5A). Section *Trifoliae* was recovered as subsection *Contortae* + (subsection *Australes* + subsection *Ponderosae*) by all FA partitions greater than 137 kbp in size; resolution based on FA partitions less than 137 kbp in size was variable, although placement of subsection *Contortae* as sister to or nested within subsection *Australes* was supported by several partitions between FA.136665 and FA.133065.

Bootstrap support for the phylogenetic position of *P. krempfii* was moderate (59-84%) until removal of the most variable 5.7 kbp (FA size 135.6 kbp), at which point bootstrap values steadily increased until peaking at 97-100% after removal of the most variable 6.3-7.8 kbp (FA size 133.6-135 kbp) (Figure 5.5B). FA phylogenetic partitions greater than 129.4 kbp in size recovered section *Quinquefoliae* as subsection *Strobus* + (*P. krempfii* + subsection *Gerardianae*); at FA partition sizes smaller than this phylogenetic position was variable.

The monophyly of *P. merkusii*/*P. latteri* was highly supported until removal of the most variable 18.2 kbp (FA size 123.2 kbp) (Figure 5.5C). Support for their resolution within section *Pinus*, however, was consistently moderate until removal of 7.2 kbp of the most variable sites (FA size 134.2). FA phylogenetic partitions prior to this point recovered the *P. merkusii*/*P. latteri* clade alternately sister to subsection *Pinaster* and subsection *Pinus*. After this point, bootstrap support rapidly increased to a peak of 96-100% between removals of 7.6-

9 kbp of the most variable sites (FA sizes 132.4-133.7 kbp), and all FA partitions in this range recovered section *Pinus* as subsection *Pinus* + (*P. merkusii*/*P. latteri* + subsection *Pinaster*).

Impact of Long-Branch Exclusion

When all alignment sites were included in analyses, long-branch exclusion strategies had essentially no impact on the topology or support of subsection *Contortae* or *Pinus krempfii*, while support increased moderately for a monophyletic (*Pinus merkusii*/*P. latteri* + subsection *Pinaster*) only with the most exclusive strategy (Table 5.2). When long-branch exclusion was used in combination with variable site removal (partition sizes FA.136665 and FA.133065), trends were reflective of variable site removal alone for partition size FA.136665 in subsection *Contortae* and *P. merkusii*/*P. latteri* (Table 5.2). In the remaining cases, trends were either non-existent (*P. krempfii* and *P. merkusii*/*P. latteri* exclusion strategies applied to FA.133065) or counter to patterns seen with variable site removal alone (*P. krempfii* exclusion strategies applied to FA.136665, subsection *Contortae* exclusion strategies applied to FA.133065) (Table 5.2).

Impact of Noise Reduction on Saturation

Correlation between paired corrected and uncorrected pairwise genetic distances was high for all strategies of variable site removal and outgroup exclusion (minimum $R^2 = 0.9997$). Based on the slopes of regression lines of corrected vs. uncorrected pairwise distances, saturation decreased similarly both with variable site removal and long-branch exclusion strategies (Table 5.3). The highest levels of saturation were observed with inclusion of all accessions, while the lowest values occurred with removal of the most variable 8.3 kbp of the alignment (FA.133065) and exclusion of at least the Pinaceae outgroups (Table 5.3).

DISCUSSION

As genome-scale datasets become increasingly common tools in evolutionary analyses, it is reasonable to expect challenges associated with highly variable or noisy data. Because of this, it is prudent to develop efficient strategies to identify and mitigate phylogenetic noise while simultaneously preserving sites and taxa carrying useful phylogenetic signal in order to most effectively capture information from large datasets. The benefit of developing such strategies has been demonstrated already, for example in placental mammals (Goremykin et al. 2010), early-diverging angiosperm lineages (Goremykin et al. 2009) and deep eukaryotic phylogeny

(Rodríguez-Ezpeleta et al. 2007). Our methodology is similar to previous efforts, but focused on two fundamental and complementary strategies, variable site removal and long-branch exclusion, and explored the dynamics of tree topology and support values to measure their impact on an infrageneric phylogenetic analysis. While the two strategies we employed were both utilized to counter the effect of phylogenetic noise, there are important contrasts between them. For example, the strict application of long-branch exclusion serves to minimize long-branch attraction artefacts, yet phylogenetic hypotheses may still be misled by evolutionary patterns at highly variable sites since all sites are still included in the analysis. In this case, removal of taxa could mask evolutionary patterns at some sites that otherwise might be more clearly interpreted (Hendy and Penny 1989, Zwickl and Hillis 2002), while the inclusion of fast-evolving sites may still mislead phylogenetic analyses (Townsend and Leuenberger 2011). On the other hand, removal of highly variable sites diminishes the impact of noise in an alignment and should increase the ability of applied models of sequence evolution to capture evolutionary patterns in phylogenetic analyses. The success of this strategy may be limited, however, as the inclusion of highly divergent taxa could still lead to long-branch artefacts when phylogenetic signal is minimal, and the broad application of variable site removal may diminish or erase phylogenetic signal in some clades (Kalersjo et al. 1999). It is therefore likely that utilizing a combination of these two strategies is prudent in many cases (Rodríguez-Ezpeleta et al. 2007), yet an overly conservative approach could still lead to the loss of essential phylogenetic signal. With our dataset and strategy, removal of variable sites appears to be a more effective tool in clarifying the evolutionary relationships of three historically problematic clades, and it seems reasonable to investigate conflicting or weakly supported phylogenetic resolution by removing the most variable 4.6 to 8.3 kb of alignment positions from our 142 kbp alignment. This range of variable site removal, corresponding to alignment partitions with lower BSM/PM values and high overall bootstrap support, is significant in two regards. First, as overall high levels of bootstrap support are maintained across this range of partitions (Figure 5.3), these sites clearly carry essential signal for the resolution of many relationships within the genus *Pinus*. Second, as variable sites are removed within this range of partitions, support for the putatively incorrect position of subsection *Contortae* diminishes substantially, while there is increasing resolution for the positions of *P. krempfii* and *P. merkusii* / *P. latteri*. Conversely, the long-branch exclusion strategies applied have little to no effect on the topology and support for these clades when applied to the full plastome

alignment, suggesting that variable site removal is more effective in mitigating the impact of phylogenetic noise in our data set.

The specific changes in position and topological support shown in our analyses are also noteworthy because they highlight the disparate resolutions previously supported by different analyses or different types of data. For example, the position of subsection *Contortae* is strongly supported (up to 100% bootstrap support) as sister to subsections *Ponderosae* and *Trifoliae* (Figure 5.1) based on previous reports using chloroplast sequence data or chloroplast restriction fragment analyses (Krupkin et al. 1996, Geada Lopez et al. 2002, Gernandt et al. 2005, Eckert and Hall 2006, Gernandt et al. 2008, Parks et al. 2009). Alternatively, other lines of evidence suggest this highly supported topology may be incorrect. For example, hybridization is possible between some members of subsections *Contortae* and *Australes*, but not between members of subsections *Contortae* and *Ponderosae* (Critchfield 1963, Saylor and Koenig 1967). Similarly, the relatively shallow fossil record of subsection *Contortae* (Miller Jr. 1992, McKown et al. 2002) suggests a more recent derivation within its section. In turn, two reports based on nrITS and four low-copy nuclear loci place the *Contortae* either nested within subsection *Australes* with moderately high support (77-82% bootstrap support in Liston et al. (2003)) or forming a polytomy with monophyletic subsections *Ponderosae* and *Australes* (Syring et al. 2005), respectively, while restriction fragment analysis including chloroplast, mitochondrial and nuclear DNA suggest a more derived position of subsection *Contortae* within section *Trifoliae* and some affinity to members of subsection *Australes* (Strauss and Doerksen 1990). The unique morphological characteristics of *Pinus krempfii* (most notably its flat, paired needles) have led to a wide range of phylogenetic resolution, including placement outside the genus *Pinus* (Chevalier 1944), in its own subgenus within *Pinus* (De Ferre 1953, Gaussen 1960, Little and Critchfield 1969), and within subgenus *Strobus*, section *Parrya* (Van der Berg 1973, Farjon 1984, Ickert-Bond 2001). At least two morphological treatments have recognized an affinity of *P. krempfii* to *P. gerardiana* and *P. bungeana* of subsection *Gerardianae* (Pilger 1926, Ickert-Bond 2001). Molecular evidence to date strongly support a position within or sister to section *Quinquefoliae* of subgenus *Strobus*, although a consistent and clear relationship of *P. krempfii* to subsections *Strobus* and *Gerardianae* of section *Quinquefoliae* has proven elusive (Figure 5.1). Some analyses based on chloroplast sequence data suggest an affinity to subsection *Gerardianae* (Wang et al. 1999, Wang et al. 2000), but support for this relationship is typically moderate to weak. Other reports based on chloroplast

or nuclear sequence data show poor resolution (Gernandt et al. 2005, Gernandt et al. 2008), place the species sister to section *Quinquefoliae* (Parks et al. 2009), or suggest inclusion within subsection *Strobus* (Liston et al. 1999, Liston et al. 2003). *Pinus merkusii* and *P. latteri* have demonstrated similarly ambiguous phylogenetic resolution relative to subsections *Pinus* and *Pinaster* of section *Pinus* (Figure 5.1), and again there is incongruence between molecular and morphological data. For example, Frankis (1993) placed *P. merkusii* within subsection *Pinaster* based on cone morphology, while most molecular analyses place *P. merkusii* as sister to subsection *Pinus* (Liston et al. 1999, Wang et al. 1999, Liston et al. 2003, Gernandt et al. 2005), albeit typically with low to moderate support. On the other hand, Wang et al. (1999, 2000) and Szmidt et al. (1996) demonstrated a clear genetic separation of *P. merkusii* from sampled Asian members of subsection *Pinus*, and suggest a divergence between these groups possibly in the early Tertiary, although this timeframe is not in accordance with the age of section *Pinus* based on molecular clock calibrations (Willyard et al. 2007, Gernandt et al. 2009).

While our results cannot be considered conclusive by themselves, they certainly add important perspectives to *Pinus* evolutionary history as well as the use of plastome-scale sequences in plant phylogenomic analyses. For the genus *Pinus* as a whole, our dataset apparently represents the maximal resolution to be gained from the plastome, although various permutations of chloroplast loci may still prove useful at different levels of phylogenetic inquiry (for example see Gernandt et al. 2009). From this point, the next target of phylogenetic interrogation will likely be larger unique portions of the nuclear genome, particularly as increases in sequence output continue to outpace increases in read length for next-generation sequencers (Alkan et al. 2011) and progress is made on the sequencing and assembly of a representative pine nuclear genome (Neale and Kremer 2011). For the three specific clades investigated in this study, the similarities in response to removal of noise from the full plastome alignment were intriguing and clear insight was gained into their evolutionary histories and relationships. In each case, decreasing the impact of phylogenetic noise by removing highly variable sites resulted in phylogenetic resolution more reflective of results based on nuclear and/or morphological data. At the same time, the impact of long-branch exclusion was less pronounced, suggesting that long-branch attraction artefacts are not prevalent at this level of the *Pinus* phylogeny. The congruent results between model-based and parsimony methods for these clades also lend support to this conclusion, as methodological

incongruence is another indication of possible long branch attraction artefacts (Bergsten 2005). This result is somewhat counter-intuitive, as all three lineages investigated have relatively long branches in chloroplast-based phylogenetic reconstruction (Supplementary Figure 5.1B) (Parks et al. 2009). It is possible that these long branches are not all reflective of the same biological processes. For subsection *Contortae*, chloroplast-based support for an early divergence in section *Trifoliae* is clearly inflated by the phylogenetic noise of highly variable sites. In this case, the pronounced effect of variable site removal combined with the relatively long branch leading to subsection *Contortae* may instead be indicative of elevated rates of evolution in this lineage, and a position sister to or within subsection *Australes* could be the final resolution of this challenging group. The long branches of *P. krempfii* and the *P. merkusii*/*P. latteri* clade, on the other hand, likely are due to relatively long periods of divergence from their sister lineages. In these cases, however, it appears that removal of accumulated noise unmasks the limited underlying signal more definitively supporting their resolution - *P. krempfii* as sister to subsection *Gerardianae* of section *Quinquefoliae*, and *P. merkusii*/*P. latteri* as sister to subsection *Pinaster* of section *Pinus*.

As demonstrated in the current study, the promise of phylogenomics is still very much palpable and (to paraphrase Mark Twain) reports of its 'demise' (Delsuc et al. 2005, Jeffroy et al. 2006) are greatly exaggerated. Still, it is equally premature in many cases to confirm phylogenetic results based on genome-scale datasets without investigating first for the presence of misleading signal (Delsuc et al. 2005, Philippe et al. 2005, Jeffroy et al. 2006, Philippe et al. 2011). This is particularly important when trying to reconcile poorly supported topologies or conflicting phylogenetic results based on different sources or types of data (Philippe et al. 2011). The present analysis and similar efforts (for example Goremykin et al. 2010) also demonstrate not only the power of large (but well-managed) datasets to increase phylogenetic resolution, but the risk of relying on single sources of data, as inconsistencies between organellar- and nuclear-based analyses can remain even with greatly increased sampling. Fortunately, sequencing capacity and read length of next-generation platforms continue to increase (Mardis 2008, Shendure and Ji 2008, Metzker 2010, Mardis 2011, Schweiger et al. 2011), and combined with increasingly effective methods of genome interrogation (Mamanova et al. 2010, Etter et al. 2011, Seeb et al. 2011, Cronn et al. Manuscript in preparation) will make it easier to capture useful sequence data from what are currently less tractable genomes (such as plant nuclear and mitochondrial genomes). However,

the development of analytical strategies to deal with noise present in large datasets will remain essential, as phylogenetic signal clearly is not always sufficient to overcome noise, even at genomic scales.

Figure 5.1. Phylogenetic hypothesis for *Pinus* showing alternate placements (indicated by dashed lines) of subsections *Contortae* and *Krempfianae*, as well as the clade consisting of *Pinus merkusii* and *P. latteri*. The most common plastid resolution of these groups is indicated in bold. Tree topology and relative branch lengths reflective of present data, Gernandt et al. (2005) and Parks et al. (2009).

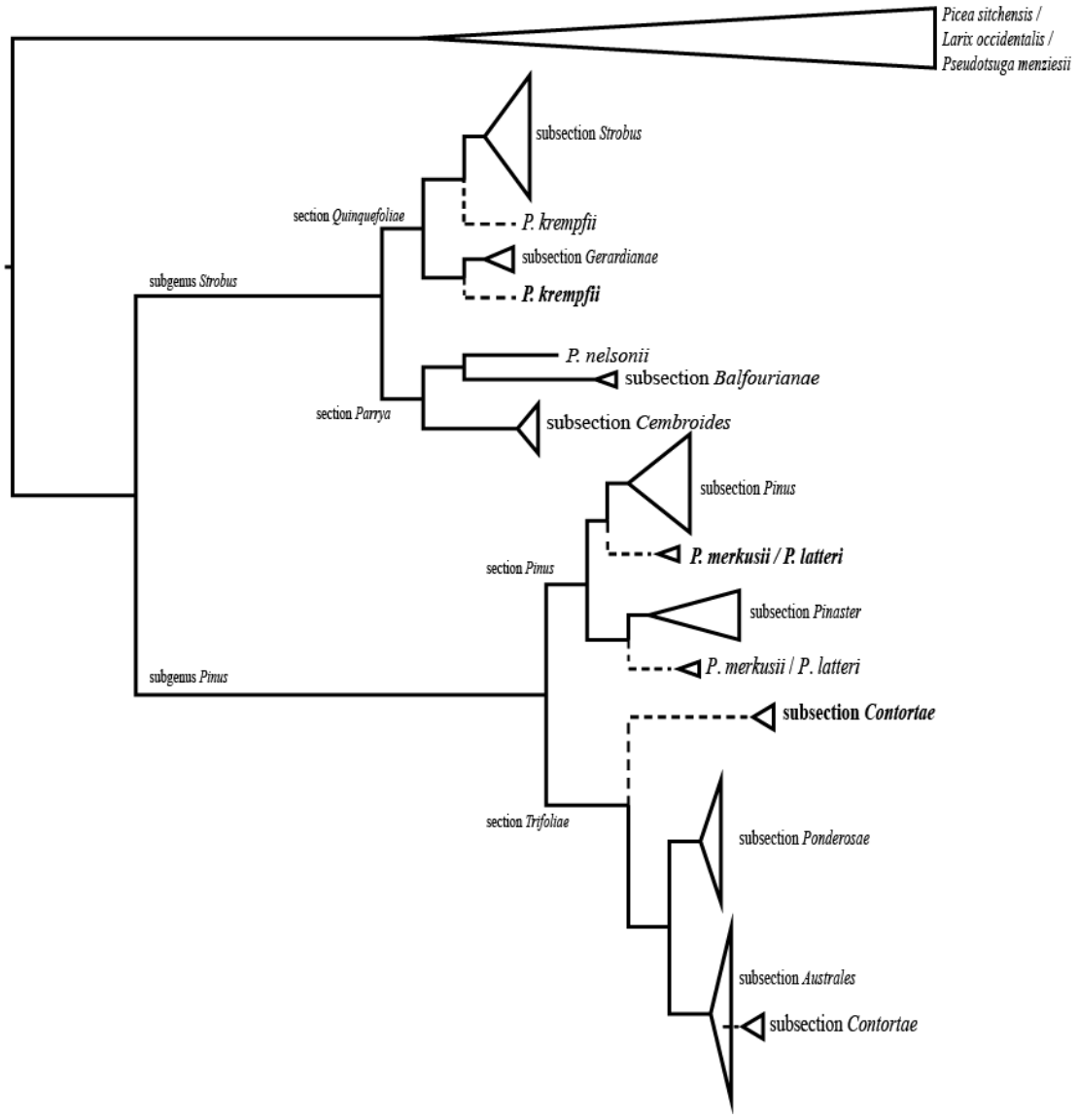


Figure 5.2. Distribution of OV for variable plastome alignment positions. Outer circle represents plastome alignment as follows: green – noncoding regions; blue – protein-coding exons; yellow – rRNA; orange – tRNA. *ycf*1 and *ycf*2 are highlighted in light blue. Inner circle represents relative OV values for sites in alignment with OV>0 as follows: red – most variable 4.6 kbp; yellow – most variable sites from 4.3 to 8.6 kbp; green – remaining sites with OV>0.

