Development of novel *Quercus rubra* chloroplast genome CAPS markers for haplotype identification

Tim Pettenkofer[1], Reiner Finkeldey[2], Markus Müller[1], Konstantin V. Krutovsky[1,3,4,5,6], Barbara Vornam[1], Ludger Leinemann[1*], Oliver Gailing[1,3,7*]

[1] Faculty of Forest Sciences and Forest Ecology, Department of Forest Genetics and Forest Tree Breeding, University of Göttingen, Büsgenweg 2, 37077 Göttingen, Germany

[2] University of Kassel, Mönchebergstraße 19, 34125 Kassel, Germany

[3] Center for Integrated Breeding Research, Georg-August University of Göttingen, Göttingen, Germany

[4] Laboratory of Forest Genomics, Genome Research and Education Center, Siberian Federal University, 660036 Krasnoyarsk, Russia

[5] Laboratory of Population Genetics, N. I. Vavilov Institute of General Genetics, Russian Academy of Sciences, 119333 Moscow, Russia

[6] Department of Ecosystem Science and Management, Texas A&M University, 2138 TAMU, College Station, TX 77843-2138, USA

[7] Michigan Technological University, School of Forest Resources and Environmental Science, Houghton, MI 49931, United States*

* Corresponding authors: Oliver Gailing, E-mail: ogailin@gwdg.de; Ludger Leinemann, E-mail: lleinem@gwdg.de

Abstract

Our main objective was to generate cost-effective chloroplast (cp) DNA markers that are easy to apply and to score. In combination with already published cpSSR markers they should increase haplotype resolution in populations. To discover new cpDNA markers, we sequenced 87-97 % of the entire chloroplast genome (except the second inverted repeat) of 8 trees representing different regions of the *Quercus rubra* L. natural range with 4,030X-6,297X coverage and assembled the genome sequences using the publicly available chloroplast genome of *Quercus rubra* L. as a reference. In total, 118 single nucleotide polymorphisms (SNPs) and 107 insertions or deletions (indels) were detected, and 15 cleaved amplified polymorphic sequence (CAPS) markers were developed for *Q. rubra*. Using these new markers together with five chloroplast microsatellite or simple sequence repeat (cpSSR) markers, we identified 10 haplotypes in our diversity panel of 19 *Q. rubra* populations. Specifically, two haplotypes based only on the cpSSR markers could now be separated in five haplotypes. These markers are useful to assess haplotype diversity with high resolution and are also transferable to a closely related species, *Quercus ellipsoidalis* E. J. Hill.

Introduction

Northern red oak (*Quercus rubra* L.) is a dominant deciduous tree species covering a large natural area in eastern North America (Borkowski et al. 2014; Fig. 1). It grows best on mesic sites with deep sandy loam, but can also be found on pug, loam or gravelly, and sandy soils, and thus covers a wide ecological range (Desmarais 1998; Nagel 2015).

At larger distances between populations, significant geographical variation can be found in oak species at chloroplast DNA (cpDNA) (Petit et al. 1993; Grivet et al. 2006; Zhang et al. 2015). Within populations, cpDNA haplotypes usually show relatively low variation supposedly due to lack of recombination and its maternal inheritance (Finkeldey and Gailing 2013; Zhang et al. 2015). In the past, cpDNA markers have successfully been used to reveal the haplotype composition of autochthonous oak populations in Germany (Gailing et al. 2009) and to trace post-glacial recolonization routes of white oaks in Europe (Petit et al. 2002). By comparing identified haplotypes

