

Pyrosequencing of the northern red oak (*Quercus rubra* L.) chloroplast genome reveals high quality polymorphisms for population management

Lisa W. Alexander · Keith E. Woeste

Received: 14 May 2013 / Revised: 31 October 2013 / Accepted: 12 November 2013 / Published online: 9 May 2014
© Springer-Verlag Berlin Heidelberg 2013

Abstract Given the low intraspecific chloroplast diversity detected in northern red oak (*Quercus rubra* L.), more powerful genetic tools are necessary to accurately characterize *Q. rubra* chloroplast diversity and structure. We report the sequencing, assembly, and annotation of the chloroplast genome of northern red oak via pyrosequencing and a combination of de novo and reference-guided assembly (RGA). Chloroplast DNA from 16 individuals was separated into four MID-tagged pools for a Genome Sequencer 20 quarter-run (Roche Life Sciences, Indianapolis, IN, USA). A four-step assembly method was used to generate the *Q. rubra* chloroplast consensus sequence: (1) reads were assembled de novo into contigs, (2) de novo contigs were aligned to a reference genome and merged to produce a consensus sequence, (3) the consensus sequence was aligned to the reference sequence and gaps between contigs were filled with reference sequence to generate a "pseudoreference", and (4) reads were mapped to the pseudoreference using RGA to generate the draft chloroplast genome. One hundred percent of the pseudoreference sequence was covered with a minimum coverage of 2× and an average coverage of 43.75×. The 161,304-bp *Q. rubra* chloroplast genome draft sequence contained 137 genes and one

rps19 pseudogene. The sequence was compared to that of *Quercus robur* and *Q. nigra* with 951 and 186 insertion/deletion or SNP polymorphisms detected, respectively. A total of 51 intraspecific polymorphisms were detected among four northern red oak individuals. The fully sequenced and annotated *Q. rubra* chloroplast genome containing locations of interspecific and intraspecific polymorphisms will be essential for studying population differentiation, phylogeography, and evolutionary history of this species as well as meeting management goals such as monitoring reintroduced populations, tracking wood products, and certifying seed lots and forests.

Keywords Forest genetics · SNP · Haplotype · Fagaceae

Introduction

Chloroplasts are organelles of prokaryotic origin located in the cytoplasm of most green plants and algae (Margulis 1970). These organelles present several features that make them uniquely useful for studies of population and evolutionary genetics. Chloroplasts contain a genome approximately 115–165 kb in size that codes for ~130 genes, comprising about 90 proteins (including essential proteins for photosynthesis), 30 tRNAs, and four rRNAs, making them sufficiently large and complex to include structural and point mutations (Raubeson and Jansen 2005). Strong purifying selection on photosynthesis-related genes imposes constraints on nucleotide and structural mutation rates, leading to high degree of conservation of chloroplast sequence content and order across, and especially within, taxa. In most angiosperms, chloroplasts are inherited maternally, while the mode of inheritance in gymnosperms is mostly paternal (Petit et al. 2005). The haploid state and uniparental transmission of chloroplasts gives chloroplast genes and genomes an effective population size of one-fourth of a nuclear locus (Birky 1978; Birky et al. 1983),

Communicated by A. Kremer

Electronic supplementary material The online version of this article (doi:10.1007/s11295-013-0681-1) contains supplementary material, which is available to authorized users.

L. W. Alexander (✉)
Department of Plant Sciences, University of Tennessee, 2431 Joe Johnson Drive, Knoxville, TN 37996, USA
e-mail: lworthen@utk.edu

K. E. Woeste
USDA Forest Service Hardwood Tree Improvement and Regeneration Center, Department of Forestry and Natural Resources, Purdue University, 715 West State Street, West Lafayette, IN 47907, USA

making them more susceptible to stochastic processes such as drift and founder events than nuclear genes and genomes. As a result, chloroplast genes display much higher genetic differentiation values as compared to differentiation at nuclear loci for the same taxa. A review of the comparative organization of organelle and nuclear diversity in plants covering 152 studies and 144 species reported an average G_{ST} of 0.637 ± 0.002 for maternally inherited genes (i.e., chloroplast genes) versus an average G_{ST} of 0.184 ± 0.002 for biparentally inherited (nuclear) genes in angiosperms (Petit et al. 2005). This contrasting pattern of nuclear and chloroplast differentiation is a property that has commonly been exploited to test hypotheses of seed and pollen dispersal, migration/colonization routes, intraspecific differentiation, and interspecific introgression (Petit 2004; Petit et al. 2001, 2004, 2005; Magni et al. 2005; Tovar-Sanchez et al. 2008).

Oaks are angiosperm trees in the family Fagaceae, in the genus *Quercus*, which consists of over 500 species worldwide. *Quercus* contains three major groups: sections *Lobatae*, *Protobalanus*, and *Cerris*; all evolved at middle latitudes of the Americas (Nixon 1993), but only *Lobatae* are unambiguously monophyletic (Manos et al. 1999). The *Lobatae*, or red oak group, contains about 195 species, with as many as 19 species sympatric with northern red oak (*Quercus rubra* L.). The taxonomy of oak species in general and red oaks in particular presents a considerable challenge, as closely related, interfertile oak species are not well differentiated (Manos et al. 1999; Moran et al. 2012), and even more distantly related congeners may share traits associated with habitat specialization due to convergent evolution (Cavender-Bares et al. 2004). For the same reasons, however, oak species are good models for studies of speciation, population genetics and gene flow, as they are primarily outcrossing, often have a wide geographical range, and they hybridize freely with congeners. Although ecophysiological studies indicated that sympatric red oaks may specialize into niches associated with soil moisture or fire regimes, sharing the landscape while maintaining diversity (Cavender-Bares and Holbrook 2001), molecular studies of communities occupied by multiple oak species reveal a complex relationship between hybridization and adaptation (Moran et al. 2012; Lind and Gailing 2013).