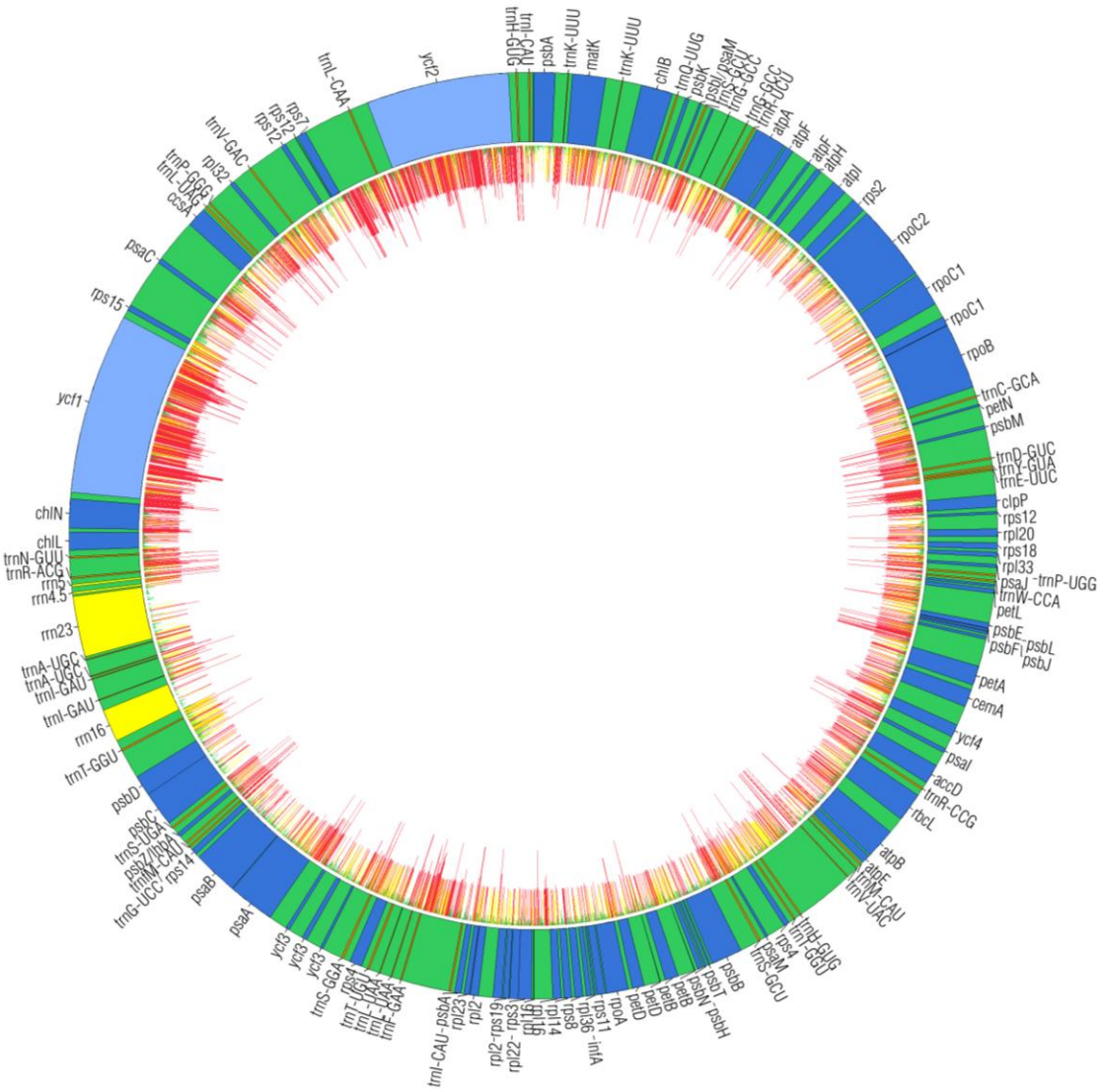


Figure 5.3. Trends in bootstrap support values for likelihood analyses of FA partitions. A) Average (triangles) and median (circles) bootstrap values for all nodes; B) Distributions of bootstrap support values for all nodes.

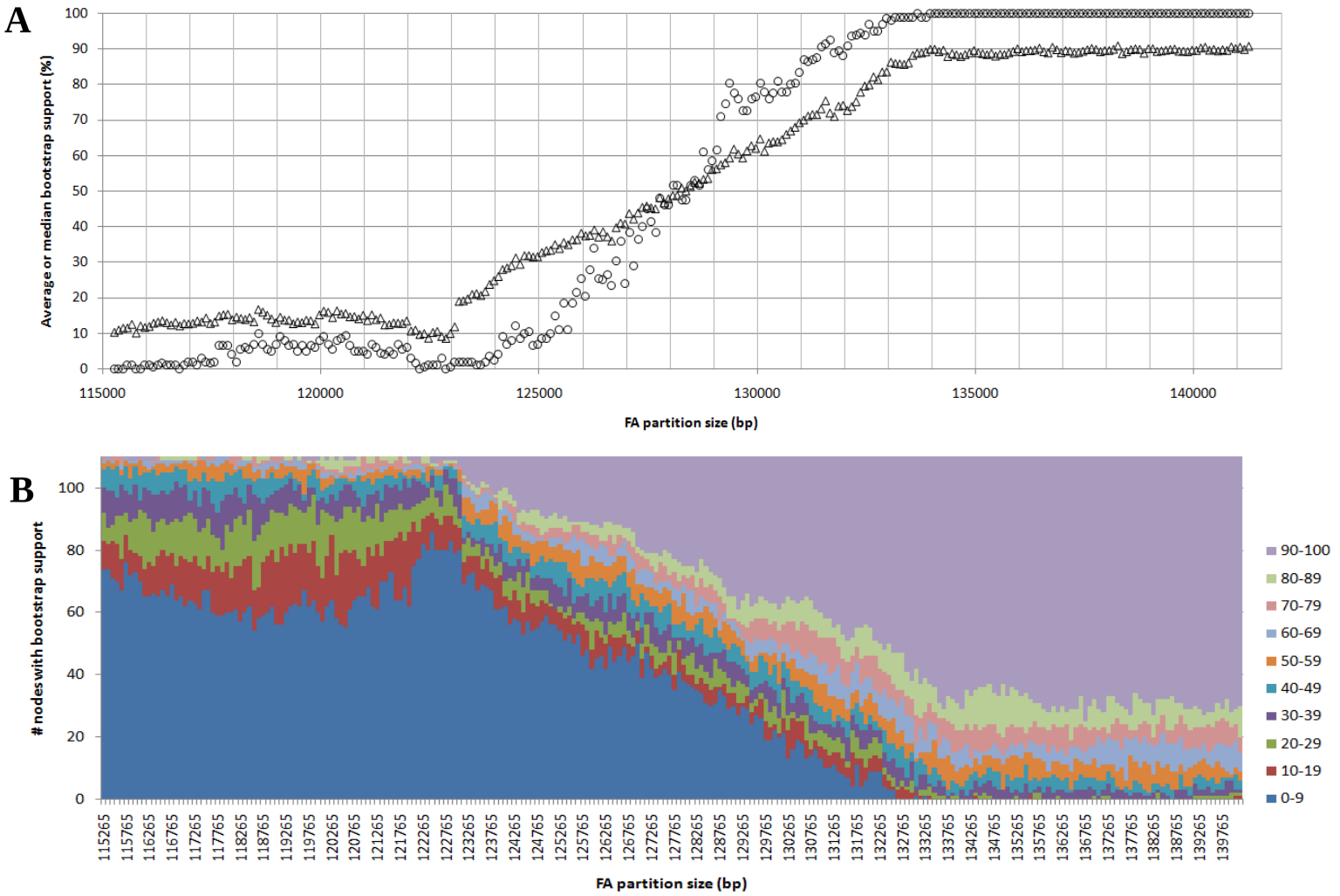


Figure 5.4. Distribution of BSM (triangles) and PM (circles) values for tests of topological congruence between FA and corresponding VS data partitions. Filled data points correspond to FA partitions sizes falling between final decrease of BSM values and start of decreases in overall bootstrap support values. PM values shown are 0.1× actual value in order to fit on same scale with BSM values.

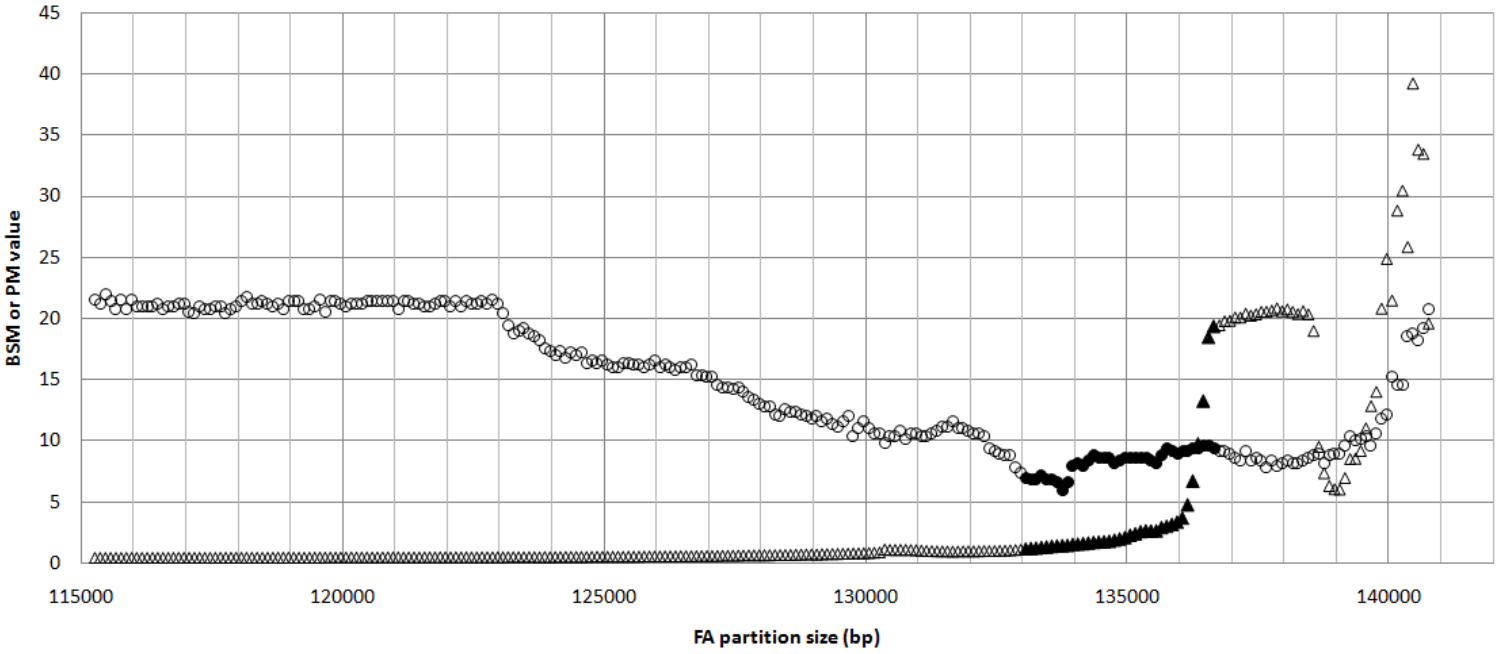


Figure 5.5. Distribution of bootstrap support values for three clades in genus *Pinus*. A) Subsection *Contortae*, B) subsection *Krempfianae* and C) *Pinus merkusii* / *P. latteri*. Circles represent bootstrap support values for monophyly of clade while triangles represent support for phylogenetic position of monophyletic clade. Filled data points correspond to FA partition sizes falling between final decrease of BSM values and start of decrease in overall bootstrap support values.

Figure 5.5.

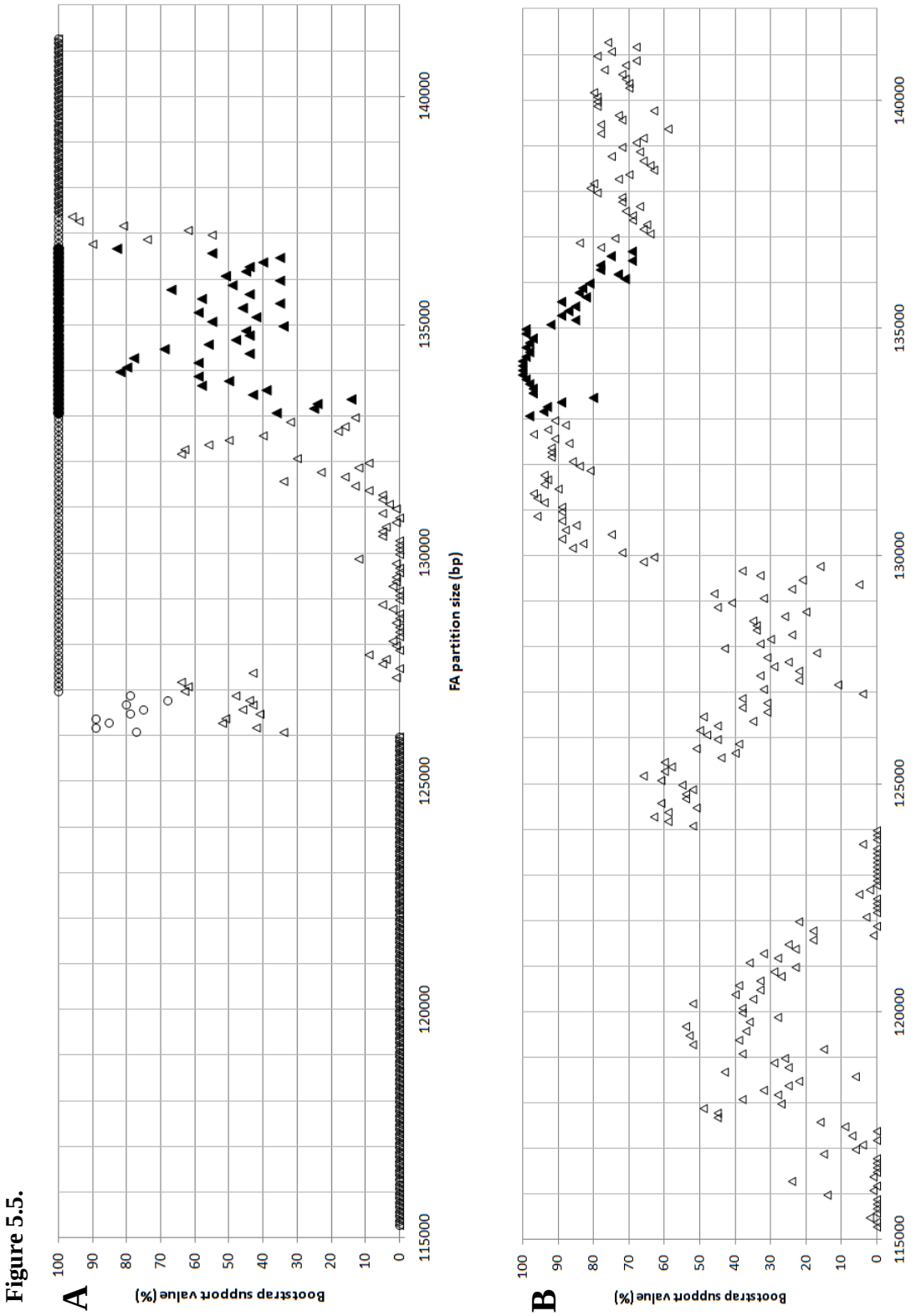

A
Bootstrap support value (%)
FA partition size (bp)
B
Bootstrap support value (%)
FA partition size (bp)

Figure 5.5.

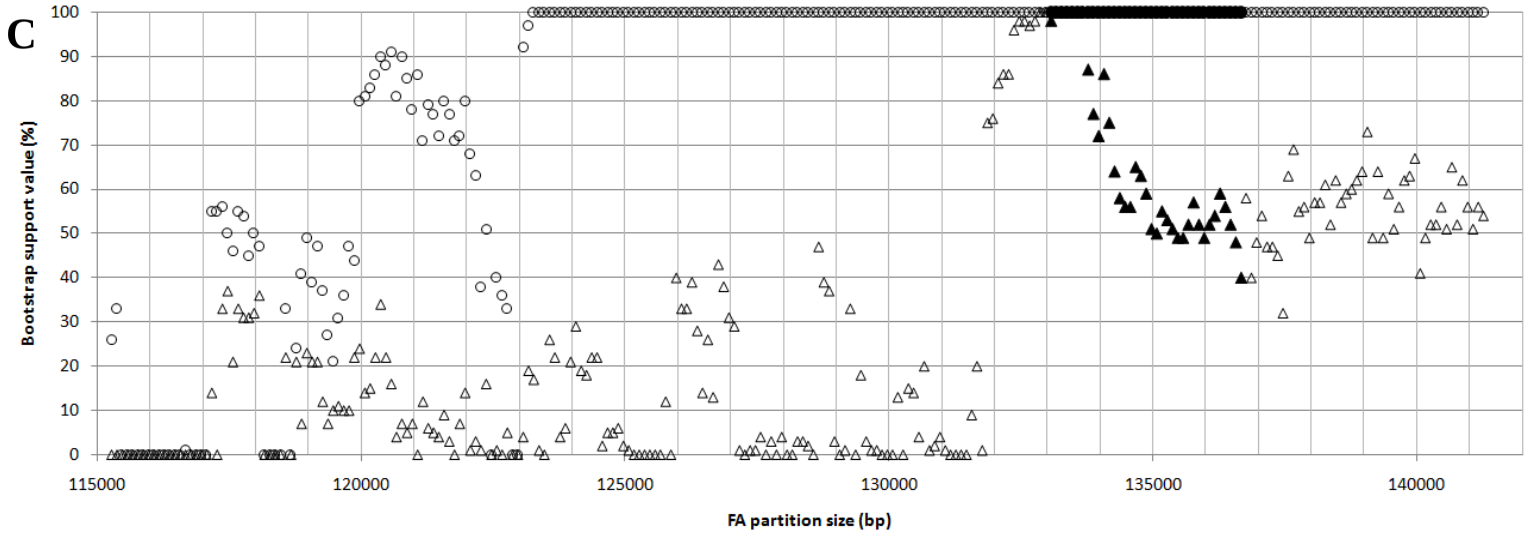

C
Bootstrap support value (%)
100
90
80
70
60
50
40
30
20
10
0
115000
120000
125000
130000
135000
140000
FA partition size (bp)

Table 5.1. Average per site OV values for protein-coding exons, introns, rRNA and tRNA genes, and noncoding regions for full plastome alignment of 113 *Pinus* and Pinaceae species. Standard deviations are given in parentheses.