and haplotype variation with reference populations within the natural range, information of their origin can be inferred. In contrary to European white oak species (Petit et al. 2002), *Q. rubra* is characterized by a relatively low geographic structure of chloroplast variation, making it difficult to distinguish between stands and geographic regions (Magni et al. 2005; Birchenko et al. 2009; Alexander and Woeste 2014). However, a relatively low number of chloroplast markers have been used in earlier studies (Magni et al. 2005; Birchenko et al., 2009; Zhang et al. 2015; Pettenkofer et al. 2019) potentially having resulted in a low resolution of haplotype diversity.
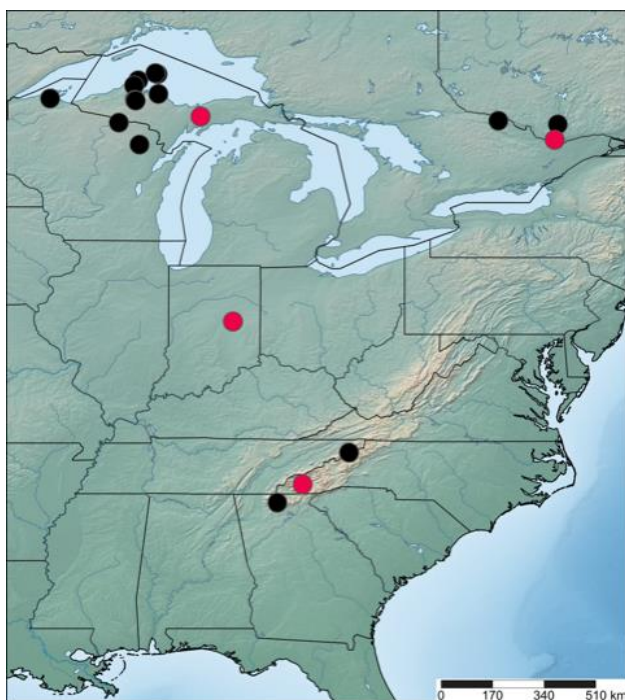

Fig. 1

Populations of *Quercus rubra* in the first set are marked in red (Liesebach and Schneck 2011), populations in the second set are marked in black (Lind and Gailing 2013; Lind-Riehl et al. 2014)

Cleaved amplified polymorphic sequence (CAPS) markers are a cost-effective and easy-to use method for the assessment of cpDNA polymorphisms and can be genotyped by analysing digested fragments in a simple agarose gel electrophoresis and do not need much DNA and expensive laboratory equipment (Kaundun and Matsumoto 2003). In case of cultivated stands of unknown origin, this tool can be applied to determine the origin of seeds or plants used to establish them if appropriate reference information for assignment is available. In combination with chloroplast microsatellite or simple sequence repeat (cpSSR) markers (Zhang et al. 2015; Pettenkofer et al. 2019), CAPS markers are expected to increase resolution of chloroplast variation. In this study, the red oak chloroplast genome was sequenced almost entirely (except the second inverted repeat region) for 8 samples representing different geographic regions, and the sequences were assembled using the publicly available chloroplast genome of *Q. rubra* as a reference (Alexander and Woeste 2014) for CAPS marker development. Only CAPs with a single polymorphic restriction site distinguishing one of the provenances from all others were selected.

Finally, the fully sequenced chloroplast genomes of trees representing all major regions within the natural range will also provide a good resource for the identification of additional polymorphic markers for the characterization of intraspecific differentiation in the future.

Our main objectives were to a) introduce a novel set of easy-to-use and cost-effective CAPS markers applicable for *Q. rubra* and closely related species such as *Q. ellipsoidalis*, and b) reveal more chloroplast haplotypes by combining CAPS with cpSSR markers.

Materials and Methods

Plant material

We used two sets of samples (Table 1). Bud samples in the first set were collected from eight trees representing four different regions (two trees per region) within the natural distribution range of *Q. rubra* (Table 1, Fig. 1) and were obtained from a provenance trial located in Luebeck, Germany. This provenance trial was established from seeds collected by A. Kremer in 1988 and then distributed to research institutes in France, Germany, Belgium, Great Britain and the USA (Liesebach and Schneck 2011). This set was used for amplicon-based sequencing of the chloroplast genome. The second set consisted of 96 samples collected from 19 different populations within the natural distribution range in North America and was used to develop and test CAPS markers (Table 1). The samples were selected to include different chloroplast haplotypes already identified in northern red oak populations in Germany and North America (Pettenkofer et al. 2019). Among 96 samples, 57 had cpSSR haplotype A, each of the haplotypes B, C, K and L was represented by eight samples, five samples had haplotype G, and two samples haplotype H. Except for the last two haplotypes, each haplotype was represented by samples from different locations of the natural range. This set included samples from populations described by Liesebach and Schneck (2011), Lind and Gailing (2013) and Lind-Riehl et al. (2014).