Northern red oak is an economically and ecologically important forest tree of North America. It is a major dominant hardwood species with a wide native range extending east from Nebraska to the Atlantic coast (60° to 96° W longitude) and from northern Ontario to southern Alabama (32° to 47° N latitude). Northern red oak has adapted to wide ranges of mean annual temperature and rainfall. Often it is the dominant oak species on lower slopes and north-facing slopes of the eastern deciduous forest (Abrams 2002), and it usually co-occurs with several other close congeners. Intraspecific variation for adaptive traits such as height growth, phenology, drought resistance, and cold-hardiness has been reported on altitudinal,

latitudinal, and longitudinal clines across the species' range (Gall and Taft 1973; McGee 1974; Schlarbaum and Bagley 1981; Kriebel 1993). Compared to other oak species, however, the neutral-marker population differentiation (G_{ST}) and chloroplast haplotype diversity of northern red oak is remarkably low (Sork et al. 1993; Hokanson et al. 1993; Magni et al. 2005; Feng et al. 2008). White oaks from the eastern United States had a mean chloroplast G_{ST} value of 0.87 ± 0.07 (Whittemore and Schaal 1991) while six species of European white oaks exhibited coefficients of chloroplast differentiation ranging from 0.78 in *Q. robur* to 0.96 in *Q. pyrenaica* (Petit et al. 2002). Magni et al. (2005) found the value of cpDNA differentiation in *Q. rubra* ($G_{ST}=0.46$) was lower than the mean G_{ST} value of 22 European forest trees and shrubs ($G_{ST}=0.54$, Petit et al. 2005) and the overall mean of angiosperms ($G_{ST}=0.76$, Dumolin-Lapegue et al. 1999), though Tovar-Sanchez et al. (2008) reported an R_{ST} (similar to G_{ST}) value of 0.398 for *Quercus crassipes* in an introgression zone.

As in most angiosperms, oak chloroplasts are inherited maternally (Dumolin et al. 1995; Romero-Severson et al. 2003). Chloroplast intergenic markers have been used extensively to examine genetic differentiation of European white oak populations (*Quercus* subgenus *Quercus*, Petit et al. 2002), to map the northward colonization of white oaks in Europe following the last glacial maximum approximately 21,000–18,000 years ago (Kremer et al. 2010; Petit et al. 2001, 2004), and to elucidate the invasion-by-hybridization model of gene flow in the *Quercus robur*–*Quercus petraea* forests of southwestern Europe (Petit 2004).

Q. rubra chloroplasts show strikingly low differentiation of among populations, and low haplotype diversity. This has been attributed to short post-glacial migration routes and a particularly high seed-mediated gene flow (Magni et al. 2005; Birchenko et al. 2009). Whatever the historical factors causing low haplotype diversity in red oak, detection of statistically robust population differentiation requires larger amounts of chloroplast DNA sequence - both in terms of number of individuals sampled and number of bases sequenced - than nuclear DNA sequence, (Small et al. 1998). Large amounts of sequence have traditionally been generated by cloning and Sanger sequencing or primer walking approaches (Jansen et al. 2005; Petit and Vendramin 2006), which can be laborious and time-consuming. Recently introduced second-generation sequencing technologies provide the possibility of acquiring entire genomes — or many simple genomes — at a fraction of the cost and time of traditional approaches. Multiplex tagging methods have the potential to spread the capacity of high-capacity sequencers across many genomes and strike an acceptable balance between coverage, throughput, and cost (Cronn et al. 2008). Second-generation sequencing platforms have proven useful in sequencing chloroplast genomes (Moore et al. 2006, 2007; Cronn et al. 2012; Parks et al. 2010; Straub et al. 2012). As a consequence, there has been a rapid increase in the number of chloroplast

genomes added to public databases in recent years (Ravi et al. 2008; Parks et al. 2012).

Given that the intraspecific chloroplast diversity detected in northern red oak is low, more powerful tools are necessary to more accurately characterize *Q. rubra* chloroplast diversity. A fully sequenced and annotated northern red oak chloroplast genome would serve as a reference for the construction of chloroplast genomes of other red oak species and provide a reference for comparative genomics within the Fagaceae. Detection of intra- and interspecific polymorphisms is essential for studying population differentiation, phylogeography, and evolutionary history of North American oak species. Tools for detecting intraspecific variation are also needed for management goals such as monitoring reintroduced populations, tracking wood products, and certifying seed lots and forests (Deguilloux et al. 2003). To this end, we report the sequencing, assembly, and annotation of the chloroplast genome of northern red oak. Our objectives were to 1) annotate and characterize the genome including the location and description of conserved regions, genes, and pseudogenes, and 2) to locate polymorphisms between *Quercus* species and within northern red oak.

Materials and methods

Branches bearing fresh leaves were collected from 17 trees in the Cherokee National Forest Watauga northern red oak seed orchard in Unicoi County, TN (36.33°N, 82.03°W; Table 1). Branches were stored in the dark at 4 °C for 7–10 days in order for starches to break down. Chloroplast isolation and cpDNA extraction were performed following the methods of Michaud et al. (1995) for chloroplast isolation in broad-leaved tree species. The method has two major components: the isolation of chloroplasts in non-aqueous solutions to avoid the activity of oxidative and phenolic compounds, and the chloroform/isoamylalcohol extraction of chloroplast DNA. Briefly, young leaves were ground in liquid nitrogen and freeze dried in vacuo. Dry leaf powder was blended with *n*-hexane/carbon tetrachloride and the homogenate filtered through nylon cloth. The filtrate was overlaid with *n*-hexane/CCl₄ and centrifuged. The chloroplast fraction, visible as a dark green band, was extracted, washed twice by pelleting in 12 ml of *n*-hexane/CCl₄ mixtures of $\rho = 1.25$ and 1.12, successively, pelleted in *n*-hexane, and desiccated until use. For cpDNA extraction, chloroplast pellets were resuspended in a lysing solution and centrifuged. Supernatant from the lysate was incubated with a pronase and CTAB buffer and the mixture was extracted two or three times with chloroform/isoamylalcohol. Nucleic acids were precipitated from the aqueous phase of the final extraction with 2-propanol, washed with cold ethanol, pelleted by centrifugation, and stored at –20°C until use.

Genome Sequencer 20 (Roche Life Sciences) library construction and sequencing were performed as described in

Table 1 Identification and origin information for 17 *Quercus rubra* ramets in the Watauga northern red oak clonal seed orchard (Watauga, TN, USA) that contributed leaf samples for chloroplast isolation

Parent number	Origin				Ramet position
	County, ST	Estimated elevation (m)	Latitude, N	Longitude, W	
323	Trigg, KY	152	36°50"	88°00"	1, 12
323 ^a	Trigg, KY ^b	152	36°50"	88°00"	2, 20
555 ^a	Morgan, TN	610	36°08"	84°28"	1, 3
580	Claiborne, TN	366	36°20"	83°35"	1, 1
629	Campbell, TN	366	36°18"	84°01"	1, 14
629	Campbell, TN	366	36°18"	84°01"	3, 5
720	Campbell, TN	305	36°16"	84°10"	6, 15
856	Overton, TN	305	36°20"	85°25"	4, 25
877	Madison, AL	457	34°45"	86°30"	4, 3
889	Madison, AL	457	34°45"	86°30"	1, 13
889	Madison, AL	457	34°45"	86°30"	5, 3
899 ^a	Washington, VA	640	36°42"	81°58"	3, 1
899	Washington, VA	640	36°42"	81°58"	4, 21
905	Henderson, TN	152	35°43"	88°17"	1, 1
913	Henderson, TN	152	35°43"	88°17"	1, 26
914 ^a	Henderson, TN	152	35°43"	88°17"	3, 26
1164	Monroe, TN	610	35°20"	84°15"	4, 23