	Noncoding regions	Protein-coding exons	Introns	tRNA	rRNA
average OV	0.04546 (0.12833)	0.03153 (0.11227)	0.02110 (0.08880)	0.00443 (0.03725)	0.00462 (0.04255)
average OV without *ycf*1		0.01907 (0.08184)			
without *ycf*1 or *ycf*2		0.01478 (0.06997)			

Table 5.2. Impact of long-branch exclusion on full alignment, FA.136665 and FA.133065 data partitions. For each combination, supported topology is given with maximum likelihood bootstrap support underneath in parentheses. Subsection and species abbreviations are as follows: C=*Contortae*, A=*Australes*, P=*Ponderosae*, K=*P. krempfii*, G=*Gerardianae*, S=*Strobus*, M/L=*P. merkusii*/*P. latteri*, Pinast.=*Pinaster*. Single outgroups used in most exclusive groups are described Methods and Materials.

	subsection *Contortae*				subsection *Krempfianae* (*Pinus krempfii*)				*Pinus merkusii* / *P. latteri*			
	all	no non-*Pinus* outgroups	subgenus only	section + single outgroup	all	no non-*Pinus* outgroups	subgenus only	section + single outgroup	all	no non-*Pinus* outgroups	subgenus only	section + single outgroup
Full alignment	C+(A+P) (100)	C+(A+P) (100)	C+(A+P) (100)	C+(A+P) (100)	(K+G)+S (79)	(K+G)+S (61)	(K+G)+S (79)	(K+G)+S (79)	(M/L+Pinast.) + Pinus (53)	(M/L+Pinast.) + Pinus (53)	(M/L+Pinast.) + Pinus (54)	(M/L+Pinast.) + Pinus (78)
FA.136665	C+(A+P) (100)	C+(A+P) (100)	C+(A+P) (100)	C+(A+P) (84)	(K+G)+S (71)	(K+G)+S (70)	(K+G)+S (71)	(K+G)+S (54)	(M/L+Pinast.) + Pinus (43)	(M/L+Pinast.) + Pinus (46)	(M/L+Pinast.) + Pinus (47)	(M/L+Pinast.) + Pinus (68)
FA.133065	P+(C+A) (37/42)	P+(C+A) (31/36)	P+(C+A) (35/42)	C+(A+P) (84)	(K+G)+S (97)	(K+G)+S (97)	(K+G)+S (97)	(K+G)+S (98)	(M/L+Pinast.) + Pinus (99)	(M/L+Pinast.) + Pinus (100)	(M/L+Pinast.) + Pinus (100)	(M/L+Pinast.) + Pinus (99)

Table 5.3. Slopes of regression lines for plots of corrected versus uncorrected pairwise distances. Slopes with decreased values below 1.0 represent increased levels of saturation in the alignment tested. Correlation was high in each test, with R^2 values ≥ 0.9997.

	All accessions	no non-*Pinus* outgroups	subgenus *Pinus*	subgenus *Strobus*	section *Pinus* + *P. ponderosa*	section *Trifoliae* + *P. thunbergii*	section *Quinquefoliae* + *P. monophylla*
Full alignment	0.9574	0.9682	0.9898	0.9893	0.9925	0.9928	0.9905
FA.136665	0.9604	0.9851	0.9935	0.9903	0.9941	0.9948	0.9916
FA.133065	0.9655	0.9954	0.9959	0.995	0.9955	0.9961	0.9952

Literature cited.

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. Current Opinion in Plant Biology, 8:135-141.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. Nat Rev Genet, 12:363-376.

Bergsten J. 2005. A review of long-branch attraction. Cladistics, 21:163-193.

Birky CW, Jr. 1978. Transmission genetics of mitochondria and chloroplasts. Annu. Rev. Genet., 12:471-512.

Brinkmann H, van der Giezen M, Zhou Y, de Raucourt GP, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol., 54:743-757.

Cai Z, Penaflor C, Kuehl JV, Leebens-Mack J, Carlson JE, dePamphilis CW, Boore JL, Jansen RK. 2006. Complete plastid genome sequences of *Drimys*, *Liriodendron*, and *Piper*: implications for the phylogenetic relationships of Magnoliids. BMC Evol. Biol., 6:77.

Chevalier A. 1944. Notes sur les coniferes de l'Indochine. Revue de Botanique Appliquee et d'Agriculture Tropicale, 24:7-34.

Critchfield WB. 1963. Hybridization of the southern pines in California. Southern Forest Tree Improvement Committee Publications, 22:40-48.

Critchfield WB. 1966. Crossability and relationships of the closed-cone pines. Silvae Genetica, 16:89-97.

Critchfield WB. 1975. Interspecific hybridization in *Pinus*: a summary review. In: Fowler DP, Yeatman CY editors. Symposium on Interspecific and Interprovenance Hybridization in Forest Trees, p. 99-105.

Critchfield WB. 1986. Hybridization and classification of the white pines (*Pinus* section *Strobus*). Taxon, 35:647-656.

Cronn R, Knaus B, Liston A, Maughan J, Parks M, Syring J, Udall J. Manuscript in preparation. Simplifying the complex: targeted sequencing strategies for population, phylogenetic, and genomic studies in plants.

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res., 36:e122.

De Ferre Y. 1953. Division du genre *Pinus* en quatre sous-genres. Academie des Sciences Compte Rendu, 236:226-228.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet, 2:762-768.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet, 6:361-375.

Dowton M, Austin AD. 2002. Increased congruence does not necessarily indicate increased phylogenetic accuracy - the behavior of the incongruence length difference test in mixed-model analyses. Syst. Biol., 51:19-31.

Eckert AJ, Hall BD. 2006. Phylogeny, historical biogeography, and patterns of diversification for *Pinus* (Pinaceae): phylogenetic tests of fossil-based hypotheses. Mol. Phylogenet. Evol., 40:166-182.

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA. 2011. Local *de novo* assembly of RAD paired-end contigs using short sequencing reads. PLoS ONE, 6:e18561.

Farjon A. 1984. Pines: drawings and descriptions of the genus. Leiden, W. Backhuys.

Farris JS, Källersjö M, Kluge AG, Bult C. 1995. Constructing a significance test for incongruence. Syst. Biol., 44:570-572.

Fauron C, Allen J, Clifton S, Newton K. 2004. Plant mitochondrial genomes. In: Daniell H, Chase C editors. Molecular Biology and Biotechnology of Plant Organelles, Springer Netherlands, p. 151-177.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology, 27:401-410.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington.

Fitch WM. 1979. Cautionary remarks on using gene expression events in parsimony procedures. Systematic Zoology, 28:375-379.

Fitch WM. 1984. Cladistic and other methods: problems, pitfalls, and potentials. In: Duncan T, Stuessy TF editors. Cladistics: Perspectives on the Reconstruction of Evolutionary History, Columbia University Press, p. 221-252.

Frankis MP. 1993. Morphology and affinities of *Pinus brutia*. International Symposium on *Pinus brutia* Ten. Ankara, Ministry of Forestry, p. 11-18.

Gaussen H. 1960. Les gymnospermes actuelles et fossiles. Fassicule VI. Les Coniferales. Chapter 11. Generalites, Genre *Pinus*. *Travaux du Toulous Universite Laboratoire Forestier*, p. 1-272.

Geada Lopez G, Kamiya K, Harada K. 2002. Phylogenetic relationships of diploxylon pines (subgenus *Pinus*) based on plastid sequence data. Int. J. Plant Sci., 163:737-747.

Gernandt DS, Hernández-León S, Salgado-Hernández E, Rosa JAPdl. 2009. Phylogenetic relationships of *Pinus* subsection *Ponderosae* inferred from rapidly evolving cpDNA regions. Syst. Bot., 34:481-491.

Gernandt DS, Lopez G, Garcia SO, Liston A. 2005. Phylogeny and classification of *Pinus*. Taxon, 54:29-42.

Gernandt DS, Magallon S, Geada Lopez G, Zeron Flores O, Willyard A, Liston A. 2008. Use of simultaneous analyses to guide fossil-based calibrations of Pinaceae phylogeny. Int. J. Plant Sci., 169:1086-1099.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotech., 27:182-189.

Goremykin V, Nikiforova S, Bininda-Emonds O. 2010. Automated removal of noisy data in phylogenomic analyses. J. Mol. Evol., 71:319-331.

Goremykin V, Viola R, Hellwig F. 2009. Removal of noisy characters from chloroplast genome-scale data suggests revision of phylogenetic placements of *Amborella* and *Ceratophyllum*. J. Mol. Evol., 68:197-204.

Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH. 2005. Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Mol. Biol. Evol., 22:1813-1822.

Grotkopp E, Rejmánek M, Sanderson MJ, Rost TL, Soltis P. 2004. Evolution of genome size in pines (*Pinus*) and its life-history correlates: supertree analyses. Evolution, 58:1705-1729.

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

Harris RS. 2007. Improved pairwise alignment of genomic DNA. Computer Science, *Dissertation*, The Pennsylvania State University.

Hendy MD, Penny D. 1989. A framework for the quantitative study of evolutionary trees. Syst. Biol., 38:297-309.

Hillis DM, Huelsenbeck JP. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. J. Hered., 83:189-195.

Hipp AL, Hall JC, Sytsma KJ. 2004. Congruence versus phylogenetic accuracy: revisiting the incongruence length difference test. Syst. Biol., 53:81-89.

Huelsenbeck JP. 1991. Tree-length distribution skewness: an indicator of phylogenetic information. Syst. Biol., 40:257-270.

Ickert-Bond S. 2001. Reexamination of wood anatomical features in *Pinus krempfii* (Pinaceae). IAWA Journal, 22:355-365.

Illumina. 2007. Protocol for whole genome sequencing using Solexa technology. Biotechniques Protocol Guide:29.

Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack J, Muller KF, Guisinger-Bellian M, Haberle RC, Hansen AK. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA, 104:19369.

Jansen RK, Kaittanis C, Saski C, Lee SB, Tomkins J, Alverson AJ, Daniell H. 2006. Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol. Biol., 6:32.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet., 22:225-231.

Kalersjo M, Albert VA, Farris JS. 1999. Homoplasy increases phylogenetic structure. Cladistics, 15:91-93.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res., 33:511-518.

Kent WJ. 2002. BLAT - the BLAST-like alignment tool. Genome Res., 4:656-664.

Klopfstein S, Kropf C, Quicke DLJ. 2010. An evaluation of phylogenetic informativeness profiles and the molecular phylogeny of Diplazontinae (Hymenoptera, Ichneumonidae). Syst. Biol., 59:226-241.

Klymiuk AA, Stockey RA, Rothwell GW. 2011. The first organismal concept for an extinct species of Pinaceae. Int. J. Plant Sci., 172:294-313.

Krupkin AB, Liston A, Strauss SH. 1996. Phylogenetic analysis of the hard pines (*Pinus* subgenus *Pinus*, Pinaceae) from chloroplast DNA restriction site analysis. Am. J. Bot., 83:489-498.

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol., 56:17-24.

Kubo T, Newton KJ. 2008. Angiosperm mitochondrial genomes and mutations. Mitochondrion, 8:5-14.

Kuhner MK, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol., 11:459-468.

Kurtz S, Phyllippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol., 5:R12.

Leebens-Mack J, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, dePamphilis CW. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol. Biol. Evol., 22:1948-1963.

Lin C-P, Huang J-P, Wu C-S, Hsu C-Y, Chaw S-M. 2010. Comparative chloroplast genomics reveals the evolution of Pinaceae genera and subfamilies. Genome Biology and Evolution, 2:504-517.

Liston A, Gernandt DS, Vining TF, Campbell CS, Pinero D. 2003. Molecular phylogeny of Pinaceae and *Pinus*. Acta Hort.:107-114.

Liston A, Robinson WA, Piñero D, Alvarez-Buylla ER. 1999. Phylogenetics of *Pinus* (Pinaceae) based on nuclear ribosomal DNA internal transcribed spacer region sequences. Mol. Phylogenet. Evol., 11:95-109.

Little EL, Critchfield WB. 1969. Subdivision of the genus *Pinus* (pines). In: U.S. Department of Agriculture FS editor. Washington, D.C., p. 1-51.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. 2010. Target-enrichment strategies for next-generation sequencing. Nat Meth, 7:111-118.

Mardis ER. 2008. Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics, 9:387-402.

Mardis ER. 2011. A decade's perspective on DNA sequencing technology. Nature, 470:198-203.

McKown AD, Stockey RA, Scheger CE. 2002. A new species of *Pinus* subgenus *Pinus* subsection *Contortae* from Pliocene sediments of Ch'ijee's Bluff, Yukon Territory, Canada. Int. J. Plant Sci., 163:687-697.

Metzker ML. 2010. Sequencing technologies - the next generation. Nat Rev Genet, 11:31-46.

Millar CI. 1998. Early evolution of pines. In: Richardson DM editor. Ecology and biogeography of *Pinus*. Cambridge, Cambridge University Press, p. 69–91.

Miller Jr. CN. 1992. Preserved cones of *Pinus* from the Neogene of Idaho and Oregon. Int. J. Plant Sci., 153:147-154.

Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci USA, 104:19363.

Neale DB, Kremer A. 2011. Forest tree genomics: growing resources and applications. Nat Rev Genet, 12:111-122.

Oda K, Yamato K, Ohta E, Nakamura Y, Takemura M, Nozato N, Akashi K, Kanegae T, Ogura Y, Kohchi T, *et al.* 1992. Gene organization deduced from the complete sequence of liverwort *Marchantia polymorpha* mitochondrial DNA : a primitive form of plant mitochondrial genome. Journal of Molecular Biology, 223:1-7.

Ortiz Garcia S. 1999. Evolucion y fiogenia en pinos y sus hongos endofitos: aspectos sistematicos de la coespeciacion. Instituto de Ecologia. Mexico City, Universidad Nacional Autonoma de Mexico.

Palmé AE, Pyhäjärvi T, Wachowiak W, Savolainen O. 2009. Selection on nuclear genes in a *Pinus* phylogeny. Mol. Biol. Evol., 26:893-905.

Palmer JD. 1990. Contrasting modes and tempos of genome evolution in land plant organelles. Trends Genet., 6:115-120.

Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu Y-L, Song K. 2000. Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proc. Acad. Nat. Sci. Phila., 97:6960-6966.

Pan X, Urban AE, Palejev D, Schulz V, Grubert F, Hu Y, Snyder M, Weissman SM. 2008. A procedure for highly specific, sensitive, and unbiased whole-genome amplification. Proc. Acad. Nat. Sci. Phila., 105:15499-15504.

Parks M, Cronn R, Liston A. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biology, 7:84.

Parks M, Liston A, Cronn R. 2011. Newly developed primers for complete *ycf*1 amplification in *Pinus* (Pinaceae) chloroplasts with possible family-wide utility. Am. J. Bot.:in press.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol, 9:e1000602.

Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. Ecology, Evolution and Systematics, 36:541-562.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol., 21:1455-1458.

Pilger R. 1926. Pinus. In: Engler A, Prantl K editors. Die Naturliche Pflanzenfamilien. Liepzig, Wilhelm Engelmann, p. 331-342.

Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol., 51:664-671.

Ratan A. 2009. Assembly algorithms for next-generation sequence data. *Dissertation*, The Pennsylvania State University.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Mathematical Biosciences, 53:131-147.

Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol., 56:389-399.

Rokas A, Carroll SB. 2006. Bushes in the tree of life. PLoS Biol, 4:e352.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, 19:1572-1574.

Saylor LC, Koenig RL. 1967. The slash x sand pine hybrid. Silvae Genetica, 16:134-138.

Schweiger M, Kerick M, Timmermann B, Isau M. 2011. The power of NGS technologies to delineate the genome organization in cancer: from mutations to structural variations and epigenetic alterations. Cancer and Metastasis Reviews, 30:199-210.

Seeb JE, Carvalho G, Hauser L, Naish K, Roberts S, Seeb LW. 2011. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Molecular Ecology Resources, 11:1-8.

Shaw J, Lickey EB, Schilling EE, Small RL. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. Am. J. Bot., 94:275.

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

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nat Biotech, 26:1135-1145.

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, *et al.* 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. The EMBO Journal, 5:2043-2049.

Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu Y-L, Chase MW, Farris JS, Stefanovic S, Rice DW, Palmer JD, *et al.* 2004. Genome-scale data, Angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. Trends Plant Sci., 9:477-483.

Stamatakis A. 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol., 57:758-771.

Stefanovic S, Rice D, Palmer J. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? BMC Evol. Biol., 4:35.

Straub SC, Fishbein M, Livshultz T, Foster Z, Parks M, Weitemier K, Cronn RC, Liston A. 2011. Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. BMC Genomics, 12:211.

Strauss SH, Doerksen AH. 1990. Restriction fragment analysis of pine phylogeny. Evolution, 44:1081-1096.

Strimmer K, von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Acad. Nat. Sci. Phila., 94:6815-6819.

Swofford DL. 2000. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland, Massachusetts, Sinauer Associates.

Syring J, Willyard A, Cronn R, Liston A. 2005. Evolutionary relationships among *Pinus* (Pinaceae) subsections inferred from multiple low-copy nuclear loci. Am. J. Bot., 92:2086-2100.

Szmidt A, Wang XR, Changtragoon S. 1996. Contrasting patterns of genetic diversity in two tropical pines: *Pinus kesiya* (Royle ex Gordon) and *P. merkusii* (Jungh et De Vriese). TAG Theoretical and Applied Genetics, 92:436-441.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol., 24:1596-1599.

Townsend JP. 2007. Profiling phylogenetic informativeness. Syst. Biol., 56:222-231.

Townsend JP, Leuenberger C. 2011. Taxon sampling and the optimal rates of evolution for phylogenetic inference. Syst. Biol.

Van der Berg J. 1973. Holzer der niederrheinischen braunkohlenformation 2. Holzer der braunkohlengruben "Maria Theresia" zu Herzogenrath, "Zukunft West" zu Eschweiler und "Victor" Zulpich mitte zu Zulpich. Nebst einer systematisch-anatomischen bearbeitung der gattung *Pinus* L. Review of Palaeobotany and Palynology, 15:73-275.

Wang XR, Szmidt AE, Nguyên HN. 2000. The phylogenetic position of the endemic flat-needle pine *Pinus krempfii* (Pinaceae) from Vietnam, based on PCR-RFLP analysis of chloroplast DNA. Plant Syst. Evol., 220:21-36.

Wang XR, Tsumura Y, Yoshimaru H, Nagasaka K, Szmidt AE. 1999. Phylogenetic relationships of Eurasian pines (*Pinus*, Pinaceae) based on chloroplast *rbc*L, *mat*K, *rpl*20-*rps*18 spacer, and *trn*V intron sequences. Am. J. Bot., 86:1742-1753.

Wendel JF. 2000. Genome evolution in polyploids. Plant Molecular Biology, 42:225-249.

Willyard A, Syring J, Gernandt DS, Liston A, Cronn R. 2007. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. Mol. Biol. Evol., 24:90-101.

Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Acad. Nat. Sci. Phila., 84:9054-9058.

Zhang ZY, Li DZ. 2004. Molecular phylogeny of section *Parrya* of *Pinus* (Pinaceae) based on chloroplast *mat*K gene sequence data. Acta Bot. Sin., 46:171-179.

Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol., 51:588-598.

CONCLUDING REMARKS

As we move into a new era of genomic-level analyses across the biological sciences, there is ample reason for both excitement and caution. Rapid increases in sequence throughput due to the technological advances of next-generation or massively parallel sequencing, and perhaps '3rd generation' sequencers, are allowing genomic exploration at an unprecedented level (Rusk 2009, Mardis 2011). Simultaneously, persistent challenges, ranging from accurately assembling sequence reads into genomic level contigs to simply storing and manipulating the massive amounts of data inherent to these technologies, will continue to require perspicacious and clever solutions (GBET 2011, Mardis 2011). As a matter of perspective, the recent sequencing of the giant panda genome, completed exclusively with Illumina sequencing, resulted in over 176 giga-base pairs (Gbp) of useful sequence information, enough to theoretically cover the ca. 2.4 Gbp genome 73 times (Li et al. 2010). The short read length (average of 52 bp, using paired-end sequencing), however, limited the assembly such that the N50 contig length was just under 40,000 bp (the N50 length represents the point at which 50% of all contig bases reside in contigs this length or greater). By comparison, the recent sequencing of the woodland strawberry genome, which utilized three next-generation sequencing platforms (Illumina, Roche/454 and SOLiD), averaged 39× coverage of the ca. 240 mega-base pair (Mbp) genome (Shulaev et al. 2011). In this case, the longer average read length due to Roche/454 sequence reads (ca. 365 bp) and longer Illumina reads (76 bp, using paired-end sequencing) helped produce a higher N50 of 1.3 Mbp. Nonetheless, although overall genomic coverage was sufficient for the assembly of the great majority of the nuclear genomes in both of these cases, the resulting assemblies were still fragmented into thousands of discrete contigs.