DNA isolation

Bud samples were collected in 2013 as fully grown, live buds from 10-20 cm long twigs. DNA was isolated subsequently with the DNeasy™ 96 Plant Kit from Qiagen (Hilden, Germany) following the instructions.

Amplicon sequencing

The red oak chloroplast genome (NCBI GenBank accession number JX970937.1) is 161,304 bp long and contains a large single copy (LSC), a short single copy (SSC) and two highly conserved inverted repeat (IR) regions (Alexander and Woeste 2014). The second inverted repeat region is the reverse complement of the first one. Excluding the second inverted repeat

Table 1

Study sites in North America

Stand	Set	Country	Location	N	Latitude	Longitude
Atomic Energy[1]	2	Canada, ON	Renfrew, Atomic Energy, Chalk River	5	-	-
Constance Bay[1]	1, 2	Canada, ON	Ottawa, Constance Bay	4	-	-
Pl. de Kazabazua[1]	2	Canada, ON	Gatineau, Plaines de Kazabazua, Basse-Lièvre	6	-	-
Chattahoochee[1]	2	USA, GA	Fannin, Chattahoochee, Toccoa	4	-	-
Anderson[1]	1, 2	USA, IN	Madison, Anderson	6	-	-
Hiawatha[1]	1, 2	USA, MI	Chippewa, Hiawatha, Soo	6	-	-
Nantahala[1]	1, 2	USA, NC	Clay, Nantahala, Tusquitee	6	-	-
Cherokee[1]	2	USA, TN	Sullivan, Cherokee, Watauga	5	-	-
FC-A[2]	2	USA, MI	Ford Forestry Center	8	46.65261	-88.50193
FC-B[2]	2	USA, MI	Ford Forestry Center	3	46.67442	-88.52427
BR1[2]	2	USA, MI	Brockway Mountain	3	47.46616	-87.91671
MTU[2]	2	USA, MI	Michigan Tech Trails	3	47.10055	-88.5475
HMR-LP[2]	2	USA, MI	Huron Mountain Res	8	46.84994	-87.83022
N-QE[3]*	2	USA, WI	Nicolet NF	4	45.32194	-88.33138
N-QR[3]	2	USA, WI	Nicolet NF	7	45.34805	-88.38805
Keweenaw[4]	2	USA, MI	Brockway Mountain	4	47.4401	-87.85658
Mine[4]	2	USA, MI	Calumet Township	2	47.25398	-88.42676
Porcupine[4]	2	USA, WI	Kentuck Lake	4	46.00003	-88.9996
C-QR[3]	2	USA, WI	Chequamegon NF	8	46.715	-91.03555

Note. MI - Michigan, WI - Wisconsin, GA - Georgia, IN - Indiana, NC - North Carolina, TN - Tennessee, ON - Ontario. [1] Liesebach and Schneck 2011, [2] Lind & Gailing 2013, [3] Lind-Riehl et al. 2014, [4] unpublished. The location of these populations in North America is shown in Fig. 1. *Population was genetically identified as *Q. ellipsoidalis* (Lind and Gailing 2013; Lind-Riehl et al. 2014), a species which is closely related to *Q. rubra*. Both species are interfertile and hybridize with each other in contact zones