Ramet position is expressed as block, row

^a Isolated chloroplasts used in genome sequencing

^b Land Between the Lakes State Park

Marguiles et al. (2005) with slight modifications as specified by Roche Life Sciences. High molecular weight DNA from the chloroplast DNA extraction was sheared by nebulization into to a size range of 300–800 bp. DNA fragment ends were repaired and phosphorylated using T4 DNA polymerase and T4 polynucleotide kinase. Adaptor oligonucleotides "A" and "B" supplied with the 454 Life Sciences sequencing reagent kit were ligated to the DNA fragments using T4 DNA ligase. Purified DNA fragments were hybridized to DNA capture beads and amplified by emulsion PCR (emPCR). DNA capture beads containing amplified DNA were deposited onto a 40 x 75 mm PicoTiterPlate equipped with an eight lane gasket. Four pools containing chloroplast DNA were assigned two lanes on a single plate (quarter-run). Each pool received a unique MID adapter (Roche Life Sciences) for use in identifying sequences within each pool in downstream applications.

Read trimming, read assembly, reference mapping, and sequence analysis was performed in CLC Bio Genomics Workbench (Finlandsgade, Denmark) and Newbler sequence assembly software. Read trimming, read assembly, and reference mapping were performed using manufacturer's settings except that long reads reference mapping similarity was increased from 0.8 to 0.9. A modified four-step assembly

method following Cronn et al. (2008) was used to generate the *Q. rubra* chloroplast consensus sequence. First, reads were assembled de novo into contigs. Second, reads were mapped against water oak (*Quercus nigra* L. section Lobatae, Moore et al. 2010) and pedunculate oak (*Quercus robur* L. section Quercus, Kremer et al. 2012). Third, contigs from de novo assembly were mapped to the *Q. nigra*-based consensus sequence. The *Q. nigra* genome was chosen as the reference as *Q. nigra* is more closely related to *Q. rubra* than is *Q. robur*; thus, more de novo contigs from *Q. rubra* mapped to the *Q. nigra* genome than the *Q. robur* genome. The new contigs were aligned to the *Q. nigra*-based consensus sequence, and gaps between contigs were filled with the *Q. nigra*-based consensus sequence to generate the "pseudoreference". Fourth, a final reference-guided assembly (RGA) was performed using the pseudoreference and the trimmed reads. This iterative method (also used by Parks et al. 2010 and Straub et al. 2012) leads to more accurate assemblies than RGA alone when the reference and target are not the same species (Straub et al. 2012). To the author's knowledge, this is the first use of this assembly method using 454 short-read data.

CLC Genomics Workbench employs the neighborhood quality standard (NQS) model for SNP calling (Altshuler et al. 2000) which incorporates the quality of the central base (SNP) and the quality of bases in a neighborhood, or window (here, 11 bp). SNP and deletion/insertion polymorphism (DIP) detection was first carried out with parameters set to manufacturer's settings, and carried out a second time with minimum detection coverage increased from 4 to 6, the minimum average quality of surrounding bases increased from 20 to 30, and minimum quality of the central base increased from 30 to 40. The minimum variant frequency for SNP and DIP detection was 35 %, i.e., a SNP or DIP was accepted when the minority base at any polymorphic position was present at a frequency of at least 35 %. Primers were designed by visual examination of the SNP and surrounding sequence using the following guidelines: at least 16 bp in length, GC content of 40 % with no more than 1 % GC difference in the forward and reverse strands, melting temperature between 52 °C and 58 °C with no more than 2 °C difference in melting temperatures between the forward and reverse strands, and at least 40 bp between the SNP and the primers.

Results

The GS 20 quarter-run produced a total of 331,364 reads comprising 115,735,252 bases. MID multiplex tagging allowed the sequencing of 16 individuals in four pools; however, three of the four pools had significant (>95 %) contamination from non-target DNA and were not included in downstream analysis. The remaining pool of four genotypes

produced a total of 159,338 reads comprising 54,304,612 bases. The average read length was 365.2 bp. Reads were trimmed for length, quality, and adapters according to manufacturer's settings; after trimming, 156,024 reads remained for de novo assembly and reference mapping with an average read length of 348.1 bp. Of the trimmed reads, 54,921 assembled into 10,441 de novo contigs with a mean length of 538 bp, comprising a total of 5,628,135 bases. 19,348 reads comprising 7,065,243 bases mapped to the *Q. nigra* reference, more reads than any other reference species. Removal of the inverted repeat region "B" (IR_B) from the *Q. nigra*-based reference allowed alignment of de novo contigs to the *Q. nigra*-based reference. Thirty-nine contigs with a mean length of 9,412±4,236 bp and an average coverage of 19.2±27.0 aligned to the *Q. nigra*-based reference. Contigs were merged into two large "supercontigs": the first covered reference bases 1–36,000 and the other bases 36,011 to 135,233 (i.e., to the end of the reference SSC region). The missing area of 11 bases in the LSC was filled in with the *Q. nigra*-based reference. No areas of conflict were found between contigs and the reference sequence within IR_A; however, six areas of misalignment or incorrect base calls were found in the small single copy (SSC) and large single copy (LSC) regions, with all six related to sequence gaps of >10 bp in the contigs relative to the reference. In all six cases, "extra" sequence contained within the *Q. nigra*-based reference was removed and the de novo contigs joined to create the 161,306-bp pseudoreference sequence. RGA based on the pseudoreference and the 156,024 trimmed reads from MID pool 4 yielded the *Q. rubra* chloroplast consensus sequence. One hundred percent of the pseudoreference sequence was covered with a minimum coverage of 2 and an average coverage of 43.75 (Table 2, Online Resource 1).