The thesis work presented herein, as well as related work (Cronn et al. 2008, Whittall et al. 2010), is similarly reflective of the benefits and challenges of new sequencing technologies, as manifested in the exploration of *Pinus* chloroplast genomes and their utility in phylogenetic analysis. In addition, it is important to recognize the implications of our work in *Pinus* for similar efforts in plant systematics in general. Our initial efforts, coinciding with the first years of commercially available massively parallel sequencers, demonstrated that large amounts of phylogenetically useful data could be generated rapidly and affordably, and resulted in substantially increased phylogenetic resolution for a subset of *Pinus* species (Cronn et al.

2008, Parks et al. 2010). In contrast, our meta-analysis of contemporary chloroplast-based infrageneric phylogenetic studies revealed that researchers using traditional Sanger sequence technology sample around 32 ingroup species and utilize less than 2600 bp of aligned chloroplast sequence (Parks et al. 2009), while resolving less than 23% of ingroup nodes with high support. Further, a significant and positive correlation between the amount of sequence applied and the proportion of nodes resolving with high (≥95%) bootstrap support suggest that the application of full plastome sequences to phylogenetic analyses in most plant genera should allow gains in resolution similar to what we demonstrated in *Pinus*. Simultaneously, limitations in short-read sequencing technologies were also demonstrated. For example, increasing genetic divergence resulted in less effective plastome assemblies in subgenus *Pinus* and Pinaceae outgroups (Parks et al. 2010), and repetitive regions in the chloroplast genome proved difficult or impossible to confidently resolve with short sequence reads (Cronn et al. 2008). We also relied on Sanger sequencing to accurately assemble the highly divergent locus *ycf*1 for our subsectional references, further demonstrating the difficulties of short-read assembly in repetitive and highly divergent regions. Finally, the analysis of phylogenetic noise in our final data matrix suggests that several thousand sites spuriously detract from the support of valid phylogenetic positioning in two taxa, while strongly supporting a putatively incorrect topology in a third taxon. It is reasonable to assume similar instances will occur in plastome-scale analyses in other taxa, and clearly the impact of noise removal should be explored. This is particularly the case when results based on chloroplast sequence conflict with those based on other types of data, such as nuclear sequence or morphological data.

Beyond strictly answering phylogenetic questions, our work has also benefitted from and contributed to the development of laboratory techniques and bioinformatic strategies that allow more efficient production and manipulation of large sequence datasets. For example, Richard Cronn developed a strategy based on the work of Gnirke et al. (2009) to effectively enrich genomic library preparations for chloroplast DNA using *Pinus* plastome sequences as hybridization probes. This technique was essential to the sequencing of the majority of our plastome accessions, allowing for the preparation and sequencing of over 80 nearly complete (ca. 120,000 bp) *Pinus* and Pinaceae plastomes in the span of less than two months. The development of a more effective short-read assembler (Ratan 2009) and assembly pipeline (Straub et al. 2011) also allowed for the rapid, automated assembly of longer and thus more accurate contigs. Both of these advancements are currently being utilized by collaborators on

the Gymnosperm Tree of Life and other projects, and should be broadly applicable (at least) for projects focusing on small genome sequencing and assembly.

Taken together, the above aspects of our work highlight both the gains and the challenges associated with applying massively parallel sequencing to phylogenetic questions in general, and clearly demonstrate that high-throughput machines are a powerful but not yet singular solution for phylogenetic pursuits. In turn, our results have several important implications specifically for *Pinus* systematics. The overall resolution seen in our final full-plastome, genus-wide phylogenetic analysis represents the most highly supported and well resolved topology to date for the worlds pine species. The application of the entire plastome results in strong support for the major divisions of the genus (subgenera, sections, subsections) documented by Gernandt et al. (2005), while high levels of overall support (ingroup nodes resolved with nearly 90% average bootstrap support) almost completely resolve the polytomies that previously dominated species-rich subsections. The removal of putative phylogenetic noise also results in important topological changes in three historically problematic clades. *Pinus krempfii* (subsection *Krempfianae*) is confidently placed as sister to subsection *Gerardianae* of section *Quinquefoliae* for the first time, and more closely allies chloroplast sequence-based results with morphological analyses (Pilger 1926, Ickert-Bond 2001). Similarly, *Pinus merkusii* and *Pinus latteri* are strongly resolved as sister to subsection *Pinaster* of section *Pinus*, also reflective of morphological analysis (Frankis 1993) and counter to previous molecular analyses (Gernandt et al. 2005, Parks et al. 2009). Resolution of subsection *Contortae*, on the other hand, was greatly diminished with removal of phylogenetic noise, and suggested some affinity to subsection *Australes* of section *Trifoliae*. Although strong support was not reached with noise removal in this case, the resulting trend still more closely reflects that seen in other sources of data, including crossability studies (Critchfield 1963, Saylor and Koenig 1967), the known fossil record (Miller Jr. 1992, McKown et al. 2002) and other molecular studies (Strauss and Doerksen 1990, Liston et al. 2003, Syring et al. 2005). Finally, the identification of the putative protein-coding locus *ycf*1 (and to a lesser extent *ycf*2) as highly variable is of consequence for future Pinaceae studies. As such, *ycf*1 has already been applied to a detailed phylogenetic analysis of subsection *Ponderosae* (Gernandt et al. 2009), and is currently being utilized as a species-identifier for studies of closely related Asian white pines (Handy et al. 2011).

To conclude, it is clear that in spite of any current limitations, advances in DNA sequencing are having a striking effect on phylogenetic and broader biological research, and will continue to do so. As read lengths and sequencing output continue to increase, it is also reasonable to expect that the assemblies of even complex genomes will become increasingly tractable and at the same time more accurate, although issues of data storage will likely present continuing challenges. Concurrently, as the field of phylogenetics transitions into the world of phylogenomics, astute approaches toward data management, screening and analysis will bring unprecedented clarity to the historical and in some cases dynamic relationships of our planet's biota.

LITERATURE CITED.

Critchfield WB. 1963. Hybridization of the southern pines in California. Southern Forest Tree Improvement Committee Publications, 22:40-48.

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res., 36:e122.

Frankis MP. 1993. Morphology and affinities of *Pinus brutia*. International Symposium on *Pinus brutia* Ten. Ankara, Ministry of Forestry, p. 11-18.

GBET. 2011. Closure of the NCBI SRA and implication for the long-term future of genomics data storage. Genome Biol., 12:402.

Gernandt DS, Hernández-León S, Salgado-Hernández E, Rosa JAPdl. 2009. Phylogenetic relationships of *Pinus* subsection *Ponderosae* inferred from rapidly evolving cpDNA regions. Syst. Bot., 34:481-491.

Gernandt DS, Lopez G, Garcia SO, Liston A. 2005. Phylogeny and classification of *Pinus*. Taxon, 54:29-42.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotech., 27:182-189.

Handy SM, Parks M, Rader JI, Diachenko GW, Callahan J, Liston A, Deeds JR. 2011. Genetic identification of pine nuts obtained from consumers experiencing dysgeusia. Manuscript in preparation.

Ickert-Bond S. 2001. Reexamination of wood anatomical features in *Pinus krempfii* (Pinaceae). IAWA Journal, 22:355-365.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, *et al.* 2010. The sequence and de novo assembly of the giant panda genome. Nature, 463:311-317.

Liston A, Gernandt DS, Vining TF, Campbell CS, Pinero D. 2003. Molecular phylogeny of Pinaceae and *Pinus*. Acta Hort.:107-114.

Mardis ER. 2011. A decade's perspective on DNA sequencing technology. Nature, 470:198-203.

McKown AD, Stockey RA, Scheger CE. 2002. A new species of *Pinus* subgenus *Pinus* subsection *Contortae* from Pliocene sediments of Ch'ijee's Bluff, Yukon Territory, Canada. Int. J. Plant Sci., 163:687-697.

Miller Jr. CN. 1992. Preserved cones of *Pinus* from the Neogene of Idaho and Oregon. Int. J. Plant Sci., 153:147-154.

Parks M, Cronn R, Liston A. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biology, 7:84.

Parks M, Liston A, Cronn R. 2010. Meeting the challenges of non-referenced genome assembly from short-read seqeunce data. Acta Horticulturae (ISHS), 859:323-332.

Pilger R. 1926. Pinus. In: Engler A, Prantl K editors. Die Naturliche Pflanzenfamilien. Liepzig, Wilhelm Engelmann, p. 331-342.

Ratan A. 2009. Assembly algorithms for next-generation sequence data. *Dissertation*, The Pennsylvania State University.

Rusk N. 2009. Cheap third-generation sequencing. Nat Meth, 6:244-244.

Saylor LC, Koenig RL. 1967. The slash x sand pine hybrid. Silvae Genetica, 16:134-138.

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, *et al.* 2011. The genome of woodland strawberry (*Fragaria vesca*). Nat Genet, 43:109-116.

Straub SC, Fishbein M, Livshultz T, Foster Z, Parks M, Weitemier K, Cronn RC, Liston A. 2011. Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. BMC Genomics, 12:211.

Strauss SH, Doerksen AH. 1990. Restriction fragment analysis of pine phylogeny. Evolution, 44:1081-1096.

Syring J, Willyard A, Cronn R, Liston A. 2005. Evolutionary relationships among *Pinus* (Pinaceae) subsections inferred from multiple low-copy nuclear loci. Am. J. Bot., 92:2086-2100.

Whittall JB, Syring J, Parks M, Buenrostro J, Dick C, Liston A, Cronn R. 2010. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. Mol Ecol, 19 Suppl 1:100-114.

COMPREHENSIVE BIBLIOGRAPHY

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. Current Opinion in Plant Biology, 8:135-141.

Ahn S-M, Kim T-H, Lee S, Kim D, Ghang H, Kim D-S, Kim B-C, Kim S-Y, Kim W-Y, Kim C, et al. 2009. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res., 19:1622-1629.

Alfaro ME, Zoller S, Lutzoni F. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov Chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol., 20:255-266.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. Nat Rev Genet, 12:363-376.

Asif M, Mantri S, Sharma A, Srivastava A, Trivedi I, Gupta P, Mohanty C, Sawant S, Tuli R. 2010. Complete sequence and organization of the *Jatropha curcas* (Euphorbiaceae) chloroplast genome. Tree Genetics & Genomes, 6:941-952.

Atherton R, McComish B, Shepherd L, Berry L, Albert N, Lockhart P. 2010. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. Plant Methods, 6:22.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 456:53.

Bergsten J. 2005. A review of long-branch attraction. Cladistics, 21:163-193.

Birky CW, Jr. 1978. Transmission genetics of mitochondria and chloroplasts. Annu. Rev. Genet., 12:471-512.

Birky CW, Jr., Maruyama T, Fuerst P. 1983. An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts, and some results. Genetics, 103:513-527.

Bouille M, Bousquet J. 2005. Trans-species shared polymorphisms at orthologous nuclear gene loci among distant species in the conifer *Picea* (Pinaceae): implications for the long-term maintenance of genetic diversity in trees. Am. J. Bot., 92:63-73.

Brinkmann H, van der Giezen M, Zhou Y, de Raucourt GP, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol., 54:743-757.

Cai Z, Penaflor C, Kuehl JV, Leebens-Mack J, Carlson JE, dePamphilis CW, Boore JL, Jansen RK. 2006. Complete plastid genome sequences of *Drimys*, *Liriodendron*, and *Piper*: implications for the phylogenetic relationships of Magnoliids. BMC Evol. Biol., 6:77.

Chaisson MJ, Pevzner PA. 2008. Short read fragment assembly of bacterial genomes. Genome Res., 18:324-330.

Chevalier A. 1944. Notes sur les coniferes de l'Indochine. Revue de Botanique Appliquee et d'Agriculture Tropicale, 24:7-34.

Chung SM, Gordon VS, Staub JE. 2007. Sequencing cucumber (*Cucumis sativus* L.) chloroplast genomes identifies differences between chilling-tolerant and-susceptible cucumber lines. Genome, 50:215-225.

Collins FS, Lander ES, Rogers J, Waterston RH, Conso I. 2004. Finishing the euchromatic sequence of the human genome. Nature, 431:931-945.

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. Nature Meth., 5:887-893.

Critchfield WB. 1963. Hybridization of the southern pines in California. Southern Forest Tree Improvement Committee Publications, 22:40-48.

Critchfield WB. 1966. Crossability and relationships of the closed-cone pines. Silvae Genetica, 16:89-97.

Critchfield WB. 1975. Interspecific hybridization in *Pinus*: a summary review. In: Fowler DP, Yeatman CY editors. Symposium on Interspecific and Interprovenance Hybridization in Forest Trees, p. 99-105.

Critchfield WB. 1986. Hybridization and classification of the white pines (*Pinus* section *Strobus*). Taxon, 35:647-656.

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res., 36:e122.

Cronn R, Knaus B, Liston A, Maughan J, Parks M, Syring J, Udall J. Manuscript in preparation. Simplifying the complex: targeted sequencing strategies for population, phylogenetic, and genomic studies in plants.

De Ferre Y. 1953. Division du genre *Pinus* en quatre sous-genres. Academie des Sciences Compte Rendu, 236:226-228.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet, 2:762-768.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet, 6:361-375.

Doorduin L, Gravendeel B, Lammers Y, Ariyurek Y, Chin-A-Woeng T, Vrieling K. 2011. The complete chloroplast genome of 17 individuals of pest species *Jacobaea vulgaris*: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. DNA Research, 18:93-105.

Douady CJ, Delsuc F, Boucher Y, Doolittle WF, Douzery EJ. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol., 20:248-254.

Dowton M, Austin AD. 2002. Increased congruence does not necessarily indicate increased phylogenetic accuracy - the behavior of the incongruence length difference test in mixed-model analyses. Syst. Biol., 51:19-31.

Drescher A, Ruf S, Calsa T, Carrer H, Bock R. 2000. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J., 22:97-104.

Eckert AJ, Hall BD. 2006. Phylogeny, historical biogeography, and patterns of diversification for *Pinus* (Pinaceae): phylogenetic tests of fossil-based hypotheses. Mol. Phylogenet. Evol., 40:166-182.

Engelmann G. 1880. Revision of the genus *Pinus*, and description of *Pinus elliottii*. Transactions of the Saint Louis Academy of Science, 4:161-189.

Erickson DL, Spouge J, Resch A, Weigt LA, Kress JW. 2008. DNA barcoding in land plants: developing standards to quantify and maximize success. Taxon, 57:1304-1316.

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA. 2011. Local de novo assembly of RAD paired-end contigs using short sequencing reads. PLoS ONE, 6:e18561.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using Phred.II. Error probabilities. Genome Res., 8:186-194.

Farjon A. 1984. Pines: drawings and descriptions of the genus. Leiden, W. Backhuys.

Farris JS, Källersjö M, Kluge AG, Bult C. 1995. Constructing a significance test for incongruence. Syst. Biol., 44:570-572.

Fauron C, Allen J, Clifton S, Newton K. 2004. Plant mitochondrial genomes. In: Daniell H, Chase C editors. Molecular Biology and Biotechnology of Plant Organelles, Springer Netherlands, p. 151-177.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology, 27:401-410.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 39:783-791.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington.

Fishbein M, Hibsch-Jetter C, Oltis DES, Hufford L. 2001. Phylogeny of Saxifragales (angiosperms, eudicots): analysis of a rapid, ancient radiation. Syst. Biol., 50:817-847.

Fitch WM. 1979. Cautionary remarks on using gene expression events in parsimony procedures. Systematic Zoology, 28:375-379.

Fitch WM. 1984. Cladistic and other methods: problems, pitfalls, and potentials. In: Duncan T, Stuessy TF editors. Cladistics: Perspectives on the Reconstruction of Evolutionary History, Columbia University Press, p. 221-252.

Frankis MP. 1993. Morphology and affinities of *Pinus brutia*. International Symposium on *Pinus brutia* Ten. Ankara, Ministry of Forestry, p. 11-18.

Gaussen H. 1960. Les gymnospermes actuelles et fossiles. Fassicule VI. Les Coniferales. Chapter 11. Generalites, Genre *Pinus*. Travaux du Toulous Universite Laboratoire Forestier, p. 1-272.

GBET. 2011. Closure of the NCBI SRA and implication for the long-term future of genomics data storage. Genome Biol., 12:402.

Geada Lopez G, Kamiya K, Harada K. 2002. Phylogenetic relationships of diploxylon pines (subgenus *Pinus*) based on plastid sequence data. Int. J. Plant Sci., 163:737-747.

Gernandt DS, Lopez G, Garcia SO, Liston A. 2005. Phylogeny and classification of *Pinus*. Taxon, 54:29-42.

Gernandt DS, Magallon S, Geada Lopez G, Zeron Flores O, Willyard A, Liston A. 2008. Use of simultaneous analyses to guide fossil-based calibrations of Pinaceae phylogeny. Int. J. Plant Sci., 169:1086-1099.

Gernandt DS, Hernández-León S, Salgado-Hernández E, Rosa JAPdl. 2009. Phylogenetic relationships of *Pinus* subsection *Ponderosae* inferred from rapidly evolving cpDNA regions. Syst. Bot., 34:481-491.

Gibbs MJ, Armstrong JS, Gibbs AJ. 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences, 16:573-582.

Gilbert MTP, Drautz DI, Lesk AM, Ho SYW, Qi J, Ratan A, Hsu CH, Sher A, Dalen L, Gotherstrom A. 2008. Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes. Proc Natl Acad Sci USA, 105:8327.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotech., 27:182-189.

Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH. 2005. Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Mol. Biol. Evol., 22:1813-1822.

Goremykin V, Viola R, Hellwig F. 2009. Removal of noisy characters from chloroplast genome-scale data suggests revision of phylogenetic placements of *Amborella* and *Ceratophyllum*. J. Mol. Evol., 68:197-204.

Goremykin V, Nikiforova S, Bininda-Emonds O. 2010. Automated removal of noisy data in phylogenomic analyses. J. Mol. Evol., 71:319-331.

Graham SW, Olmstead RG. 2000. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. Am. J. Bot., 87:1712-1730.

Grotkopp E, Rejmánek M, Sanderson MJ, Rost TL, Soltis P. 2004. Evolution of genome size in pines (*Pinus*) and its life-history correlates: supertree analyses. Evolution, 58:1705-1729.

Gupta PK. 2009. Single-molecule DNA sequencing technologies for future genomics research. Trends in Biotechnology, 26:602-611.

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

Handy SM, Parks M, Rader JI, Diachenko GW, Callahan J, Liston A, Deeds JR. 2011. Genetic identification of pine nuts obtained from consumers experiencing dysgeusia. Manuscript in preparation.

Harris RS. 2007. Improved pairwise alignment of genomic DNA. Computer Science, Dissertation, The Pennsylvania State University.