region due to lack of variation (Alexander and Woeste 2014), a total of 32 primer pairs (Table 2) were designed using BioEdit v. 7.2.6.1 (Hall 1999) and OligoCalc v. 3.27 (Kibbe 2007) software to amplify 132,042 bp (96.87 %) of the genome (see Table 2 and Fig. 2).
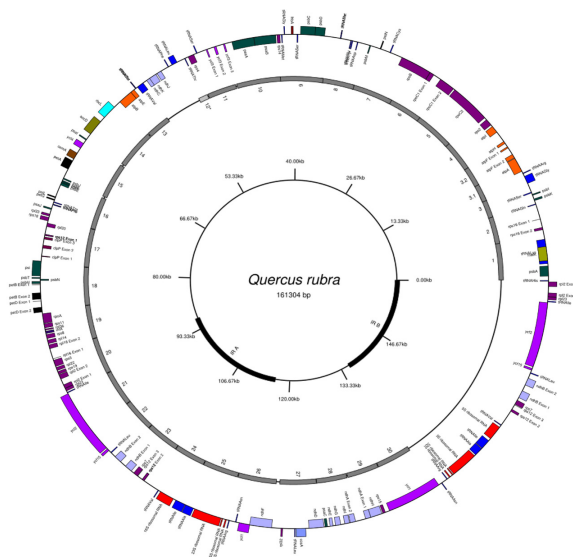

Fig. 2

Chloroplast genome of *Quercus rubra* assembled and annotated by Alexander and Woeste (2014). Bold lines in the inner circle represent inverted repeat regions. The middle circle indicates positions of primers and gaps. The second inverted repeat region was excluded from amplicon sequencing. Modified after Alexander and Woeste (2014)

The targeted cpDNA regions were PCR amplified in a 19 μl volume containing 1 μl of the template genomic DNA (~0.6 ng/μl), 10 μl HotStarTaq Master Mix from Qiagen (containing 5 units / μl HotStarTaq DNA Polymerase, PCR buffer with 3 mM $MgCl_2$ and 400 μM of each dNTP) (Hilden, Germany), 2.3 μl Q-Solution from Qiagen (Hilden, Germany), 1.1 μl $MgCl_2$ (25 mM), 2.3 μl of each forward and reverse primers (5 pM/μl).

PCR reactions were conducted in a MJ Research PTC-200 thermocycler (St. Bruno, Canada). The PCR profile started with 15 min initial denaturation at 95 °C, followed by 38 cycles of 1 min denaturation at 94 °C, 30 sec annealing at 55 °C and 6 min extension at 68 °C, and ended with a final 20 min extension step at 72 °C.

The PCR products were checked in an agarose gel electrophoresis at 90 volts for 20 to 30 minutes using a 1 % agarose gel with TAE as running buffer. The DNA was stained with Roti-Safe GelStain from Roth (Karlsruhe, Germany). Then, amplicon bands were cut out of the gel under UV light. DNA was extracted from the gel by using the innuPREP Gel Extraction Kit from Analytik Jena (Jena, Germany). The concentration of the eluted DNA was measured with the NanoDrop 2000 from Thermo Fisher Scientific (Waltham, USA).

The collected amplicons representing 32 cpDNA fragments were pooled for each sample and sequenced by the Transcriptome and Genome Analysis Laboratory (TAL) at the University of Göttingen using the Illumina MiSeq sequencing system. Amplicon sizes ranged from 2239 to 5666 bp (average 4295 bp).

Table 2
PCR primer sequences designed in the study and used to amplify the chloroplast genome regions (except the second inverted repeat region)