The 161,304-bp *Q. rubra* chloroplast genome consensus sequence (GenBank JX970937) was annotated using DOGMA (Wyman et al. 2004), an online database for searching and annotating chloroplast genomes. During

Table 2 Reference mapping coverage statistics for a *Quercus nigra*-based reference-guided assembly of the *Q. rubra* chloroplast genome via the Genome Sequencer 20 pyrosequencing platform

Total reference length	161,306
Total read count	19,475
Mean read length	364.91
Total read length	7,106,590
% GC	36.79
Total consensus length	161,304
Fraction of reference covered	1
Minimum coverage	2
Maximum coverage	118
Average coverage	43.75
Standard deviation	18.89

annotation, each putative coding sequence (hereafter, gene) was visually inspected, start and stop codons were manually selected, and putative exons were examined and joined. Annotation results confirmed by BLAST searches of *Q. rubra* chloroplast genes against other Viridiplantae plastid genomes revealed 11 major classes of proteins coded by 138 genes in the *Q. rubra* chloroplast: ribosomal RNA (eight genes), transfer RNA (41), photosystem I (five), photosystem II (15), cytochrome b6/f (six), ATP synthase (seven), NADH dehydrogenase (12), ribosomal small and large subunits (13 and 11), RNA polymerase (four), hypothetical open reading frames (eight), and eight miscellaneous proteins (Fig. 1, Online Resource 2). Ribosomal small subunit gene *rps19* contained the IRB/LSC border such that its duplicate at the IRA/LSC border appears truncated into a pseudogene. The *Q. rubra* chloroplast genome LSC and SSC contain 91,121 and 20,150 bp, respectively, while the IR is 25051 bp in length. Almost 60 % (57.4) of the *Q. rubra* chloroplast genome is comprised of coding sequence (Table 3).

SNPs were detected using standard and high quality (HQ) settings and each SNP and DIP was visually examined for accuracy. A total of 737 (99 HQ) SNPs and 214 DIPs were detected between *Quercus robur* and *Q. rubra*; 23 (5 HQ) SNPs and 163 DIPs were detected between *Q. nigra* and *Q. rubra*; and eight (6 HQ) SNP and 45 DIPs were detected within the genomes of the four sequenced *Q. rubra* individuals. All polymorphisms were in the LSC and SSC; no polymorphisms were found in the inverted repeat regions (Fig. 2). Of the 23 SNPs detected between *Q. nigra* and *Q. rubra*, nine were located within genes. Of the six HQ SNPs detected within *Q. rubra*, three were located in the coding region and three in the noncoding regions; ten of 45 DIPs were found in coding regions. *Q. rubra* chloroplast genes *rpoC2* and *ycf1* showed the most variability. Including monomers, 233 repeat regions of 9 to 108 bp were detected in the chloroplast genome (Table 4).

Discussion

The most common method of isolating plant plastid DNA involves sucrose-based chloroplast isolation followed by rolling-circle amplification (RCA). It is significantly less expensive than BAC library construction or primer walking (Cronn et al. 2008). However, like our method, the RCA-based method leads to non-chloroplast DNA contamination, with approximately 10–40 % of the resulting product being non-target (mostly nuclear) DNA. This contamination penalty must be overcome in Sanger-based sequencing by increasing sequencing capacity, thereby partially mitigating the significant savings that could come with reducing sequence coverage.

The same contamination reduces plastid genome coverage in GS 20 sequencing runs, but it does not impede the recovery of essentially complete plastid genomes of high accuracy, as demonstrated by the chloroplast genome sequencing of *Platanus occidentalis* and *Nandina domestica* (Moore et al. 2006). Moore et al. (2006) also noted that although they observed a higher percentage of non-cpDNA contamination in the RCA product of *Platanus* (44 %) versus *Nandina* (18 %), and although that led to a lower average coverage of *Platanus*, it did not lead to a higher sequencing error rate in *Platanus* compared to *Nandina*. While the contamination rate in our sequencing project was much higher (>80 %), we recovered the essentially complete chloroplast genome at an average coverage of 43 \times , and found high-quality structural and point polymorphisms for further analysis.

In theory, multiplex tags would have allowed us to distribute the capacity of the pyrosequencing platform across 16 genomes and strike a favorable balance between throughput, coverage, and cost, and include even more individuals in the discovery of polymorphisms. High levels of non-target DNA contamination (primarily nuclear) constrained our sequencing to one pool of four individuals. The method we used to isolate and lyse whole chloroplasts for cpDNA extraction relies on the separation of organelles in non-aqueous solutions of differing specific gravities. Leaves with low chloroplast yields need solutions of higher specific gravity than high-yielding leaves to obtain a given amount of chloroplasts; however, solutions of higher specific gravity (e.g., $\rho = 1.33$ used here) attract more nuclear DNA (Bowman and Dyer 1982). New, partially expanded leaves would likely have a higher concentration of chloroplasts, although Michaud et al. (1995) noted that cork oak (*Quercus suber*) has leaf chloroplast concentrations low enough to make this method unsuitable at any stage of leaf growth. Another method, such as chloroplast primer walking, may be a more efficient method for isolating and sequencing *Quercus* cpDNA, though primer walking does not guarantee complete chloroplast coverage (Ferris et al. 1993).

Recently published methods for chloroplast genome sequencing such as "genome skimming" from shotgun sequencing data (Straub et al. 2012) and targeted enrichment of chloroplast sequence, e.g., by hybridization (Cronn et al. 2012) overcome contamination pitfalls by eliminating non-target DNA before sequencing. For example, Straub et al. (2012) reported the mean percentage of cpDNA sequence in genomic libraries of Apocynaceae species to be 11.4 %. In the present study, 19,348 of 156,024 reads (about 12 %) mapped to the chloroplast. Thus, the method we used did not enrich chloroplast sequence beyond what is likely to be present in a genomic library. Even though our plastid prep was contaminated with high levels of nuclear DNA, the 43 \times coverage depth we obtained was more than the 30 \times coverage recommended for plastome assembly (Straub et al. 2012). Chloroplast genomes are being added to public databases at