Hendy MD, Penny D. 1989. A framework for the quantitative study of evolutionary trees. Syst. Biol., 38:297-309.

Herman DS, Hovingh GK, Iartchouk O, Rehm HL, Kucherlapati R, Seidman JG, Seidman CE. 2009. Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. Nature Meth.

Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J. 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res., 18:802-809.

Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, et al. 2008. Whole-genome sequencing and variant discovery in *C. elegans*. Nature Meth., 5:183-188.

Hillis DM, Huelsenbeck JP. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. J. Hered., 83:189-195.

Hillis DM, Bull JJ. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol., 42:182-192.

Hipp AL, Hall JC, Sytsma KJ. 2004. Congruence versus phylogenetic accuracy: revisiting the incongruence length difference test. Syst. Biol., 53:81-89.

Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, Bank Mvd, Chase MW, Cowan RS, Erickson DL, Fazekas AJ, et al. 2009. A DNA barcode for land plants. Proc Natl Acad Sci USA, 106:12794-12797.

Holt RA, Jones SJM. 2008. The new paradigm of flow cell sequencing. Genome Res., 18:839.

Hudson ME. 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. Molecular Ecology Resources, 8:3-17.

Huelsenbeck JP. 1991. Tree-length distribution skewness: an indicator of phylogenetic information. Syst. Biol., 40:257-270.

Ickert-Bond S. 2001. Reexamination of wood anatomical features in *Pinus krempfii* (Pinaceae). IAWA Journal, 22:355-365.

Illumina. 2007. Protocol for whole genome sequencing using Solexa technology. Biotechniques Protocol Guide:29.

Jansen RK, Kaittanis C, Saski C, Lee SB, Tomkins J, Alverson AJ, Daniell H. 2006. Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol. Biol., 6:32.

Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack J, Muller KF, Guisinger-Bellian M, Haberle RC, Hansen AK. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA, 104:19369.

Jansen RK, Saski C, Lee S-B, Hansen AK, Daniell H. 2011. Complete plastid genome sequences of three rosids (*Castanea, Prunus, Theobroma*): evidence for at least two independent transfers of *rpl*22 to the nucleus. Mol. Biol. Evol., 28:835-847.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet., 22:225-231.

Kalersjo M, Albert VA, Farris JS. 1999. Homoplasy increases phylogenetic structure. Cladistics, 15:91-93.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res., 33:511-518.

Kent WJ. 2002. BLAT - the BLAST-like alignment tool. Genome Res., 4:656-664.

Klopfstein S, Kropf C, Quicke DLJ. 2010. An evaluation of phylogenetic informativeness profiles and the molecular phylogeny of Diplazontinae (Hymenoptera, Ichneumonidae). Syst. Biol., 59:226-241.

Klymiuk AA, Stockey RA, Rothwell GW. 2011. The first organismal concept for an extinct species of Pinaceae. Int. J. Plant Sci., 172:294-313.

Kovach A, Wegrzyn J, Parra G, Holt C, Bruening G, Loopstra C, Hartigan J, Yandell M, Langley C, Korf I, et al. 2010. The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics, 11:420.

Kriebel HB. 1985. DNA sequence components of the *Pinus strobus* nuclear genome. Can. J. For. Res., 15:1-4.

Krupkin AB, Liston A, Strauss SH. 1996. Phylogenetic analysis of the hard pines (*Pinus* subgenus *Pinus*, Pinaceae) from chloroplast DNA restriction site analysis. Am. J. Bot., 83:489-498.

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol., 56:17-24.

Kubo T, Newton KJ. 2008. Angiosperm mitochondrial genomes and mutations. Mitochondrion, 8:5-14.

Kuhner MK, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol., 11:459-468.

Kurtz S, Phyllippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol., 5:R12.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W. 2001. Initial sequencing and analysis of the human genome. Nature, 409:860-921.

Langmead B, Trapnell C, Pop M, Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol., 10:R25.

Leebens-Mack J, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, dePamphilis CW. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol. Biol. Evol., 22:1948-1963.

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature, 456:66-72.

Librado P, Rozas J. 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics, 25:1451-1452.

Lidholm J, Gustafsson P. 1991. The chloroplast genome of the gymnosperm *Pinus contorta*: a physical map and a complete collection of overlapping clones. Curr. Genet., 20:161-166.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25:1754-1760.

Li R, Li Y, Kristiansen K, Wang J. 2008. SOAP: short oligonucleotide alignment program. Bioinformatics, 24:713-714.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, et al. 2010. The sequence and de novo assembly of the giant panda genome. Nature, 463:311-317.

Lin C-P, Huang J-P, Wu C-S, Hsu C-Y, Chaw S-M. 2010. Comparative chloroplast genomics reveals the evolution of Pinaceae genera and subfamilies. Genome Biology and Evolution, 2:504-517.

Liston A, Gernandt DS, Vining TF, Campbell CS, Pinero D. 2003. Molecular phylogeny of Pinaceae and *Pinus*. Acta Hort.:107-114.

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*. Mol. Ecol., 16:3926-3937.

Liston A, Robinson WA, Piñero D, Alvarez-Buylla ER. 1999. Phylogenetics of *Pinus* (Pinaceae) based on nuclear ribosomal DNA internal transcribed spacer region sequences. Mol. Phylogenet. Evol., 11:95-109.

Little EL, Critchfield WB. 1969. Subdivision of the genus *Pinus* (pines). In: U.S. Department of Agriculture FS editor. Washington, D.C., p. 1-51.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. 2010. Target-enrichment strategies for next-generation sequencing. Nat Meth, 7:111-118.

Mardis ER. 2008. Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics, 9:387-402.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet., 24:133-141.

Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K, Koboldt DC, Fulton RS, Delehaunty KD, McGrath SD, et al. 2009. Recurring mutations found by sequencing an acute myeloid leukemia genome. New England Journal of Medicine, 361:1058-1066.

Mardis ER. 2011. A decade's perspective on DNA sequencing technology. Nature, 470:198-203.

Martin DP, Posada D, Crandall KA, Williamson C. 2005a. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Research & Human Retroviruses, 21:98-102.

Martin DP, Williamson C, Posada D. 2005b. RDP2: recombination detection and analysis from sequence alignments, 21:260-262.

Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences, 16:562-563.

McKown AD, Stockey RA, Scheger CE. 2002. A new species of *Pinus* subgenus *Pinus* subsection *Contortae* from Pliocene sediments of Ch'ijee's Bluff, Yukon Territory, Canada. Int. J. Plant Sci., 163:687-697.

Metzker ML. 2010. Sequencing technologies - the next generation. Nat Rev Genet, 11:31-46.

Millar CI. 1998. Early evolution of pines. In: Richardson DM editor. Ecology and biogeography of *Pinus*. Cambridge, Cambridge University Press, p. 69–91.

Miller Jr. CN. 1992. Preserved cones of *Pinus* from the Neogene of Idaho and Oregon. Int. J. Plant Sci., 153:147-154.

Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Folta KM, Soltis DE. 2006. Rapid and accurate pyrosequencing of angiosperm plastid genomes. BMC Plant Biol., 6:17.

Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci USA, 104:19363.

Neale DB, Kremer A. 2011. Forest tree genomics: growing resources and applications. Nat Rev Genet, 12:111-122.

Neubig KM, Whitten WM, Carlsward BS, Blanco MA, Endara L, Williams NH, Moore M. 2009. Phylogenetic utility of *ycf*1 in orchids: a plastid gene more variable than *mat*K. Plant Syst. Evol., 277:75-84.

Nock CJ, Waters DLE, Edwards MA, Bowen SG, Rice N, Cordeiro GM, Henry RJ. 2011. Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnology Journal, 9:328-333.

Nylander JAA. 2004. MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

Oda K, Yamato K, Ohta E, Nakamura Y, Takemura M, Nozato N, Akashi K, Kanegae T, Ogura Y, Kohchi T, et al. 1992. Gene organization deduced from the complete sequence of liverwort *Marchantia polymorpha* mitochondrial DNA : a primitive form of plant mitochondrial genome. Journal of Molecular Biology, 223:1-7.

Ortiz Garcia S. 1999. Evolucion y fiogenia en pinos y sus hongos endofitos: aspectos sistematicos de la coespeciacion. Instituto de Ecologia. Mexico City, Universidad Nacional Autonoma de Mexico.

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. 2008. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res., 18:2024.

Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new geminiviruses by frequent recombination. Virology, 265:218-225.

Palmé AE, Pyhäjärvi T, Wachowiak W, Savolainen O. 2009. Selection on nuclear genes in a *Pinus* phylogeny. Mol. Biol. Evol., 26:893-905.

Palmer JD. 1985. Comparative organization of chloroplast genomes. Annu. Rev. Genet., 19:325-354.

Palmer JD. 1990. Contrasting modes and tempos of genome evolution in land plant organelles. Trends Genet., 6:115-120.

Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu Y-L, Song K. 2000. Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proc. Acad. Nat. Sci. Phila., 97:6960-6966.

Pan X, Urban AE, Palejev D, Schulz V, Grubert F, Hu Y, Snyder M, Weissman SM. 2008. A procedure for highly specific, sensitive, and unbiased whole-genome amplification. Proc. Acad. Nat. Sci. Phila., 105:15499-15504.

Parks M, Cronn R, Liston A. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biology, 7:84.

Parks M, Liston A, Cronn R. 2010. Meeting the challenges of non-referenced genome assembly from short-read seqeunce data. Acta Horticulturae (ISHS), 859:323-332.

Parks M, Liston A, Cronn R. 2011. Newly developed primers for complete *ycf*1 amplification in *Pinus* (Pinaceae) chloroplasts with possible family-wide utility. Am. J. Bot.:in press.

Patenaude NJ, Portway VA, Schaeff CM, Bannister JL, Best PB, Payne RS, Rowntree VJ, Rivarola M, Baker CS. 2007. Mitochondrial DNA diversity and population structure among southern right whales (*Eubalaena australis*). J. Hered., 98:147-157.

Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. Ecology, Evolution and Systematics, 36:541-562.

Philippe H, Frederic, D., Henner, B., and Lartillot, N. 2005. Phylogenomics. Annu. Rev. Ecol. Evol. Syst., 36:541-C-542.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol, 9:e1000602.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol., 21:1455-1458.

Pilger R. 1926. *Pinus*. In: Engler A, Prantl K editors. Die Naturliche Pflanzenfamilien. Liepzig, Wilhelm Engelmann, p. 331-342.

Pollard DA, Iyer VN, Moses AM, Eisen MB, McAllister BF. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genet, 2:e173.

Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol., 51:664-671.

Pop M, Salzberg SL. 2008. Bioinformatics challenges of new sequencing technology. Trends Genet., 24:142-149.

Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, et al. 2007. Multiplex amplification of large sets of human exons. Nature Meth., 4:931-936.

Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics, 14:817-818.

Price RA, Liston A, Strauss SH. 1998. Phylogeny and Systematics of *Pinus*. In: Richardson DM editor. Ecology and Biogeography of *Pinus*. Cambridge, Cambridge University Press, p. 49-68.

Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. 2008. A large genome center's improvements to the Illumina sequencing system. Nature Meth., 5:1005-1010.

Ratan A. 2009. Assembly algorithms for next-generation sequence data. Dissertation, The Pennsylvania State University.

Richardson DM, Rundel PW. 1998. Ecology and biogeography of *Pinus*: an introduction. In: Richardson DM editor. Ecology and Biogeography of *Pinus*. Cambridge, Cambridge University Press.

Richterich P. 1998. Estimation of errors in “raw” DNA sequences: a validation study. Genome Res., 8:251-259.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Mathematical Biosciences, 53:131-147.

Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol., 56:389-399.

Rokas A, Carroll SB. 2006. Bushes in the tree of life. PLoS Biol, 4:e352.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, 19:1572-1574.

Rusk N. 2009. Cheap third-generation sequencing. Nat Meth, 6:244-244.

Saylor LC, Koenig RL. 1967. The slash x sand pine hybrid. Silvae Genetica, 16:134-138.

Schweiger M, Kerick M, Timmermann B, Isau M. 2011. The power of NGS technologies to delineate the genome organization in cancer: from mutations to structural variations and epigenetic alterations. Cancer and Metastasis Reviews, 30:199-210.

Seeb JE, Carvalho G, Hauser L, Naish K, Roberts S, Seeb LW. 2011. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Molecular Ecology Resources, 11:1-8.

Shaw GR. 1914. The genus *Pinus*. Forage Village, The Murray Printing Co.

Shaw GR. 1924. Notes on the genus *Pinus*. Journal of the Arnold Arboretum 5:225-227.

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

Shaw J, Lickey EB, Schilling EE, Small RL. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. Am. J. Bot., 94:275.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nat Biotech, 26:1135-1145.

Shen R, Mockler TC. RGA - a reference-guided assembler. http://rga.cgrb.oregonstate.edu/rga_about.html. Manuscript in preparation.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol, 16:1114-1116.

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, et al. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. The EMBO Journal, 5:2043-2049.

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, et al. 2011. The genome of woodland strawberry (*Fragaria vesca*). Nat Genet, 43:109-116.

Simon SA, Zhai J, Nandety RS, McCormick KP, Zeng J, Mejia D, Meyers BC. 2009. Short-read sequencing technologies for transcriptional analyses. Annual Review of Plant Biology, 60:305-333.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: A parallel assembler for short read sequence data. Genome Res., 19:1117-1123.

Smith JM. 1992. Analyzing the mosaic structure of genes. J. Mol. Evol., 34:126-129.

Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu Y-L, Chase MW, Farris JS, Stefanovic S, Rice DW, Palmer JD, et al. 2004. Genome-scale data, Angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. Trends Plant Sci., 9:477-483.

Stamatakis A. 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol., 57:758-771.

Stefanovic S, Rice D, Palmer J. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? BMC Evol. Biol., 4:35.

Straub SC, Fishbein M, Livshultz T, Foster Z, Parks M, Weitemier K, Cronn RC, Liston A. 2011. Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. BMC Genomics, 12:211.

Strauss SH, Doerksen AH. 1990. Restriction fragment analysis of pine phylogeny. Evolution, 44:1081-1096.

Strimmer K, von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Acad. Nat. Sci. Phila., 94:6815-6819.

Suzuki Y, Glazko GV, Nei M. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc Natl Acad Sci USA, 99:16138-16143.

Swofford DL. 2000. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland, Massachusetts, Sinauer Associates.

Syring J, Willyard A, Cronn R, Liston A. 2005. Evolutionary relationships among *Pinus* (Pinaceae) subsections inferred from multiple low-copy nuclear loci. Am. J. Bot., 92:2086-2100.

Syring J, Farrell K, Businsky R, Cronn R, Liston A. 2007. Widespread genealogical nonmonophyly in species of *Pinus* subgenus *Strobus*. Syst. Biol., 56:163-181.

Szmidt A, Wang XR, Changtragoon S. 1996. Contrasting patterns of genetic diversity in two tropical pines: *Pinus kesiya* (Royle ex Gordon) and *P. merkusii* (Jungh et De Vriese). TAG Theoretical and Applied Genetics, 92:436-441.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol., 24:1596-1599.

Tangphatsornruang S, Sangsrakru D, Chanprasert J, Uthaipaisanwong P, Yoocha T, Jomchai N, Tragoonrung S. 2010. The chloroplast genome sequence of mungbean (*Vigna radiata*) determined by high-throughput pyrosequencing: structural organization and phylogenetic relationships. DNA Research, 17:11-22.

Townsend JP. 2007. Profiling phylogenetic informativeness. Syst. Biol., 56:222-231.

Townsend JP, Leuenberger C. 2011. Taxon sampling and the optimal rates of evolution for phylogenetic inference. Syst. Biol.

Van der Berg J. 1973. Holzer der niederrheinischen braunkohlenformation 2. Holzer der braunkohlengruben "Maria Theresia" zu Herzogenrath, "Zukunft West" zu Eschweiler und "Victor" Zulpich mitte zu Zulpich. Nebst einer systematisch-anatomischen bearbeitung der gattung *Pinus* L. Review of Palaeobotany and Palynology, 15:73-275.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA. 2001. The sequence of the human genome, 291:1304-1351.

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*. Proc Natl Acad Sci USA, 91:9794-9798.

Wang XR, Szmidt AE, Nguyên HN. 2000. The phylogenetic position of the endemic flat-needle pine *Pinus krempfii* (Pinaceae) from Vietnam, based on PCR-RFLP analysis of chloroplast DNA. Plant Syst. Evol., 220:21-36.

Wang XR, Tsumura Y, Yoshimaru H, Nagasaka K, Szmidt AE. 1999. Phylogenetic relationships of Eurasian pines (*Pinus*, Pinaceae) based on chloroplast *rbc*L, *mat*K, *rpl*20-*rps*18 spacer, and *trn*V intron sequences. Am. J. Bot., 86:1742-1753.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J. 2008. The diploid genome sequence of an Asian individual. Nature, 456:60-65.

Wendel JF. 2000. Genome evolution in polyploids. Plant Molecular Biology, 42:225-249.

Whiteford N, Haslam N, Weber G, Prugel-Bennett A, Essex JW, Roach PL, Bradley M, Neylon C. 2005. An analysis of the feasibility of short read sequencing. Nucleic Acids Res., 33:e171.

Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol., 22:258-265.

Whittall JB, Syring J, Parks M, Buenrostro J, Dick C, Liston A, Cronn R. 2010. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. Mol Ecol, 19 Suppl 1:100-114.

Willyard A, Syring J, Gernandt DS, Liston A, Cronn R. 2007. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for Pinus. Mol. Biol. Evol., 24:90-101.

Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Acad. Nat. Sci. Phila., 84:9054-9058.

Wortley AH, Rudall PJ, Harris DJ, Scotland RW. 2005. How Much Data are Needed to Resolve a Difficult Phylogeny? Case Study in Lamiales. Syst. Biol., 54:697-709.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics, 20:3252-3255.

Yang M, Zhang X, Liu G, Yin Y, Chen K, Yun Q, Zhao D, Al-Mssallem IS, Yu J. 2010. The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). PLoS ONE, 5:e12762.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res., 18:821-829.

Zhang ZY, Li DZ. 2004. Molecular phylogeny of section *Parrya* of *Pinus* (Pinaceae) based on chloroplast *mat*K gene sequence data. Acta Bot. Sin., 46:171-179.

Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol., 51:588-598.

APPENDICES

Appendix A

Chapter II Supplementary Table

Appendix Table 2.1. Meta-analysis details.

Appendix Table 2.1.