Primer Pair	Forward primer (5′-3′)	Reverse primer (5′-3′)	Nucleotide position in the reference genome		Amplicon length, bp
			start	end	
1	GTGGATTCACAATCCACTGCCT	AAACCCAAGCATTGCGATACTTT	47	5057	5010
1.1*	GTGGATTCACAATCCACTGCCT	CCCTTGGATTGCTGTTGCATATTC	47	1181	1134
1.2*	CCCCCACTCCCTGTATGTAGT	AAAGGAGCAATGTCAACCCTCTTG	97	496	399
2	CCCAGCAAATCCGACGAATCCATC	GTAAGGCAACGGGTTTTGGTCCCG	4722	8177	3455
3	GGTATGCTCTGGGACGGAAGG	TCCGCCCATAACATCTATGTCAGC	8139	11057	2918
3.1	CTCATTAGTCTCCTTGGCCAAAGC	TGATCCTTCCAAACAAGCACAGGC	10353	12592	2239
3.2*	CCGGCGGTCTTCGTAATAGAAGAG	GAAGGACTGATTGGAGAATGGGGA	12554	15093	2539
4	GGGAAGGAAGAAGGCGAGTC	AGAGAATTGAGGGTTGGAACAAG	15062	19248	4186
4.1*	TGATGACCCTCCATGGATTCGCCT	TTGCAAACATATCTGGGCGGGATC	16670	17890	1220
5	TTGCTAAAACTCGAGCAGTTTCTT	GGTCCCTCCCTTTCATTACAT	18875	23831	4956
5.2*	CCATGAGTAGGACTTCGGCCA	CGGGAGCTAGAGGAAATGCATCTC	21853	22378	525
6	TCGAAGCACGATGTTGTCTAATTA	CCAGAGACTGAATCTTGATATACC	23746	28272	4526
7	GCCCATAGTGACTAGGATGGA	GGGAGGTTCGAACTTACCTTTT	27838	33018	5180
8*	GAAGATCCTTTATCCATACCGAA	TCCCCAGACATATTCCTCTCG	32795	36929	4134
8.1*	CCTTGAAAGAGAGATGTCCTGAAC	AAGCCCCTTATCGGATTTGAACCG	34539	35674	1135
8.2*	CCGTTTTGAGTTTACCCATAATCAC	CATTCGCCACTATGAGTCTAGTGC	35372	35565	193
9	AGTGCGTTGCTTGCTTTGAACCGC	CGTCGATCCCCCAAAAAAGAA	36686	42009	5323
10	CGGTGGGGATTGTAACTTTCC	GTCAAAATAGAAGGTCCTCCGG	41946	47387	5441
11	GAGTGCTCCTATTCGAAACGCCTTG	GATTGGGTATGGCCTCGACTATTCCC	47541	51141	3600
12	GGCTTTTCCGGCAATACGAACG	GTTTATGAAAGAGCCCAATGCAC	51231	56897	5666
13	GATAAAGAGATGGCTCCGTGTGCTCGG	CCAACGAAATCTAGTGCAGTGCG	56965	59963	2998
14	GCGGATCAATAAAAGTCTTGAG	TTGTCATAAAATGGATACCTGG	60077	64991	4914
15	ATAAAATATGTGATGTGGTTTCCCCCG	CATTAAATCCCATCCCTCTTTTGCGG	65200	69687	4487
16	GATCATTAAGAACTCAACGGG	CATTCCAAAATAACTGTTACTCG	69896	74012	4116
17	GCTAATAAGGAATCAAAGAGG	CATGGAAATACCCCTTTAATCAACG	74060	78977	4917
17.1*	CCATGAGCTTGGGCTTCTGTTGCT	GGCCAAGAGGTTGATAGCGATCTC	76010	77115	1105
18	TATCGTGTTCATACCGTCGTATTG	CCAACTTAGCTCTTAATACAG	79014	83810	4796
19	TTATTATGGAGAACCCGCATGG	GAAAGGAATATCTTATCGAGGTAATCG	83850	88101	4251
20	GTTTACGGAATCTGGCTCTTTTGG	CAGACAAAAGTATTCGGTTATTG	88144	92771	4627
21	CTTCCCTATGAGTTTTAGTCTCAATAAG	AGAATTCAGCTCAGATGTAGG	92833	97328	4495
22	AGTTGCTCTGGAACAATTAGGAG	CCAAGAAATAACCCCTCACGTGCG	97372	101386	4014
23	GATAGAAACAACGCTCGTAAGGAG	CGCGGGCTTTACGCAATCGATCGG	101454	105630	4176
24	CGGAACCGGGGAAAGTATACAG	GGCAAGTCTTTGTGAAATAACTCCG	105661	109545	3884
25	GCACACTTGGAGAGCGCAGTACACCGG	AGGGGACACCAAAGGCCTCTG	109596	114020	4424
26	GCCTTTTATTCGTTCATGGTTCGATATTCTG	ATGGTTCATACCTTTCATTCCAG	114093	118847	4754
26.1*	GCGGTGACGATACTGTAGGGGAGG	TCTTCGGCGCACGTGGGCTTTTCC	113976	118571	4595
27	GCTCGAAAAAAGAGTGCACCTAATTG	TCCTGATGTTAGCAATGTATAGCG	119283	123777	4494
28	GAATCTCGAGTAACCGGCCAAG	TAAGTGGTCGAATTCGGGGCCG	123815	127591	3776
29*	ATTTGCTGAGTAATAACTTTGAG	GAGTCTGGATATTCTGATGGTCCG	127696	131625	3929
30	GGATTTCTCAGTCTAAGCAGGAG	CTTATTAGATCTAATAAGTACCTTGTG	131688	136055	4367
30.1*	AAGCCGTTTTCGGGGATAACC	CTGACGAGGTAGCTTTGATCCG	133645	134411	766