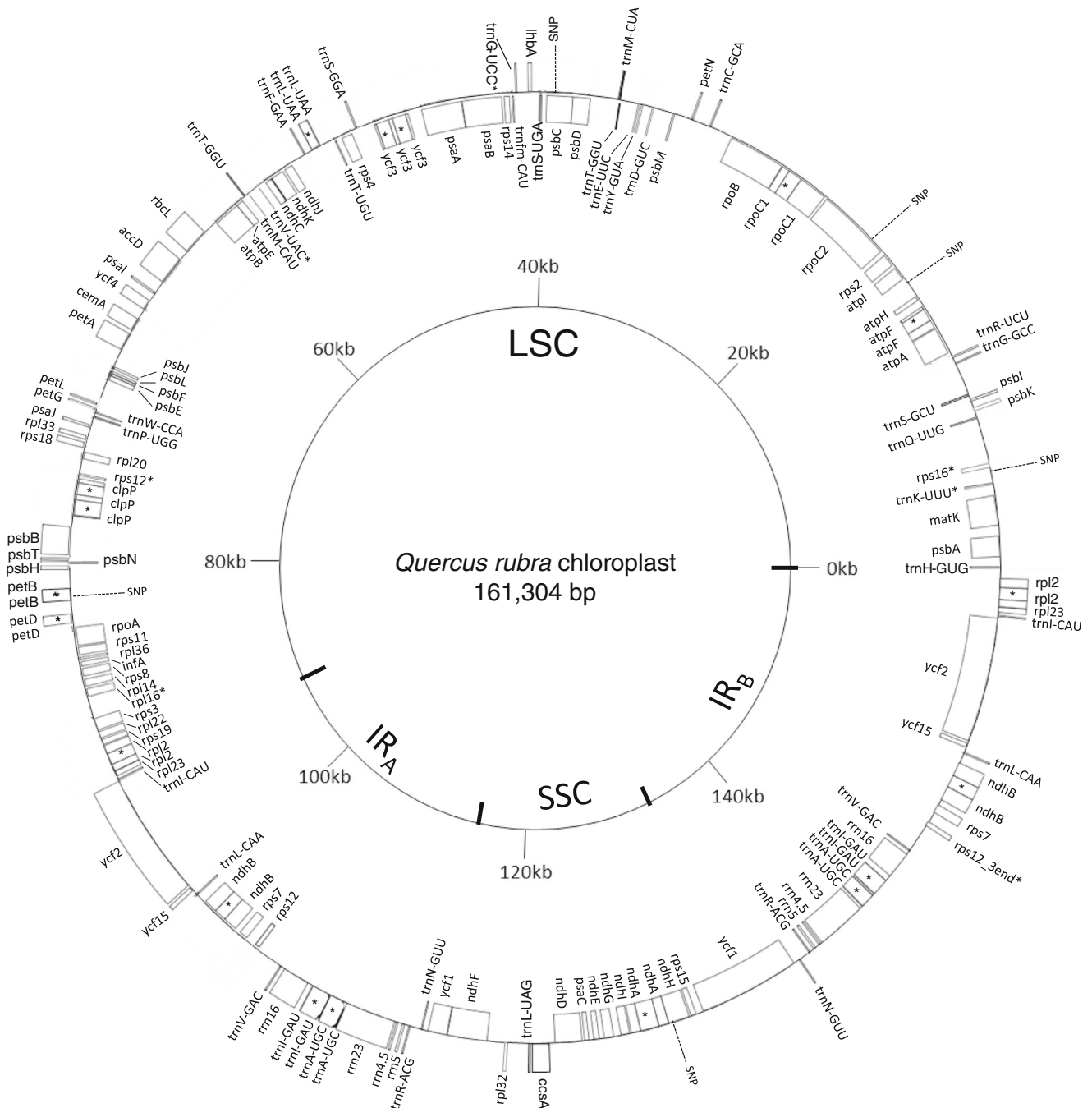


Fig. 1 Circular representation of the *Quercus rubra* chloroplast genome. Black bars in the inner circle denote boundaries between the large single copy (LSC), small single copy (SSC) and inverted repeat (IR) regions. Genes are denoted by boxes scaled to size; boxes outside and inside the circle represent genes on the plus and minus strands, respectively. Introns

are represented by asterisks (*) either within the gene or next to the gene name. Distribution of six single nucleotide polymorphisms (SNPs) within the chloroplast genomes of four chloroplast *Q. rubra* individuals detected via reference-guided assembly of chloroplast GS 20 reads are shown with dotted lines

a rapid rate (Ravi et al. 2008; Straub et al. 2012). As of January 2013, there were 311 complete chloroplast genomes published in the NCBI organelle genomes database; 19 of those were managed forest tree species (<http://www.ncbi.nlm.nih.gov/genomes>). Many partially complete chloroplast genomes are also available; for example, over

75 *Pinus* cp genomes were assembled by Parks et al. (2012) and estimated to be >98 % complete. These sequences are housed in the NCBI nucleotide database, making it difficult to accurately assess the number of functionally complete chloroplast genomes currently available.

Table 3 Basic characteristics of the *Quercus rubra* chloroplast genome

<i>Quercus rubra</i> chloroplast genome	
Total genome length	161,304
IR length	25,869
SSC length	19,025
LSC length	90,542
Total length (%) of coding sequence	92,558 (57.4)
Total length (%) of noncoding sequence	68,746 (42.6)

Lengths are given in base pairs (bp). Total length of coding and non-coding sequence include both IRs

IR inverted repeat region, SSC small single-copy region, LSC large single-copy region

Q. nigra, water oak, is native to the USA, a congener with *Q. rubra*, and the two species often co-occur and freely hybridize, making it the closest related reference sequence (Solomon 1983). After Cronn et al. (2008) we mapped trimmed GS 20 reads to *Q. nigra* then mapped de novo contigs back to the *Q. nigra*-based consensus to create a

Table 4 Repeat motifs in the *Quercus rubra* chloroplast genome

Number of repeats	No. bases in repeat	Tandem size range
122	1	9–15
20	2	10–16
77	3	9–12
9	4	12–16
4	5	15
1	6	18

pseudoreference sequence. Eleven base positions (36,001–36,011) were filled in using the *Q. nigra*-based reference, while 87 base positions were deleted by manually joining contigs rather than using reference sequence. Mapping to this pseudoreference increased the number of reads used, increased minimum coverage from 0 to 2, and decreased the number of zero-coverage areas from 174 to 0, relative to mapping to the *Q. nigra* reference. This method is better than de novo or RGA alone when an exact reference is not available, especially in difficult repeat regions. Comparison of