Study Citation	Genus	# of Loci	Aligned bp	# of taxa from genus
Andersson, L., Kocsis, M. and Eriksson, R. 2006 Relationships of the genus Azorella (Apiaceae) and other hydrocotyloids inferred from sequence variation in three plastid markers. Taxon 55: 270-280.	*Azorella*	3	3470	15
Bellstedt, D.U., van Zyl, L., Marais, E.M., Bytebier, B., de Villiers, C.A., Makwarela, A.M. and Dreyer, L.L. 2008 Phylogenetic relationships, character evolution and biogeography of southern African members of Zygophyllum (Zygophyllaceae) based on three plastid regions. Mol. Phylogenet. Evol. 47: 932-950.	*Zygophyllum*	2	2311	31
Bellusci, F., Pellegrino, G., Palermo, A.M. and Musacchio, A. 2008 Phylogenetic relationships in the orchid genus Serapias L. based on noncoding regions of the chloroplast genome. Mol. Phylogenet. Evol. 47: 986-991.	*Serapias*	4	1946	15
Bessega, C., Hopp, H.E. and Fortunato, R.H. 2008 Toward a phylogeny of Mimosa (Leguminosae: Mimosoidae): a preliminary analysis of southern South American species based on chloroplast DNA sequence. Ann. Mo. Bot. Garden 95: 567-579.	*Mimosa*	2	1147	28
Bruneau, A., Starr, J.R. and Joly, S. 2007 Phylogenetic relationships in the genus Rosa: new evidence from chloroplast DNA sequences and an appraisal of current knowledge. Syst. Bot. 32: 366-378.	*Rosa*	4	1294	72
Calvino, C.L. and Downie, S.R. 2007 Circumscription and phylogeny of Apiaceae subfamily Saniculoideae based on chloroplast DNA sequences. Mol. Phylogenet. Evol. 44: 175-191.	*Eryngium*	3	4846	35
	Sanicula / Hacquetia	3	4846	15
	Alepidea	3	4846	13
Carlsward, B.S., Whitten, W.M., Williams, N.H. and Bytebier, B. 2006 Molecular phylogenetics of Vandeae (Orchidaceae) and the evolution of leaflessness. Am. J. Bot. 93: 770-786.	*Aerangis*	2	3173	16
Chacon, J., Madrinan, S., Chase, M.W. and Bruhl, J.J. 2006 Molecular phylogenetics of Oreobolus (Cyperaceae) and the origin and diversification of the American species. Taxon 55: 359-366.	*Oreobolus*	1	1028	14
Chiapella, J. 2007 A molecular phylogenetic study of Deschampsia (Poaceae: Aveneae) inferred from nuclear ITS and plastid trnL sequence data: support for the recognition of Avenella and Vahlodea. Taxon 56: 55-64.	*Deschampsia*	1	584	15
Choong, C.Y., Wickneswari, R., Norwati, M. and Abbott, R.J. 2008 Phylogeny of Hopea (Dipterocarpaceae) inferred from chloroplast DNA and nuclear PgiC sequences. Mol. Phylogenet. Evol. 48: 1238-1243.	*Hopea*	3	2351	18
Chung, K.-F. 2007 Inclusion of the South Pacific alpine genus Oreomyrrhis (Apiaceae) in Chaerophyllum based on nuclear and chloroplast DNA sequences. Syst. Bot. 32: 671-681.	*Oreomyrrhis*	2	1342	34
	Piptochaetium	2	2090	22
	Nassela	2	2090	21
Covreur, T.L.P., Hahn, W.J., de Granville, J.-J., Pham, J.-L., Ludena, B., Pintaud, J.-C. 2007 Phylogenetic relationships of the cultivated neotropical palm Bactris Gasipaes (Arecaceae) with its wild relatives inferred from chloroplast and nuclear DNA polymorphisms. Syst. Bot. 32: 519-530.	*Bactris*	2	2023	30
Cuenca, A. and Asmussen-Lange, C.B. 2007 Phylogeny of the palm tribe Chamaedoreeae (Arecaceae) based on plastid DNA sequences. Syst. Bot. 32: 250-263.	*Chamaedorea*	4	4746	27

Appendix Table 2.1.

Driscoll, H.E. and Barrington, D.S. 2007 Origin of Hawaiian Polystichum (Dryopteridaceae) in the context of a world phylogeny. Am. J. Bot. 94: 1413-1424.	*Polystichum*	2	1699	51
Duangjai, S., Wallnofer, B., Samuel, R., Munzinger, J. and Chase, M.W. 2006 Generic delimitation and relationships in Ebenaceae sensu lato: evidence from six plastid DNA regions. Am. J. Bot. 93: 1808-1827.	*Diospyros*	6	6440	87
Eckert, A.J., and Hall, B.D. 2006 Phylogeny, historical biogeography, and patterns of diversification for Pinus (Pinacease): phylogenetic tests of fossil-based hypotheses. Mol. Phylogenet. Evol. 40: 166-182.	*Pinus*	4	3288	83
Ekenas, Catarina, Baldwin, B.G. and Andreasen, K. 2007 A molecular phylogenetic study of Arnica (Asteraceae): low chloroplast DNA variation and problematic subgeneric classification. Syst. Bot. 32: 917-928.	*Arnica*	5	3710	27
Erkens, R.H.J., Chatrou, L.W., Koek-Noorman, J., Maas, J.W. and Maas, P.J.M. 2007 Classification of a large and widespread genus of Neotropical trees, Guatteria (Annonaceae) and its three satellite genera Guatteriella, Guatteriopsis and Heteropetalum. Taxon 56: 757-774.	*Guatteria*	4	4000	113
Essi, L., Longhi-Wagner, H.M., de Souza-Chies, T.T. 2008 Phylogenetic analysis of the Briza complex (Poaceae). Mol. Phylogenet. Evol. 47: 1018-1029.	*Briza*	2	1028	29
Ford, K.A., Ward, J.M., Smissen, R.D., Wagstaf, S.J. and Breitwieser, I. 2007 Phylogeny and biogeography of Craspedia (Asteraceae: Gnaphalieae) based on ITS, ETS and psbA-trnH sequence data. Taxon 56: 783-794.	*Craspedia*	1	603	29
Fritsch, P.W., Cruz, B.C., Almeda, F., Wang, Y. and Shi, S. 2006 Phylogeny of Symplocos based on DNA sequences of the chloroplast trnC–trnD intergenic region. Syst. Bot. 31: 181-192.	*Symplocos*	1	3177	73
Gaskin, J.F. and Wilson, L.M. 2007 Phylogenetic relatiohships among native and naturalized Hieracium (Asteraceae) in Canada and the United States based on plastid DNA sequences. Syst. Bot. 32: 478-485.	*Hieracium*	4	2066	37
Gehrke, B., Brauchler, C., Romoleroux, K., Lundberg, M., Heubl, G. and Eriksson, T. 2008 Molecular phylogenetics of Alchemilla, Aphanes and Lachemilla (Rosaceae) inferred from plastid and nuclear intron and spacer DNA sequences, with comments on generic classification. Mol. Phylogenet. Evol. 47: 1030-1044.	*Alchemilla (and Aphanes, Lach*	2	1241	85
Gernandt, D.S., Lopez, G.G., Garcia, S.O., and Liston, A. 2005 Phlyogeny and classification of Pinus. Taxon 54: 29-42.	*Pinus*	2	2817	101
Ghebretinsae, A.J., Thulin, M. and Barber, J.C. 2007 Relationships of cucumbers and melons unraveled: molecular phylogenetics of Cucumis and related genera (Benincaseae, Cucurbitaceae). Am. J. Bot. 94: 1256-1266.	*Cucumis*	1	818	36
Grose, S.O. and Olmstead, R.G. 2007 Evolution of a charismatic neotropical clade: molecular phylogeny of Tabebuia s.l., Crescentieae, and allied genera (Bignoniaceae). Syst. Bot. 32: 650-659.	*Tabebuia*	2	3106	15
Gussarova, G., Popp, M., Vitek, E. and Brochmann, C. 2008 Molecular phylogeny and biogeography of the bipolar Euphrasia (Orobanchaceae): recent radiations in an old genus. Mol. Phylogenet. Evol. 48: 444-460.	*Euphrasia*	3	1864	50
Hansen, A.K., Gilbert, L.E., Simpson, B.B., Downie, S.R., Cervi, A.C. and Jansen, R.K. 2006 Phylogenetic relationships and chromosome number evolution in Passiflora. Syst. Bot. 31: 138-150.	*Passiflora*	1	631	57

Appendix Table 2.1.

Hennequin, S., Ebihara, A., Ito, M., Iwatsuki, K. and D., J.-Y. 2006 New insights into the phylogeny of the genus Hymenophyllum s.l. (Hymenophyllaceae): revealing the polyphyly of Mecodium. Syst. Bot. 31: 271-284.	*Hymenophyllum*	3	3900	24
Huang, M., Crawford, D.J., Freudenstein, J.V., and Cantino, P.D. 2008 Systematics of Trichostema (Lamiaceae): evidence from ITS, ndhF, and morphology. Syst. Bot. 33: 437-446.	*Trichostema*	1	2147	18
Huertas, M.L., Schneider, J.V. and Zizka, G. 2007 Phylogenetic analysis of Palaua (Malveae, Malvaceae) based on plastid and nuclear sequences. Syst. Bot. 32: 157-165.	*Palaua*	1	621	16
Janssens, S., Geuten, K., Yuan, Y.-M., Song, Y., Kupfer, P. and Smets, E. 2006 Phylogenetics of Impatiens and Hydrocera (Balsaminaceae) using chloroplast atpB-rbcL spacer sequences. Syst. Bot. 31: 171-180.	*Impatiens*	1	961	84
Jobson, R.W. and Luckow, M. 2007 Phylogenetic study of the genus Piptadenia (Mimosoideae: Leguminosae) using plastid trnl-f and trnk/matk sequence data. Syst. Bot. 32: 569-575.	*Piptadenia*	2	4145	14
Kadereit, J.W., Repplinger, M., Schmalz, N., Uhink, C.H. and Worz, A. 2008 The phylogeny and biogeography of Apiaceae subf. Saniculoideae tribe Saniculeae: from south to north and south again. Taxon 57: 365-382.	*Eryngium*	1	1177	15
Kathriarachchi, H., Samuel, R., Hoffmann, P., Mlinarec, J., Wurdack, K.J., Ralimanana, H., Stuessy, T.F. and Chase, M.W. 2006 Phylogenetics of tribe Phyllantheae (Phyllanthaceae; Euphorbiaceae sensu lato) based on nrITS and plastid matK DNA sequence data. Am. J. Bot. 93: 637-655.	*Phyllanthus*	2	2050	84
Khan, S.A., Razafimandimbison, S.G., Bremer, B. and Liede-Schumann, S. 2008 Sabiceeae and Virectarieae (Rubiaceae, Ixoroideae): one or two tribes? New tribal and generic circumscriptions of Sabiceeae and biogeography of Sabicea s.l.. Taxon 57: 7-23.	*Sabicea*	1	2348	30
Kim, S.-T, and Donoghue, M.J. 2008 Molecular phylogeny of Persicaria (Persicarieae, Polygonaceae). Syst. Bot. 33: 77-86.	*Pesicaria*	4	3667	15
Kirkpatrick, R.E.B. 2007 Investigating the monophyly of Pellaea (Pteridaceae) in the context of a phylogenetic analysis of Cheilanthoid ferns. Syst. Bot. 32: 504-518.	*Pellaea*	3	1722	15
	Cheilanthes	3	1722	12
Kocyan, A., de Vogel, E.F., Conti, E. and Gravendeel, B. 2008 Molecular phylogeny of Aerides (Orchidaceae) based on one nuclear and two plastid markers: a step forward in understanding the evolution of the Aeridinae. Mol. Phylogenet. Evol. 48: 422-443.	*Aerides*	2	3101	19
Korall, P., Conant, D.S., Metzgar, J.S., Schneider, H. and Pryer, K.M. 2007 A molecular phylogeny of scaly tree ferns (Cyatheaceae). Am. J. Bot. 94: 873-886.	*Alsophila*	5	5135	25
	Cyathea	5	5135	21
	Sphaeropteris	5	5135	17
Krings, A., Thomas, D.T. and Xiang, Q.-y. 2008 On the generic circumscription of Gonolobus (Apocynaceae, Asclepiadoideae): evidence from molecules and morphology. Syst. Bot. 33: 403-415.	*Metalea*	2	1735	31
	Gonolobus	2	1735	26

Appendix Table 2.1.

Kurata, K., Jaffre, T. and Setoguchi, H. 2008 Genetic diversity and geographical structure of the pitcher plant Nepenthes vieillardii in New Caledonia: a chloroplast DNA haplotype analysis. Am. J. Bot. 95: 1632-1644.	*Nepenthes*	2	4660	17
Kwembeya, E.G., Bjora, C.S., Stedje, B. and Nordal, I. 2007 Phylogenetic relationships in the genus Crinum (Amaryllidaceae) with emphasis on tropical African species: evidence from trnL-F and nuclear ITS DNA sequence data. Taxon 56: 801-810.	*Crinum*	1	844	35
Lahaye, R., Klackenberg, J., Kallersjo, M., van Campo, E. and Civeyrel, L. 2007 Phylogenetic relationships between derived Apocynaceae s.l. and within Secamonoideae based on chloroplast sequences. Ann. Mo. Bot. Garden 94: 376-391.	*Secamone*	4	3557	20
Lhomann, L.G. 2006 Untangling the phylogeny of neotropical lianas (Bignonieae, Bignoniaceae). Am. J. Bot. 93: 304-318.	*Arrabidaea*	1	2125	20
Li, J.-M. and Wang, Y.-Z. 2007 Phylogenetic relationships of Zeugites (Poaceae: Centothecoideae) inferred from plastid and nuclear DNA sequences and morphology. Syst. Bot. 32: 888-898.	*Chiritopsis / Chirita*	1	893	22
Lo, E.Y.Y., Stefanovic, S. and Dickinson, T.A. 2007 Molecular reappraisal of relationships between Crataegus and Mespilus (Rosaceae, Pyreae) - two genera or one? Syst. Bot. 32: 596-616.	*Crataegus*	4	2085	33
Lu, J.-M., Barrington, D.S. and Li, D.-Z. 2007 Molecular phylogeny of the Polystichoid ferns in Asia based on rbcL sequences. Syst. Bot. 32: 26-33.	*Polystichum s.s*	1	1320	16
	Cyrtomium s.s.	1	1320	15
Luebert, F., and Wen, J. 2008 Phylogenetic analysis and evolutionary diversification of Heliotropium sect. Cochranea (Heliotropiaceae) in the Atacama desert. Syst. Bot. 33: 390-402.	*Heliotropium*	1	2017	26
Manning, J., Forest, F. and Vinnersten, A. 2007 The genus Colchicum L. redefined to include Androcymbium Willd. based on molecular evidence. Taxon 56: 872-882.	*Colchicum*	3	3830	27
Marazzi, B., Endress, P.K., Paganucci de Queiroz, L. and Conti, E. 2006 Phylogenetic relationships within Senna (Leguminosae, Cassiinae) based on three chloroplast DNA regions: patterns in the evolution of floral symmetry and extrafloral nectaries. Am. J. Bot. 93: 288-303.	*Senna*	3	2909	83
Martinez, A.M.S., Salazar, G.A. and Aranda, P.A. 2007 Phylogenetic relationships of Zeugites (Poaceae: Centothecoideae) inferred from plastid and nuclear DNA sequences and morphology. Syst. Bot. 32: 722-730.	*Zeugites*	1	639	11
Micheneau, C., Carlsward, B.S., Fay, M.F., Bytebier, B., Pailler, T. and Chase M.W. 2008 Phylogenetics and biogeography of Mascarene angraecoid orchids (Vandeae, Orchidaceae). Mol. Phylogenet. Evol. 46: 908-924.	*Angraecum (and Bonniera)*	4	4747	52
	Jumellea	4	4747	21
	Aeranthes	4	4747	12
Monro, A.K. 2006 The revision of species-rich genera: a phylogenetic framework for the strategic revision of Pilea (Urticaceae) based on cpDNA, nrDNA, and morphology. Am. J. Bot. 93: 426-441.	*Pilea*	1	1409	94
Moody, M.L. and Les, D.H. 2007 Phylogenetic systematics and character evolutin in the angiosperm family Haloragaceae. Am. J. Bot. 94: 2005-2025.	*Myriophyllum*	2	2298	35
	Gonocarpus	2	2298	22
	Haloragis	2	2298	14

Appendix Table 2.1.

Ohi-Toma, T., Sugawara, T., Murata, H., Wnke, S., Neinhuis, C. and Murata, J. 2006 Molecular phylogeny of Aristolochia sensu lato (Aristolochiaceae) based on sequences of rbcL, matK, and phyA genes, with special reference to differentiation of chromosome numbers. Syst. Bot. 31: 481-492.	*Aristilochia*	1	1425	20
Olmstead, R.G., Bohs, L., Migic, H.A., Santiago-Valentin, E., Garcia, V.F. and Collier, S.M. 2008 A molecular phylogeny of the Solanaceae. Taxon 57: 1159-1181.	*Solanum*	2	3885	11
Prince, L.M. and Kress, W.J. 2006 Phylogenetic relationships and classification in Marantaceae: insights from plastid DNA sequence data. Taxon 55: 281-296.	*Calathea*	3	2493	18
Ran, J.-H., Wei, X.-X. and Wang, X.-Q. 2007 Molecular phylogeny and biogeography of Picea (Pinaceae): implications for phylogeographical studies using cytoplasmic haplotypes. Mol. Phylogenet. Evol. 41: 405-419.	*Picea*	2	4167	34
Rouhan, G., Rakotondrainibe, F. and Moran R.C. 2007 Elaphoglossum nidusoides (Dryopteridaceae), a new species of fern from Madagascar with an unusual phylogenetic position in the Squamipedia group. Syst. Bot. 32: 227-235.	*Elaphoglossum*	3	1911	31
Salino, A., almeida, T.E., Smith, a.R., gomez, A.N., Kreier, H.-P. and Schneider, H. 2008 A new species of Microgramma (Polypodiaceae) from Brazil and recircumscription of the genus based on phylogenetic evidence. Syst. Bot. 33: 630-635.	*Microgramma*	4	2545	12
Salvo, G., Bacchetta, G., Ghahremaninejad, F. and Conti, E. 2008 Phylogenetic relationships of Ruteae (Rutaceae): new evidence from the chloroplast genome and comparisons with non-molecular data. Mol. Phylogenet. Evol. 49: 736-748.	*Haplophyllum*	3	3594	22
Samuel, R., Gutermann, W., Stuessy, T.F., Ruas, C.F., Lack, H.-W., Tremetsberger, K., Talavera, S., Hermanowski, B. and Ehrendorfer, F. 2006 Molecular phylogenetics reveals Leontodon (Asteraceae, Lactuceae) to be diphyletic. Am. J. Bot. 93: 1193-1205.	*Leontodon*	1	938	36
Selvi, F., Bigazzi, M., Hilger, H.H. and Papini, A. 2006 Molecular phylogeny, morphology and taxonomic re-circumscription of the generic complex Nonea/Elizaldia/Pulmonaria/Paraskevia (Boraginaceae-Boragineae). Taxon 55: 907-918.	*Nonea*	1	472	18
Simoes, A.O., Endress. M.E., van der Niet, T., Kinoshita, L.S. and Conti, E. 2006 Is Mandevilla (Apocynaceae, Mesechiteae) monophyletic? Evidence from five plastid DNA loci and morphology. Ann. Mo. Bot. Garden 93: 565-591.	*Mandevilla*	5	6344	64
Simpson, B., Larkin, L., Weekds, A., and McDill, J. 2006 Phylogeny and biogeography of Pomaria (Caesalpinioideae: Leguminosae). Syst. Bot. 31: 792-804.	*Pomaria*	3	1559	19
Soejima, A. and Wen, J. 2006 Phylogenetic analysis of the grape family (Vitaceae) based on three chloroplast markers. Am. J. Bot. 93: 278-287.	*Ampelopsis*	1	1189	12
	Vitis	1	1189	10
Su, Y.C.F, Smith, G.J.D. and Saunders, R.M.K. 2008 Phylogeny of the basal angiosperm genus Pseuduvaria (Annonaceae) inferred from five chloroplast DNA regions, with interpretation of morphological character evolution. Mol. Phylogenet. Evol. 48: 188-207.	*Pseuduvaria*	5	4473	52
Torke, B.M. and Schaal, B.A. 2008 Molecular phylogenetics of the species-rich neotropical genus Swartzia (Leguminosae, Papilionoideae) and related genera of the swartzioid clade. Am. J. Bot. 95: 215-228.	*Swartzia*	3	1403	81