* used also for CAPS analysis (see also Fig. 2).

Sequence alignment and detection of SNPs and indels

Sequence alignment and detection of SNPs and indels were accomplished with the CLC Genomics Workbench v. 9.0.1 software (Qiagen, Aarhus, Denmark). Sequence reads were mapped to the *Q. rubra* chloroplast reference genome available from the NCBI GenBank nucleotide database (accession number JX970937.1). The BioEdit v. 7.2.6.1 (Hall, 1999) and Oligo-Calc v. 3.27 (Kibbe, 2007) software were used to identify restriction sites and develop CAPS markers. The identified restriction enzymes were subsequently tested using the 96 population samples from different regions within the natural distribution area. The aligned sequences of the 8 samples can be accessed via their NCBI GenBank accession numbers MK706519-MK706526.

Restriction digestion

Directly after PCR amplification, 1 μl of restriction enzyme and 2 μl of associated buffer were added to 8 μl of the PCR reaction. Then, the mixtures were incubated following the

recommended protocol provided by the manufacturer (Thermo Fisher Scientific, Waltham, USA). After incubation, an agarose gel electrophoresis was performed at 90 volts for 20 to 30 minutes on a 1 %, 1.5 % or 2.5 % agarose gel (depending on fragment sizes) with TAE as running buffer.

Identification of haplotypes

Chloroplast microsatellite haplotypes for each sample were identified in Pettenkofer et al. (2019). They were analysed together with the CAPS haplotypes identified for each sample in this study. Haplotypes based on both cpSSR and CAPS polymorphisms were inferred using MS Excel.

Data analyses

The cpSSR haplotype data were converted into a binary 0/1-matrix using MS Excel to combine them with CAPS markers for haplotype network analysis. The Arlequin v. 3.5 software (Excoffier and Lischer 2010) was used to compute the haplotype network based on both cpSSR and CAPS markers.

Results

Depending on the sample, 87-97 % of the reference genome was sequenced with 4,030-6,297X coverage (Table 3). In total, 118 SNPs and 107 indels were identified, which were subsequently used to identify restriction sites and design CAPS markers. In this study, we selected only restriction sites that were present in one provenance while absent in the others. The PCR primers were designed to target only a single restriction site to simplify the scoring procedure. From 23 primer-enzyme combinations tested initially, 15 showed easily-interpreted results and were validated in a larger dataset (Table 4). These 15 CAPs markers together with cpSSR markers developed by Pettenkofer et al. (2019) increased haplotype resolution within the natural range of *Q. rubra* and allowed to discover 10 different haplotypes (Table 5). The five CAPS markers 1.1, 4.1, 5.2, 11 and 17.1 and cpSSR *udt1* were sufficient to distinguish these 10 haplotypes. The most frequent cpSSR haplotype A splits up into three different haplotypes, while haplotype K splits into two haplotypes (Figs 3 and 4, Table 5). All developed CAPS markers were successfully amplified also in samples of *Q. ellipsoidalis* (population N-QE).