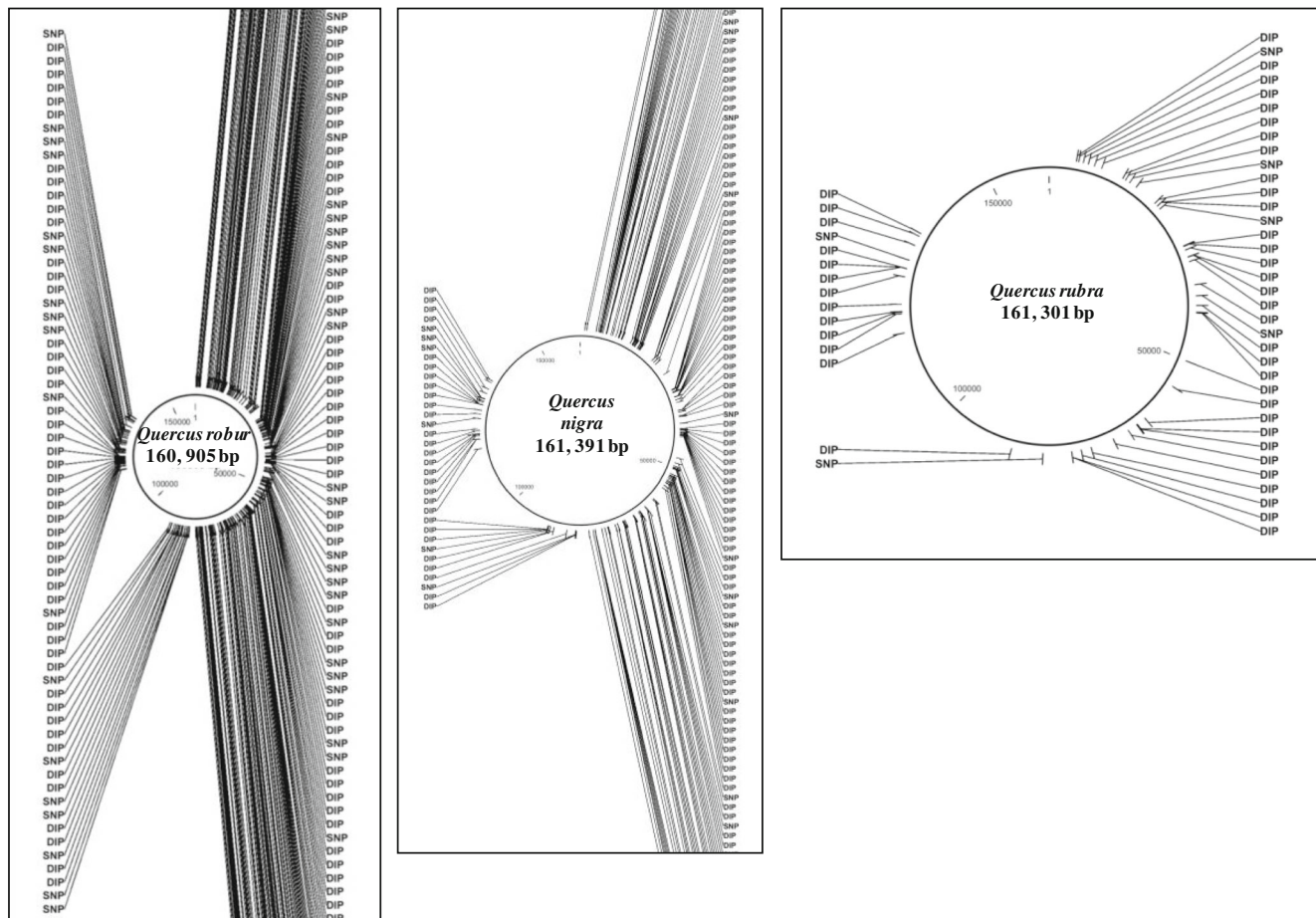


Fig. 2 Overview of chloroplast genome single nucleotide polymorphism (SNP) and deletion/insertion polymorphism (DIP) distribution between *Quercus rubra* and *Q. robur* (left; 737 SNPs, 214 DIPs), *Q. rubra* and *Q.*

nigra (center; 23 SNPs, 163 DIPs) and within four *Q. rubra* individuals (right; six SNPs, 45 DIPs). Polymorphisms are clustered in the large and small single copy regions and notably absent in inverted repeat regions

contigs assembled de novo relative to the reference can be used to identify putative insertions, deletions, and rearrangements that would otherwise be incorrectly assembled in RGA (Parks et al. 2010). In this case, generating a *Q. rubra* chloroplast consensus sequence using reference mapping alone led to a 99.4 % sequence similarity to the *Q. nigra* cp genome. Including de novo contigs to generate a pseudoreference decreased the similarity between genomes to 98.8 %, indicating that 0.6 % (~970) more base positions in the final consensus sequence were from *Q. rubra* (rather than the reference species). This iterative method represents an improvement over reference mapping alone, particularly when searching for polymorphisms between closely related species.

Pine chloroplast genomes assembled by Parks et al. (2012) and Cronn et al. (2012) using an Illumina platform and the assembly methods herein were estimated to be 92 % and 98 % complete, respectively. Using the GS 20 platform as described here, the assemblies of Moore et al. (2006) were estimated to be >99.6 % complete. Although no resequencing or in-group/out-group analysis is available to quantify the completeness of our assembly, we expect it is within the 92–99 % published range due to the close relationship between the reference species *Q. nigra* and *Q. rubra* and the resulting high coverage of the reference genome (only 0.04 % of base positions were represented by less than 30× coverage with no zero-coverage areas). In order to fully assemble the *Q. rubra* chloroplast genome, IR boundaries should be sequenced using available chloroplast universal primers.

The northern red oak chloroplast genome consensus sequence was 161,304 bp, the second smallest oak chloroplast that has been reported (Kremer et al. 2012). Only the chloroplast of *Q. robur* (161,295 bp) is smaller. Chloroplast structure and gene order are conserved as in most plant chloroplasts, with an IR region of 25 kb separating large (91 kb) and small (20 kb) single-copy regions (Palmer 1991; Raubeson and Jansen 2005). The 138 chloroplast genes code for 32 tRNA genes, four rRNA genes, and 82 protein-coding genes, a total of 118 proteins, higher than the average of 90 for green plants (Ravi et al. 2008), 113 for *Nandina* and American sycamore (Moore et al. 2006), and 127 for *Castanea* (Jansen et al. 2010). Unlike many angiosperm taxa, Northern red oak appears to contain a functional *ycf1* gene. A functional *ycf1* gene was also reported for *Prunus persica* and *Castanea mollissima* (Jansen et al. 2010). Four genes have been reported to show variation within Fagaceae: *infA*, *accD*, *rps18*, and *petA* (Kremer et al. 2012). In *Q. rubra*, *infA* appears functional as in other oak species. The *accD* gene is composed of 542 codons, shorter than the 553 reported for most oaks, but longer than the same gene in *Fagus* and *Castanea*. The *rps18* gene reportedly displays a length reduction of about 20 amino acids compared to other *Quercus* species; we found the *rps18* gene to

contain 102 amino acids. We found no evidence of a large, direct repeat in the *petA* gene as has been reported for other Fagaceae (Kremer et al. 2012).