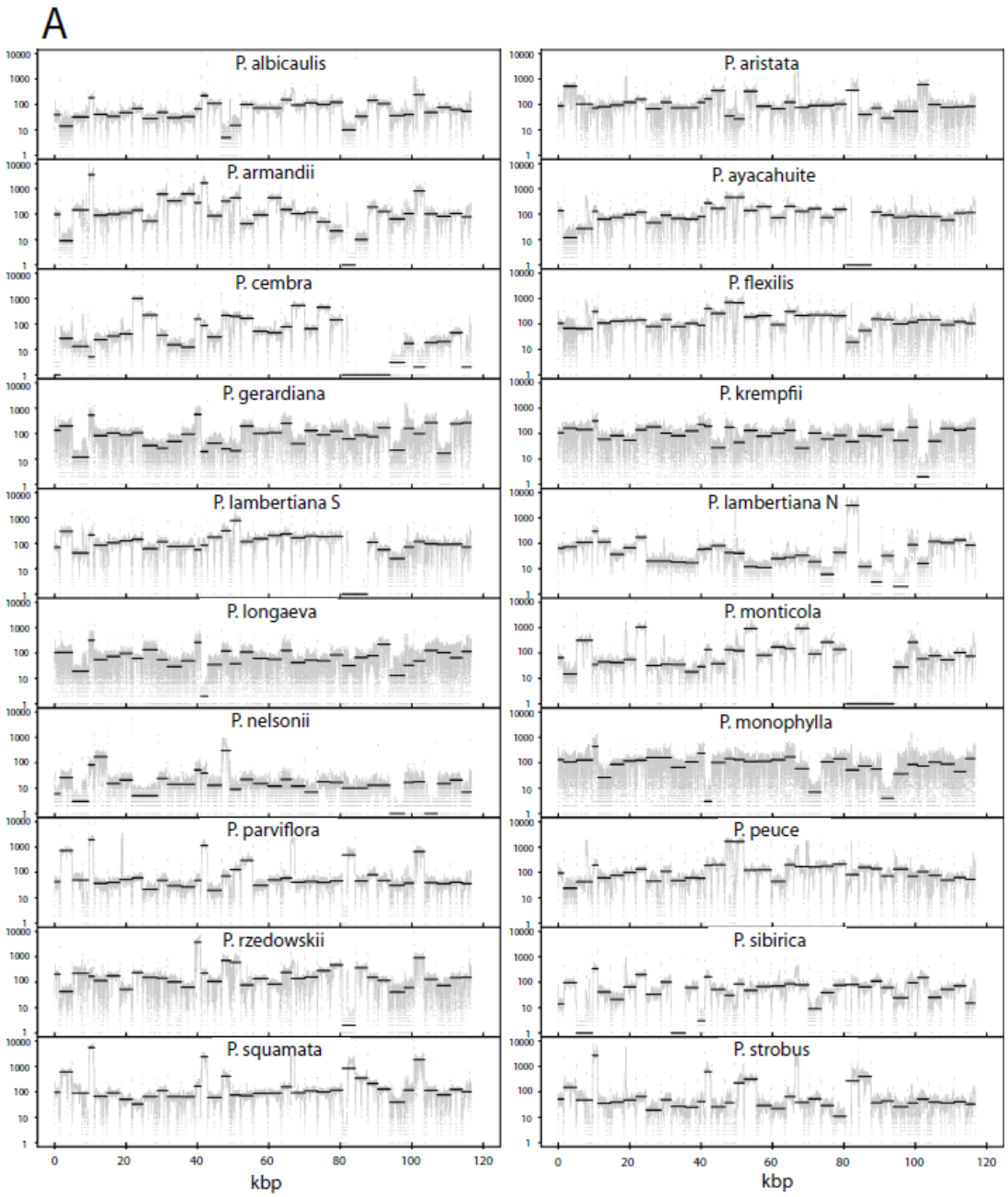
Appendix Table 2.1.

Trenel, P., Gustafsson, M.H.G., Baker, W.J., Asmussen-Lange, C.B., Dransfield, J. and Borchsenius, F. 2007 Mid-Tertiary dispersal, not Gondwanan vicariance explains distribution patterns in the wax palm subfamily (Ceroxyloideae: Arecaceae). Mol. Phylogenet. Evol. 45: 272-288.	*Ravenea*	3	3523	13
Tu, T., Dillon, M.O., Sun, H. and Wen, J. 2008 Phylogeny of Nolana (Solanaceae) of the Atacama and Peruvian deserts inferred from sequences of four plastid markers and the nuclear LEAFY second intron. Mol. Phylogenet. Evol. 49: 561-573.	*Nolana*	4	5196	64
Vorontsova, M.S., Hoffman, P., Maurin, O. and Chase, M.W. 2007 Molecular phylogenetics of tribe Poranthereae (Phyllanthaceae; Euphorbiaceae sensu lato). Am. J. Bot. 94: 2026-2040.	*Andrachne*	1	2137	17
Wagstaff, S.J., Breitwieser, L. and Swenson, U. 2006 Origin and relationships of the austral genus Abrotanella (Asteraceae) inferred from DNA sequences. Taxon 55: 95-106.	*Abrotonella*	2	1000	20
Wang, X.-R., Tsumura, Y., Yoshimaru, H., Nagasaka, K., and Szmidt, A.E. 1999 Phylogenetic relationshiips of Eurasian pines (Pinus, Pinaceae) based on chloroplast rbcL, matK, rpl20-rps18 spacer, and trnV intron sequences. Am. J. Bot. 86: 1742-1753.	*Pinus*	4	3570	32
Wanntorp, L., Kocyan, A., van Donkelaar, R. and Renner, S.S. 2006 Towards a monophyletic Hoya (Marsdenieae, Apocynaceae): inferences from the chloroplast trnL region and the rbcL-atpB spacer. Syst. Bot. 31: 586-596.	*Hoya*	2	1787	39
Weeks, A. and Simpson, B.B. 2007 Molecular phylogenetic analysis of Commiphora (Burseraceae) yields insight on the evolution and historical biogeography of an "impossible" genus. Mol. Phylogenet. Evol. 42: 62-79.	*Commiphora*	2	1569	36
	Bursera	2	1569	12
Weese, T.L. and Bohs, L. 2007 A three-gene phylogeny of the genus Solanum (Solanaceae). Syst. Bot. 32: 445-463.	*Solanum*	1	2277	102
Whipple, I.G., Barkworth, M.E. and Bushman, B.S. 2007 Molecular insights into the taxonomy of Glyceria (Poaceae: Meliceae) in North America. Am. J. Bot. 94: 551-557.	*Glyceria*	3	2019	16
Xiang, Q.-Y., Thomas, D.T., Zhang, W., Manchester, S.R. and Murrell, Z. 2006 Species level phylogeny of the genus Cornus (Cornaceae) based on molecular and morphological evidence - implications for taxonomy and Tertiary intercontinental migration. Taxon 55: 9-30.	*Cornus*	1	1548	55
Yi, T., Miller, A.J. and Wen, J. 2007 Phylogeny of Rhus (Anacardiaceae) based on sequences of nuclear Nia-i3 intron and chloroplast trnC-trnD. Syst. Bot. 32: 379-391.	*Rhus*	3	5573	22
Yuan, Y.-W. and Olmstead, R.G. 2008 A species-level phylogenetic study of the Verbena complex (Verbenaceae) indicates two independent intergeneric chloroplast transfers. Mol. Phylogenet. Evol. 48: 23-33.	*Glandularia*	6	4387	13
Zhu, Y.-P., Wen, J., Zhang, Z.-Y. and Chen, Z.-D. 2006 Evolutionary relationships and diversification of Stachyuraceae based on sequences of four chloroplast markers and the nuclear ribosomal ITS region. Taxon 55: 931-940.	*Stachyurus*	4	5562	11
Zuloaga, F.O., Giussani, L.M. and Morrone, O. 2006 On the taxonomic position of Panicum aristellum (Poaceae: Panicoideae: Paniceae). Syst. Bot. 31: 497-505.	*Panicum*	1	2061	19
Zuloaga, F.O., Giussani, L.M. and Morrone, O. 2007 Hopia, a new monotypic genus segregated from Panicum (Poaceae). Taxon 56: 145-156.	*Paspalum*	1	2061	11

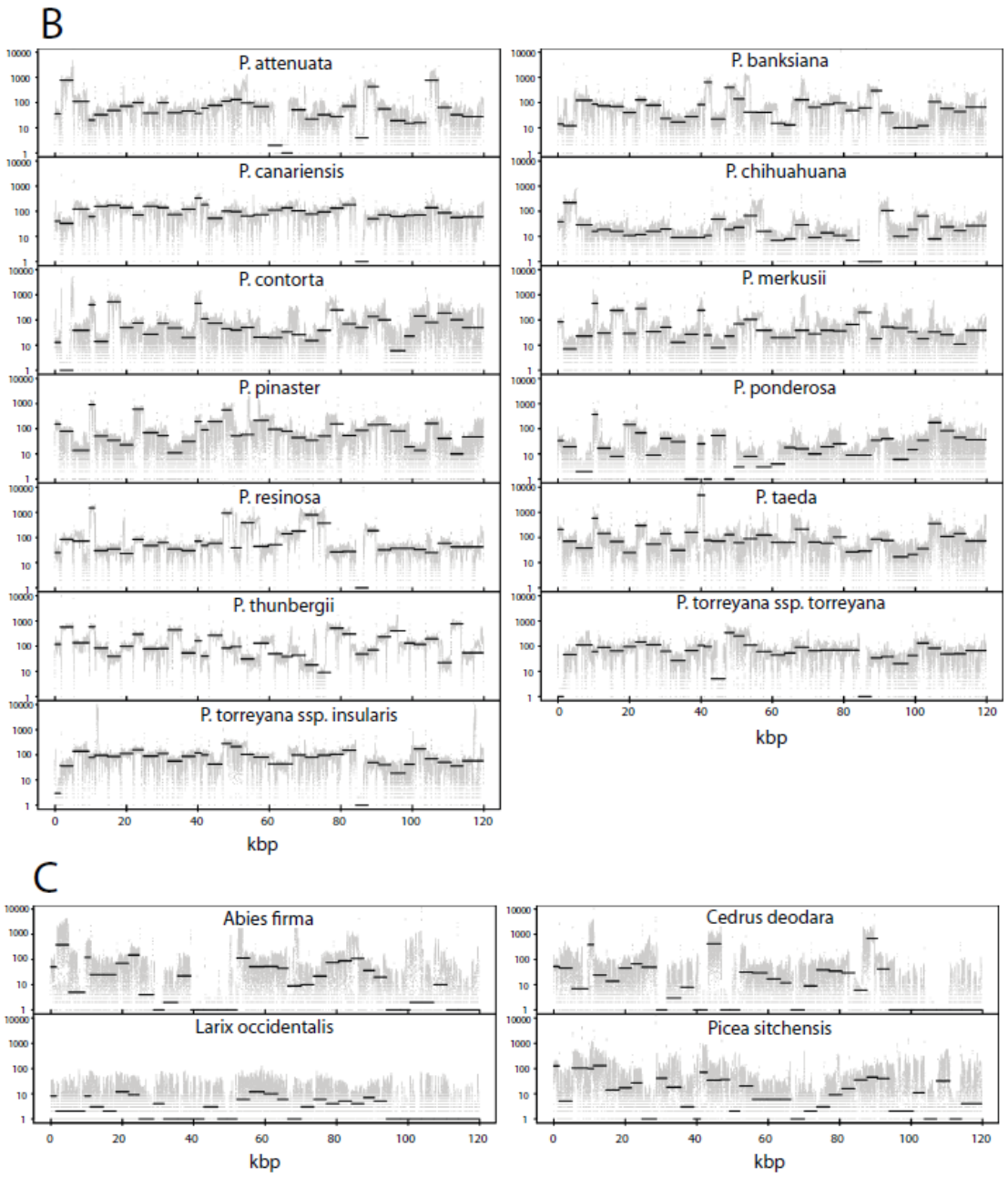
Appendix B

Chapter III Supplementary Figure

Appendix Figure 3.1. Amplicon coverage densities. A) Subgenus *Strobus*. B) Subgenus *Pinus*. C) Outgroups. Horizontal bars in charts indicate median coverage level for an amplicon.



Appendix Figure 3.1.



Appendix Figure 3.1.

Appendix C

Chapter IV Supplementary Table

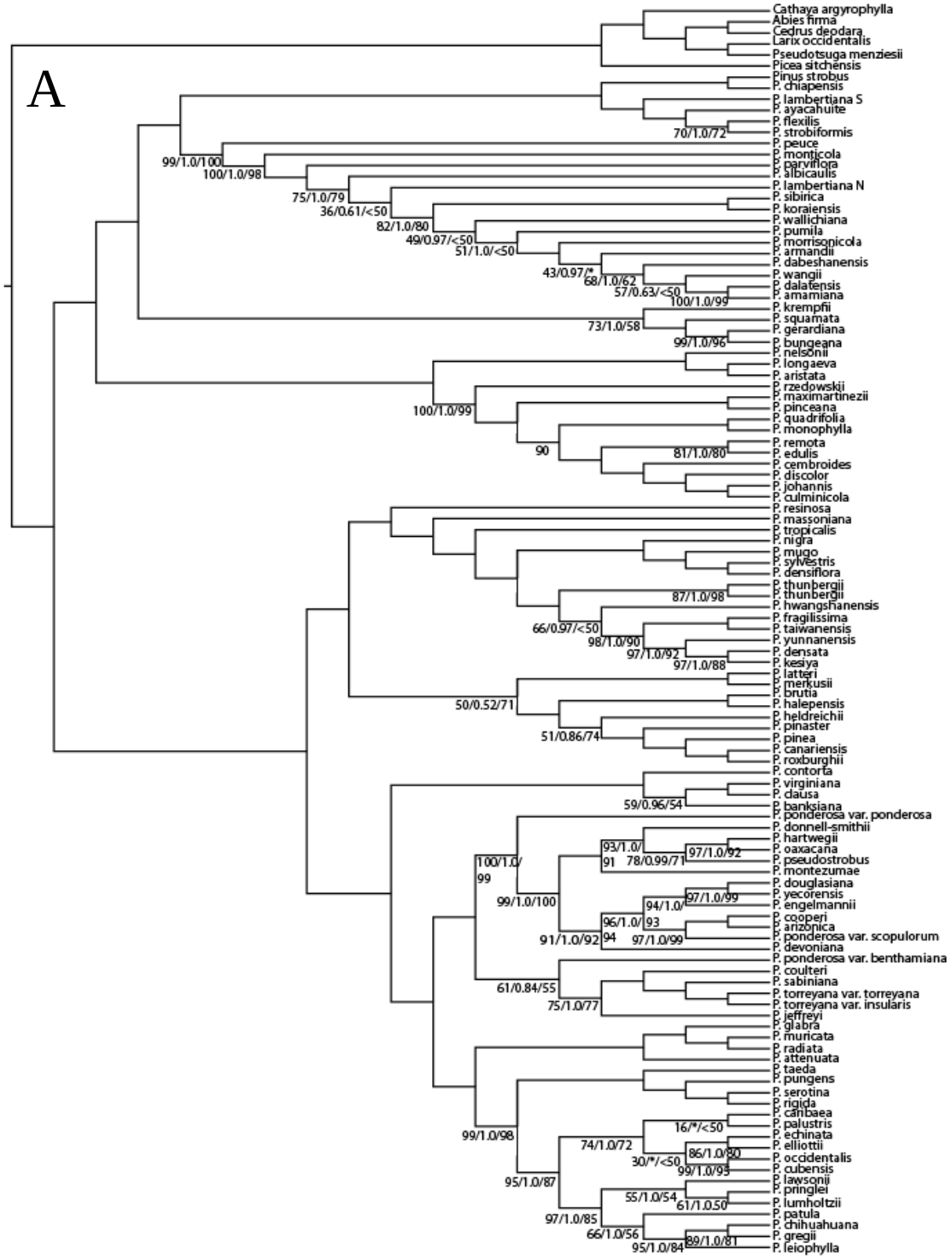
Appendix Table 4.1. Voucher information for species of *Pinus* used in *ycf*1 amplifications and sequencing.

Species	Collection information (Herbarium Location)	GenBank Number
Pinus aristata Engelmann	K. Farrell 30 (OSC213681)	FJ899567
Pinus canariensis C. Smith	A. Ensoll, M. Maquire, F. Nelson & A. Wright 1 (E)	FJ899572
Pinus contorta Douglas ex. Loudon	A. Liston 1315 (OSC218800)	EU998740
Pinus flexilis James	K. Farrell 28 (OSC213724)	FJ899576
Pinus gerardiana Wallich ex. D. Don	R. Businsky 41123 (RILOG)	EU998741
Pinus krempfii Lecomte	RBG Edinburgh First Darwin Expedition 242 (E)	EU998742
Pinus lambertiana Douglas	USFS Region 5 Camino Seed Orchard (OSC213763)	EU998743
Pinus monophylla Torrey & Fremont	D. Gernandt 479 (OSC213555)	EU998745
Pinus nelsonii Shaw	D. Gernandt 10198-15098 (OSC202126)	EU998746
Pinus pinaster Aiton	R. Businsky 36157a (RILOG)	FJ899583
Pinus ponderosa Douglas ex P. & C. Lawson	USFS R1 Geneticist (OSC213789)	FJ899555
Pinus taeda L.	Southern Institute of Forest Genetics, clone 7-56, LB-A10-05 (OSC226921)	FJ899561

Appendix D

Chapter V Supplementary Figure and Table

Appendix Figure 5.1. Phylogenetic relationships within genus *Pinus* as determined from full plastome alignment. A) Cladogram showing support values below branches as ML bootstrap support / Bayesian posterior probability / parsimony bootstrap support. Support values are shown only for nodes with less than 100% bootstrap support and/or posterior probabilities less than 1.0; single values indicate either ML bootstrap support or Bayesian posterior probability. * indicates branch not supported in analysis. B) Phylogram with branch lengths determined from maximum likelihood analysis. Scale corresponds to probability of change per position.



Appendix Figure 5.1

Appendix Figure 5.1

Appendix Table 5.1. Taxonomic and assembly information for novel accessions used in present study.

Appendix Table 5.1.