In a rangewide study consisting of 290 individuals collected from 66 natural populations, Magni et al. (2005) used five chloroplast primer/enzyme combinations to evaluate northern red oak population subdivision. They found 12 total haplotypes with one haplotype present in 75 % of individuals, and estimated a population subdivision value (G_{ST}) of 0.46 ± 0.08 . No spatial structure of chloroplast genetic diversity was detected. The estimate of population subdivision was higher ($G_{ST}=0.53$) for 35 populations in the northwestern part of the native range, however, in a series of three studies covering the northwestern range of northern red oak, only five haplotypes were found (Romero-Severson et al. 2003; Feng et al. 2008). We found a promising number of polymorphisms among *Quercus* species and within *Q. rubra* in four northern red oak individuals, and primers were designed for the *Q. rubra* intraspecific SNPs for future assays of these and other populations.

The annotated chloroplast sequence of northern red oak is a valuable genetic tool with interspecific utility in providing new insights into the population and evolutionary genetics of oaks. The *Q. rubra* chloroplast map provides an ordered list of genes and their sequences that will be valuable as a reference for the construction of other Fagaceae chloroplast genomes. This map will increase the ease of identifying SNPs and performing population genetic analysis and studies of interspecific variability. In combination with nuclear markers, these polymorphisms should be useful for understanding the complex interaction of ecophysiology, morphology and adaptation that occurs in oak hybrid zones (Moran et al. 2012). This map can help elucidate, for example, the significance of hybridization and introgression in species evolution, where precise patterns of gene exchange in a species complex must be known. Quantifying cytoplasmic exchanges using sequence divergence rather than presence or absence of a few haplotypes will increase the resolution to detect introgression that informs whether the long term evolution and management of oak species should be considered independently (Dumolin-Lapegue et al. 1999). Similarly, calculations of chloroplast genetic distances (CGD) are restricted in resolution due to the low number of RFLP sites between species (Kremer et al. 2012). Improved CGD measures based on sequence divergence will improve partial correlation coefficients that describe the relative influences of biogeography and natural selection on population differentiation. Fragmentation of forests constitutes a threat to genetic diversity and viability of forest communities worldwide, and may be exacerbated in many areas by a warming climate. Forest managers can combine forest area and structure information with high resolution chloroplast sequence data of oak species to understand the influence of fragmentation on oak species and complexes. In

Europe, genetic variation of several oak species increased as fragmentation increased, likely due to addition by humans of seedlings with varying genotypes (Petit et al. 2002). This relationship may not hold true in the range of northern red oak, where man's record of impact is less intense, and species diversity is much higher than in Europe (Sauer 1988). Forest managers may also use annotated chloroplast sequences for identification purposes. Goals such as monitoring reintroduced populations (Worthen et al. 2010a,b), tracking wood products, and certifying oak seed lots and forest stands will be aided by this valuable tool.

Acknowledgments S. Schlarbaum provided access to the Watauga Seed Orchard; M. Moore and G.G. Vendramin provided sequence data; P. San Miguel provided sequencing and assembly support. The authors would like to thank R.C. Cronn, M. Moore, and three anonymous reviewers for their valuable improvements to the manuscript. The Fred van Eck foundation for Purdue University is gratefully acknowledged as the source of funding for this project. Mention of a trademark, proprietary product, or vendor does not constitute a guarantee or warranty of the product by the U.S. Dept. of Agriculture and does not imply its approval to the exclusion of other products or vendors that also may be suitable.

Data archiving statement The sequence generated herein was deposited in the NCBI nucleotide database 16 January 2013 (accession JX970937).

References

- Abrams MD (2002) The postglacial history of oak forests in eastern North America. In: McShea WJH, Healey WM (eds) Oak forest ecosystems: ecology and management for wildlife. Johns Hopkins University Press, Baltimore, pp 34–45
- Altshuler D, Pollara VJ, Cowles CR, Etten WJV, Baldwin J, Linton L, Lander ES (2000) A SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407:513–516
- Birchenko I, Feng Y, Romero-Severson J (2009) Biogeographical distribution of chloroplast diversity in northern red oak (*Quercus rubra* L.). *Am Midl Nat* 161(1):134–145
- Birky CW (1978) Transmission genetics of mitochondria and chloroplasts. *Annu Rev Genet* 12:471–512
- Birky CW, 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
- Bowman CM, Dyer TA (1982) Purification and analysis of DNA from wheat chloroplasts isolated in nonaqueous media. *Anal Biochem* 122:108–118
- Cavender-Bares J, Holbrook NM (2001) Hydraulic properties and freezing-induced cavitation in sympatric evergreen and deciduous oaks with contrasting habitats. *Plant Cell Environ* 24:1243–1256
- Cavender-Bares J, Ackerly DD, Baum DA, Bazzaz FA (2004) Phylogenetic overdispersion in Floridian oak communities. *Am Nat* 163:823–843
- 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* 36:19
- Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV, Udall J (2012) Targeted enrichment strategies for next generation plant biology. *Am J Bot* 99(2):291–311
- Deguilloux MF, Dumolin-Lapegue S, Gielly L, Grivet D, Petit RJ (2003) A set of microsatellite primers for the amplification of chloroplast microsatellites in *Quercus*. *Mol Ecol Notes* 3:24–27
- Dumolin S, Demesure B, Petit RJ (1995) Inheritance of chloroplast and mitochondrial genomes in pedunculate oak investigated with an efficient PCR method. *Theor Appl Genet* 91:1253–1256
- Dumolin-Lapegue S, Kremer A, Petit RJ (1999) Are chloroplast and mitochondrial DNA variation species independent in oaks? *Evolution* 53:1406–1413
- Feng Y, Sun YL, Romero-Severson J (2008) Heterogeneity and spatial autocorrelation for chloroplast haplotypes in three old growth populations of northern red oak. *Silvae Genet* 57(4–5):212–220
- Ferris C, Oliver RP, Davy AJ, Hewitt GM (1993) Native oak chloroplasts reveal an ancient divide across Europe. *Mol Ecol* 2:337–344
- Gall WR, Taft KA (1973) Variation in height growth and flushing of northern red oak (*Quercus rubra* L.) In Proceedings of the 12th southern forest tree improvement conference, Baton Rouge, pp 190–199
- Hokanson SC, Isebrands JG, Jensen RJ, Hancock JF (1993) Isozyme variation in oaks of the Apostle Islands in Wisconsin: genetic structure and levels of inbreeding in *Quercus rubra* and *Q. ellipsoidalis* (Fagaceae). *Am J Bot* 80:1349–1357
- Jansen RK, Raubeson LA, Boore JL, dePamphilis CW, Chumley TW, Haberle RC et al (2005) Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods Enzymol* 395:348–384
- Jansen RK, Sasaki C, Lee SB, Hansen AK, Daniell H (2010) 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
- Kremer A, Le Corre V, Petit R, Ducousso A (2010) Historical and contemporary dynamics of adaptive differentiation in European oaks. In: DeWoody JA, Bickham JW, Michler CH, Nichols K, Rhodes OE, Woeste KE (eds) Molecular approaches in natural resource conservation and management. Cambridge University Press, New York, pp 101–122
- Kremer A, Abbott AG, Carlson JE, Manos PS, Plomion C, Sisco P, Staton ME, Ueno S, Vendramin GG (2012) Genomics of Fagaceae. *Tree Genet Genome* 8:583–610
- Kriebel HB (1993) Intraspecific variation of growth and adaptive traits in North American oak species. *Ann For Sci* 50:153s–165s
- Lind JF, Gailing O (2013) Genetic structure of *Quercus rubra* L. and *Quercus ellipsoidalis* E. J. Hill populations at gene-based EST-SSR and nuclear SSR markers. *Tree Genet Genomes* 9:707–722
- Magni CR, Ducousso A, Caron H, Petit RJ, Kremer A (2005) Chloroplast DNA variation of *Quercus rubra* L. in North America and comparison with other Fagaceae. *Mol Ecol* 14:513–524
- Manos PS, Doyle JJ, Nixon KC (1999) Phylogeny, biogeography, and processes of molecular differentiation in *Quercus* subgenus *Quercus* (Fagaceae). *Mol Phylog Evol* 12:333–349
- Marguiles M, Egholm M, Altman WE, Attiya S, Bader JS, Bembem LA et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380
- Margulis L (1970) Origin of eukaryotic cells. Yale University Press, New Haven
- McGee CE (1974) Elevation of seed sources and planting sites affects phenology and development of red oak seedlings. *For Sci* 20:160–164
- Michaud H, Lumaret R, Ripoll JP, Toumi L (1995) A procedure for the extraction of chloroplast DNA from broad-leaved tree species. *Plant Mol Biol Report* 13:131–137
- Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Foltis KM, Soltis DE (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biol* 6:17. doi:10.1186/1471-2229-6-17
- 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 U S A* 104:19363–19368
- Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE (2010)

- Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc Natl Acad Sci U S A* 107:4623–4628
- Moran EV, Willis J, Clark JS (2012) Genetic evidence for hybridization in red oaks (*Quercus* sect. *Lobatae*, Fagaceae). *Am J Bot* 99:92–100
- Nixon KC (1993) Infrageneric classification of *Quercus* (Fagaceae) and typification of sectional names. *Ann Sci For Suppl 1* (Paris) 50:25s–34s
- Palmer JD (1991) Plastid chromosomes: structure and evolution. In: Bogorad L, Vasil IK (eds) *Cell culture and somatic cell genetics of plants*, vol 7A, Molecular biology of plastids. Academic Press, San Diego, pp 5–53
- Parks M, Liston A, Cronn RC (2010) Meeting the challenges of non-referenced genome assembly from short-read sequence data. *Acta Horticult* 859:323–332
- Parks M, Cronn RC, Liston A (2012) Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenetic data set from *Pinus* L. (Pinaceae). *BMC Evol Biol* 12:100
- Petit RJ (2004) Biological invasions at the gene level. *Divers Distrib* 10:159–165
- Petit RJ, Vendramin GG (2006) Phylogeography of organelle DNA in plants: an introduction. In: Weiss S, Ferrand N (eds) *Phylogeography of Southern European refugia*. Springer, Dordrecht, pp 23–100
- Petit RJ, Bialozyt R, Brewer S, Cheddadi R, Comps B (2001) From spatial patterns of genetic diversity to postglacial migration processes in forest trees. In: Silvertown J, Antonovics J (eds) *Integrating ecology and evolution in a spatial context*. Blackwell Science, Oxford, pp 295–318
- Petit RJ, Latouche-Hall C, Pemonge MH, Kremer A (2002) Chloroplast DNA variation in oaks in France and the influence of forest fragmentation on genetic diversity. *Forest Ecol Manag* 156:115–129
- Petit RJ, Bialozyt R, Garnier-Gere P, Hampe A (2004) Ecology and genetics of tree invasions: from recent introductions to Quaternary migrations. *For Ecol Manag* 197:117–137
- Petit RJ, Duminil J, Fineschi S, Hampe A, Salvini D, Vendramin GG (2005) Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Mol Ecol* 14:689–701
- Raubeson LA, Jansen RK (2005) Chloroplast genomes of plants. In: Henry RJ (ed) *Plant diversity and evolution: genotypic and phenotypic variation in higher plants*. CABI, Cambridge, pp 45–68
- Ravi V, Khurana JP, Tyagi AK, Khurana P (2008) An update on chloroplast genomes. *Plant Syst Evol* 271:101–122
- Romero-Severson J, Aldrich P, Feng Y, Sun W, Michler C (2003) Chloroplast DNA variation of northern red oak (*Quercus rubra* L.) in Indiana. *New Forest* 26:43–49
- Sauer JD (1988) *Plant Migration: The dynamics of geographic patterning in seed plant species*. University of California Press, Berkeley
- Schlarbaum SE, Bagley WT (1981) Intraspecific genetic variation of *Quercus rubra* L., northern red oak. *Silvae Genet* 30(2–3):50–56
- Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendal JF (1998) The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogeny reconstruction in a recently diverged plant group. *Am J Bot* 85:1301–1315
- Solomon AM (1983) Pollen morphology and plant taxonomy of red oaks in eastern North America. *Am J Bot* 70:495–507
- Sork VL, Huang S, Wiener E (1993) Macrogeographic and fine-scale genetic structure in a North American oak species, *Quercus rubra* L. *Ann For Sci* 50:261s–270s
- Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A (2012) Navigating the tip of the genomics iceberg: next-generation sequencing for plant systematics. *Am J Bot* 99:349–364
- Tovar-Sanchez E, Mussali-Galante P, Esteban-Jimenez R, Pinero D, Arias DM, Dorado O, Oyama K (2008) Chloroplast DNA polymorphism reveals geographic structure and introgression in the *Quercus crassifolia* x *Quercus crassipes* hybrid complex in Mexico. *Botany* 86:228–239
- Whittemore AT, Schaal BA (1991) Interspecific gene flow in sympatric oaks. *Proc Natl Acad Sci U S A* 88:2540–2544
- Worthen LM, Woeste KE, Michler CH (2010a) Breeding American chestnuts for blight resistance. *Plant Breed Rev* 33:305–339
- Worthen LM, Michler CH, Woeste KE (2010b) Wildlife reintroductions: the conceptual development and application of theory. In: DeWoody JA, Bickham JW, Michler CH, Nichols K, Rhodes OE, Woeste KE (eds) *Molecular approaches in natural resource conservation and management*. Cambridge University Press, New York, pp 101–122
- Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20:3252–3255