Species	Accession	Source	GenBank ID	# Reads in assembly pool (without adapters)	Estimated # chloroplast reads in pool (%)	# Contigs in Assembly	Avg. contig length (bp)	Estimated plastome length (bp) (% determined)	Reference used in assembly (GenBank ID)
Pinus amamiana	AMAM02	Kagoshima Pref., 30.324 N 130.4 E, Kyushu region, Japan; Businský	tbd	2475032	1610825 (65.08%)	12	9689.08	116329 (98.64%)	Pinus koraiensis (NC_004677.2)
P. arizonica	ARIZ01	House Creek, 33.133 N 108.0 W, New Mexico, United States; Gerald Rehfeldt	tbd	1744843	965660 (55.34%)	7	17246.57	120858 (98.45%)	P. ponderosa (FJ899555.2)
P. brutia	BRUT01	mountains of SE periphery of Ala Daglar Mts., 37.497 N 35.379 E, Adana, Turkey; Businský	tbd	1784995	920744 (51.58%)	14	8494.79	120279 (98.7%)	P. pinaster (FJ899583.2)
P. bungeana	BUNG05	Corvallis, Oregon, United States	tbd	1042349	471720 (45.26%)	11	10686.45	118277 (98.63%)	P. gerardiana (EU998741.4)
P. caribaea	CARI01	Marbajita, P. del Rio, 19.36 N 85.0297 W, Cuba	tbd	1013031	476897 (47.08%)	5	24159.40	120902 (99.59%)	P. taeda (FJ899561.2)
P. cembroides	CEMD04	Jeff Davis Co., near Fort Davis, Hwy17 to McDonald Observatory, 30.673 N 104.031 W, Texas, USA; David Gernandt	tbd	1557162	1074436 (69.00%)	11	10689.55	116847 (99.29%)	P. monophylla (EU998745.4)
P. chiapensis	CHIA35	Yerba Santa; Tree #1, Sierra Sur, 17.517 N 99.96 W, Guerrero, Mexico; John Syring, Rafael del Castillo	tbd	1235599	1103528 (89.31%)	14	8361.43	116405 (99.27%)	P. koraiensis (NC_004677.2)
P. clausa	CLAU02	Key Vista Park, Ballies Rd, Holiday, Pasco Co., 28.196 N 82.784 W, Florida, United States; Ann Willyard	tbd	2517778	1862252 (73.96%)	4	30080.50	120183 (99.18%)	P. contorta (EU998740.4)
P. cooperi	COOP01	El Salto, 23.8 N 105.4 W, Durango, Mexico; Institute of Forest Genetics	tbd	991639	629118 (63.44%)	10	11994.60	120844 (98.07%)	P. ponderosa (FJ899555.2)
P. coulteri	COUL03	California Seed Zone 997, elev. 1667, 33.5 N 116.5 W, California, United States; National Tree Seed Centre	tbd	1248611	655978 (52.54%)	8	15034.38	120945 (98.42%)	P. ponderosa (FJ899555.2)
P. cubensis	CUBE01	Estación Pinares de Mayarí,Sierra de Nipe, 20.490 N 75.800 W, Holguín province, Cuba; Tom Parchman	tbd	1133366	522829 (46.13%)	2	60338.00	120630 (99.38%)	P. taeda (FJ899561.2)
P. culminicola	CULM01	Cerro Potosi, Nuevo Leon, Mexico; David S. Gernandt	tbd	831130	621304 (74.75%)	12	9720.00	116676 (98.64%)	P. monophylla (EU998745.4)
P. dabeshanensis	DABE01	Yuexi Co., 30.997 N 116.075 E, Anhui - SW part at Hubei border, China; Businský	tbd	1775508	1134886 (63.92%)	84	1355.95	117562 (86.73%)	P. koraiensis (NC_004677.2)
P. dalatensis	DALA03	Ngoc Linh mt. region, 15.047 N 107.815 E, Gia Lai - Kon Tum Prov. (N corner), Vietnam; Businský	tbd	486850	333888 (68.58%)	53	2163.00	115247 (94.56%)	P. koraiensis (NC_004677.2)
P. densiflora	DENS02	Pyeongchang-gun; seedling grown at OSU, Kangwon, South Korea; Korea Forest Service Dept. of Genetic Resources	tbd	1951002	810348 (41.53%)	8	14991.63	119799 (99.71%)	P. thunbergii (NC_001631.1)

Appendix Table 5.1.

P. densata	DENT01	Diqing Prefecture, 27.884 N 99.617 E, NW Yunnan, China; Businský	tbd	745401	369067 (49.51%)	4	29979.25	119722 (99.25%)	P. thunbergii (NC_001631.1)
P. devoniana	DEVO02	Jujucato, Zirahoen, 19.42 N 101.82 W, Michoacan, Mexico; Jesus Vargas Hernandez	tbd	1753454	795317 (45.36%)	8	15128.50	120876 (99.13%)	P. ponderosa (FJ899555.2)
P. discolor	DISC02	Fort Huachuca, Cochise Co.; voucher from seedling grown at OSU, Arizona, USA; Frank Hammond	tbd	1626160	842453 (51.81%)	7	16619.00	116785 (99.1%)	P. monophylla (EU998745.4)
P. donnell-smithii	DONN02	South of Quetzaltenango, 14.8 N 91.517 W, Guatemala; Logan Sander	tbd	1999252	1313079 (65.68%)	11	10989.55	120676 (99.43%)	P. ponderosa (FJ899555.2)
P. douglasiana	DOUG01	Atenquique, 19.533 N 103.517 W, Jalisco, Mexico; Institute of Forest Genetics	tbd	4072566	1995919 (49.01%)	10	12080.40	121012 (99.11%)	P. ponderosa (FJ899555.2)
P. echinata	ECHI01	Cl.29, G.31, Bl.40B, Pt.24, For.280124, E. Ouachita Nat'l Forest; voucher from seedling grown at OSU, 34.7 N 93.8 W, Arkansas, United States; USDAFS Womble Ranger District	tbd	1201341	883067 (73.51%)	16	7544.63	120166 (99.03%)	P. taeda (FJ899561.2)
P. edulis	EDUL08	E of Zion NP, Kane Co., elev. 5710ft, 37.296 N 112.614 W, Utah, United States; David Gernandt	tbd	1526146	105095 (6.89%)	24	4839.71	116443 (96.03%)	P. monophylla (EU998745.4)
P. elliotti	ELLI01	Clone 22, Block 8, Row 9, Erambert Seed Orchard, Brooklyn; voucher from seedling grown at OSU, 33.173 N 90.486 W, Mississippi, United States; USDAFS Erambert Seed Orchard	tbd	1402469	848975 (60.53%)	6	20099.00	120481 (99.61%)	P. taeda (FJ899561.2)
P. engelmannii	ENGE02	Florida Canyon, 31.733 N 110.833 W, Arizona, United States; Gerald Rehfeldt	tbd	1723668	1036447 (60.13%)	5	24180.60	120908 (99.51%)	P. ponderosa (FJ899555.2)
P. fragilissima	FRAG01	below the great bend of Southern Cross-Island, 23.178 N 121.033 E, Taitung Co., Taiwan; Businský	tbd	1304942	821573 (62.96%)	7	17176.71	119704 (99.52%)	P. thunbergii (NC_001631.1)
P. glabra	GLAB01	Mississippi, United States; Southern Institute of Forest Genetics	tbd	966316	597201 (61.80%)	34	3529.06	120344 (96.84%)	P. taeda (FJ899561.2)
P. gregii	GREG02	Madrono, Queretaro, Mexico; Jesus Vargas Hernandez	tbd	2213548	1533046 (62.96%)	8	15123.13	120597 (99.69%)	P. taeda (FJ899561.2)
P. halepensis	HALE03	E slopes on right side of, 43.883 N 7.175 E, Alpes Maritimes Co. (dept. 06), France; Businský	tbd	1059136	325997 (30.78%)	15	7848.53	119860 (97.25%)	P. pinaster (FJ899583.2)
P. hartwegii	HART07	Yerba Santa, 17.52 N 99.96 W, Guerrero, Mexico; John Syring	tbd	1023159	472531 (46.18%)	10	12056.40	120907 (98.4%)	P. ponderosa (FJ899555.2)
P. heldreichii	HELD07	Pollino Mts., 39.899 N 16.126 E, Basilicata & Calabria border, Italy; Businský	tbd	947902	632442 (66.72%)	25	4769.64	120262 (95.96%)	P. pinaster (FJ899583.2)

Appendix Table 5.1.

P. hwangshanensis	HWAN01	1282-79C; Chin. Acad. Forestry, China	tbd	1446970	605318 (41.83%)	6	19979.00	119774 (99.86%)	P. thunbergii (NC_001631.1)
P. jeffreyi	JEFF04	Bishop, Inyo County, 37.366 N 118.390 W, California, United States;	tbd	1782057	1059552 (59.46%)	3	40192.67	120797 (99.62%)	P. ponderosa (FJ899555.2)
P. johannis	JOHA01	Cerro Pena Nevada, 23.817 N 99.883 W, Nuevo Leon, Mexico; RBG Edinburgh	tbd	6480263	3465568 (53.48%)	18	6613.44	116497 (98.73%)	P. monophylla (EU998745.4)
P. kesiya	KESI11	East Khasi Hills district, 25.635 N 91.900 E, Meghalaya, India; Businský	tbd	953853	578516 (60.65%)	10	12001.50	119748 (98.88%)	P. thunbergii (NC_001631.1)
P. latteri	LATT02	Nam Dan, Dai Hue Forest Enterprise, Nghe An, Vietnam; RBG Edinburgh	tbd	950283	425101 (44.73%)	5	24059.60	119917 (99.28%)	P. pinaster (FJ899583.2)
P. lawsonii	LAWS02	no collection data	tbd	2107814	986679 (46.81%)	3	40223.00	120393 (99.9%)	P. taeda (FJ899561.2)
P. leiophylla	LEIO03	north of Oaxaca city, east of Ixtlan, Oaxaca, Mexico; John Syring	tbd	1720365	1091164 (63.43%)	11	10991.18	120917 (99.21%)	P. taeda (FJ899561.2)
P. lumholtzii	LUMH07	Mexico; David Gernandt	tbd	1056895	685974 (64.90%)	9	13114.22	120442 (97.34%)	P. taeda (FJ899561.2)
P. massoniana	MASS02	Lang Son, Vietnam; RBG Edinburgh	tbd	1457463	999367 (68.57%)	4	29940.50	119634 (99.99%)	P. thunbergii (NC_001631.1)
P. maximartinezii	MAXZ01	Arbol 078, Juchipila, 21.417 N 103.117 W, Zacatecas, Mexico; Institute of Forest Genetics	tbd	2623261	1669960 (63.66%)	5	23371.60	116383 (99.67%)	P. monophylla (EU998745.4)
P. montezumae	MONZ01	Mpio. Epazoyucan, Real del Monte, 20.112 N 98.605 W, Hidalgo, Mexico; David Gernandt, UAEH	tbd	1709527	1145050 (66.98%)	13	9263.77	120787 (98.33%)	P. ponderosa (FJ899555.2)
P. morrisonicola	MORR01	Taiwan-bulk collection, 25.310 N 122.140 E, , Taiwan; USFS Dorena Genetic Resource Center	tbd	1430059	1203552 (84.16%)	14	8224.86	116579 (97.81%)	P. koraiensis (NC_004677.2)
P. mugo	MUGO01	Mount Beguinjstica, Draga, Slovenia; US National Arboretum	tbd	1110839	845197 (76.09%)	13	9164.69	119849 (98.54%)	P. thunbergii (NC_001631.1)
P. muricata	MURI01	Santa Cruz Island; voucher from seedling grown at OSU, 34.017 N 119.717 W, California, United States; Rancho Santa Ana Botanic Garden at Claremont	tbd	476779	184202 (38.63%)	13	9253.62	120796 (97.36%)	P. taeda (FJ899561.2)
P. nigra	NIGR20	Tahtali Daglari Mts., 38.134 N 36 0.111 E, Adana, Turkey; Businský	tbd	987325	445102 (45.08%)	8	14956.75	119806 (98.85%)	P. thunbergii (NC_001631.1)
P. oaxacana	OAXA02	Calpulalpan de Mendez, El Polvorin, 17.317 N 96.433 W, Oaxaca, Mexico; Gernandt	tbd	820251	455060 (55.48%)	10	12048.00	120771 (98.81%)	P. ponderosa (FJ899555.2)
P. occidentalis	OCCI02	Constanza, National Park In the Cordiiela Central, Dominican Republic; Tom Parchman	tbd	1159018	608453 (52.50%)	8	15102.88	120887 (99.06%)	P. taeda (FJ899561.2)
P. palustris	PALU02	Lake George Ranger District, Silver Springs, 29.2 N 82.05 W, Florida, United States; R8 USDAFS	tbd	1237697	674019 (54.46%)	3	40288.33	120397 (99.4%)	P. taeda (FJ899561.2)

Appendix Table 5.1.

P. patula	PATU01	Mpio. Mineral Real del Monte. Steep hillside at Penas Cargadas, 20.114 N 98.627 W, Hidalgo, Mexico; David Gernandt, UAEH	tbd	2233464	1383405 (61.94%)	7	17231.86	120384 (99.91%)	P. taeda (FJ899561.2)
P. pinceana	PINC13	Gran Cepeda, 25.317 N 101.667 W, Coahuila, Mexico; Institute of Forest Genetics	tbd	1790152	1109233 (61.96%)	29	4013.38	117040 (96.44%)	P. monophylla (EU998745.4)
P. pinea	PINE03	Pisa Prov., Toscana, Italy; Businský	tbd	1736009	953113 (54.90%)	9	13327.89	119861 (99.35%)	P. pinaster (FJ899583.2)
P. ponderosa var. benthamiana	POND21	Hwy 32 NE of Chico, Butte Co., 39.691 N 121.694 W, California, United States; David Gernandt	tbd	549958	263587 (47.93%)	11	10970.82	120901 (98.24%)	P. ponderosa (FJ899555.2)
P. ponderosa var. scopulorum	POND59	44.295 N 103.828 W, South Dakota, United States; Richard Halse	tbd	1422580	791205 (55.62%)	10	12055.10	120760 (99.06%)	P. ponderosa (FJ899555.2)
P. pringlei	PRIN02	Cuidad Hidalgo, Michoacan, Mexico; Jesus Vargas Hernandez	tbd	688261	342197 (49.72%)	18	6701.06	120831 (96.86%)	P. taeda (FJ899561.2)
P. pseudostrobus	PSEU03	Patio de Bolas, 15.38 N 91.43 W, Chiantla, Guatemala; ECODESA-FUNDAP	tbd	148338	67999 (45.84%)	109	1072.77	120487 (85.74%)	P. ponderosa (FJ899555.2)
P. pumila	PUMI07	Kawayu, Teshikaga, Hokkaido, Japan, Japan; Yasayuki Watano	tbd	1300700	1008815 (77.56%)	17	6761.35	116475 (96.84%)	P. koraiensis (NC_004677.2)
P. pungens	PUNG01	Penn State Stone Valley Experimental Forest, 40.671 N 77.899 W, Pennsylvania, United States; Pennsylvania State Universtiy	tbd	844752	471511 (55.82%)	2	60255.00	120427 (99.83%)	P. taeda (FJ899561.2)
P. quadrifolia	QUAD02	Pinyon Point, Cleveland National Forest, 32.871 N 116.418 W, California, United States; Kirsten Winter	tbd	1548293	1012929 (65.42%)	11	10614.36	116377 (99.28%)	P. monophylla (EU998745.4)
P. radiata	RADI02	Monterey County, 36.251 N 121.252 W, California, United States; David Gernandt, UAEH	tbd	786013	495765 (63.07%)	19	6349.58	120953 (97.78%)	P. taeda (FJ899561.2)
P. remota	REMO05	Val Verde Co., E of Hwy 277, 29 mi N of Del Rio, 29.746 N 100.816 W, Texas, United States	tbd	1261129	891919 (70.72%)	8	14584.38	116687 (98.98%)	P. monophylla (EU998745.4)
P. rigida	RIGI01	45 N 73.75 W, Quebec, Canada; Quebec Department of Forest Research	tbd	611781	301367 (49.26%)	15	7982.07	120845 (98.02%)	P. taeda (FJ899561.2)
P. roxburghii	ROXB04	Dirang region, 27.377 N 92.272 E, West Kameng district, India; Businský	tbd	1279512	748169 (58.47%)	3	39960.67	120050 (99.78%)	P. pinaster (FJ899583.2)
P. sabiniana	SABI04	across road from 9021 Placer Rd., Redding, Shasta Co., 40.547 N 122.460 W, California, United States; Ann Willyard	tbd	1721755	413242 (24.00%)	6	19881.17	120727 (98.17%)	P. ponderosa (FJ899555.2)
P. serotina	SERO01	Worcestor county, 38.25 N 75.55 W, Maryland, United States; Institute of Forest Genetics	tbd	1039759	555533 (53.43%)	15	8052.93	120880 (98.78%)	P. taeda (FJ899561.2)

Appendix Table 5.1.

P. strobiformis	STRF17	Jeff Davis Co., Mt. Livermore, 30.656 N 104.109 W, Texas, USA; David Gernandt	tbd	1236325	641302 (51.87%)	11	10670.36	116724 (99.2%)	P. koraiensis (NC_004677.2)
P. sylvestris	SYLV02	Cherkasy, Ukraine; US National Arboretum	tbd	889264	523094 (58.82%)	5	24027.20	119793 (99.67%)	P. thunbergii (NC_001631.1)
P. taiwanensis	TAIW04	SE slopes of the central island ridge, 23.267 N 120.961 E, Taitung Co.- NW corner, Taiwan; Businský	tbd	623175	436560 (70.05%)	23	5220.13	119706 (98.02%)	P. thunbergii (NC_001631.1)
P. tropicalis	TROP02	Ceja del Negro, Vinales, Cuba; Gretel Geada-Lopez	tbd	887229	552983 (62.33%)	21	5703.95	119665 (98.29%)	P. thunbergii (NC_001631.1)
P. virginiana	VIRG01	Monroe Co. Cherokee NF, Tellico Ranger District 13-G-37 #4, 35.833 N 83.0 W, Tennessee, United States; R8 Beech Creek Seed Orchard	tbd	1148954	627006 (54.57%)	56	2147.79	120771 (94.42%)	P. contorta (EU998740.4)
P. wallichiana	WALL02	Nanga Parbat mt. region, 35.2 N 74.667 E, Northern Areas (Gilgit), Pakistan; Businský	tbd	1379889	939139 (68.06%)	15	7760.27	116814 (98.25%)	P. koraiensis (NC_004677.2)
P. wangii	WANG05	no collection data	tbd	369269	97062 (26.28%)	41	2767.51	117166 (91.53%)	P. koraiensis (NC_004677.2)
P. yecorensis	YECO02	Yecora: 7.9 km E Yecora, 28.379 N 108.869 W, Sonora, Mexico; George Ferguson	tbd	2661988	2003250 (75.25%)	11	10401.82	120962 (93.66%)	P. ponderosa (FJ899555.2)
P. yunnanensis	YUNN01	Qiaojia Co., 26.883 N 103.013 E, NE Yunnan, China; Businský	tbd	1513538	895675 (59.18%)	30	3989.20	119703 (97.62%)	P. thunbergii (NC_001631.1)