Discussion

Alexander and Woeste (2014) found only six SNPs and 45 indels in the entire chloroplast genome of four red oak individuals that represented mostly the south and north-east of the natural range. We found much more markers (118 SNPs and 107 indels) in a larger and more diverse sample. This number is rather large considering the conservation of the chloroplast genome. Generally, the number of SNPs increases with taxonomic divergence (Alexander and Woeste 2014; Nguyen et al. 2018).

Table 3

Mapping of sequence reads to the *Q. rubra* chloroplast reference genome

Population[1]	Sample ID	Reference covered by reads, %	Total number of reads mapped	Mean read length, bp	Total read length, bp	Average coverage
Nantahala	3	92	3,325,327	226.75	754,029,642	5,488
	4	94	3,298,374	222.61	734,237,502	5,335
Anderson	7	97	3,336,145	204.31	681,614,046	4,965
	8	96	2,557,784	215.95	552,357,621	4,030
Hiawatha	10	93	3,018,802	214.73	348,233,509	4,718
	12	87	3,052,790	222.67	679,766,646	4,959
Constance Bay	13	93	3,881,031	222.43	863,256,125	6,297
	14	94	3,374,986	207.92	701,724,192	5,103

Note. GC-content of sequence reads was 36% in all samples. Total length of genome used as reference was 136299 bp. [1]The locations of these populations in North America are presented in Fig. 1.
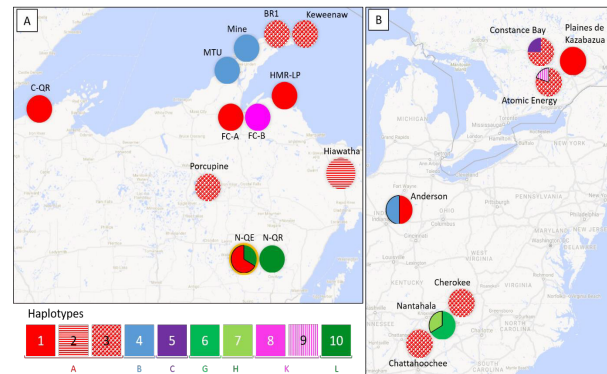

Fig. 3

Haplotypes identified in the studied populations of *Quercus rubra*. Numbers indicate haplotypes based on both cpSSR and CAPS markers (see also Fig. 4 and Table 4). Letters indicate cpSSR-based haplotypes. A –Haplotypes identified in the area of the Great Lakes. Population of *Q. ellipsoidalis* is marked with a yellow circle. B - Haplotypes identified in North American populations (see Tables 1 and 4 for details). The cpSSR haplotype A was further separated into haplotypes 1, 2, and 3, and the cpSSR haplotype K was separated into haplotype 8 and 9 (see also Table 4)

Table 4

Primer-enzyme combinations designed and tested in the study

#	Primer pair	Nucleotide position in the reference genome[1]		Amplicon size, bp	Restriction enzyme	SNP position in the reference genome	Number of fragments after cutting	Sense sequence (5'-3')	Antisense sequence (3'-5')
		start	end						
1	1.1	47	1181	1134	*HinP1*I	557	2	-	*psbA* exon
2	1.2	97	496	399	*Aci*I	200	2	-	intergenic *trnH-psbA*
3	3.2	12554	15093	2539	*Bmg*BI	12918	2	-	*atpA* exon
4[2]	4.1	16670	17890	1220	*Hae*III	17341	2[3]	-	*atpI* exon
5[2]	5.2	21853	22378	525	*Aci*I	22029	2	-	*rpoC2* exon
6	8	32795	36929	4134	*Hha*I	35513	2	intergenic *petN-trnT*	intergenic *trnS-trnE*
7	8.1	34539	35674	1135	*Cla*I	34889	2	intergenic *petN-trnT*	intergenic *trnS-trnE*
8	8.2	35372	35565	193	*Mse*I	35426	2	intergenic *petN-trnT*	intergenic *trnS-trnE*
9	11	47541	51141	3600	*Eco*RV	48925	2	-	*ycf3* intron
10[2]	17.1	76010	77115	1105	*Apa*I	76425	2	*rps12*	*clpP* intron, *rps12*
11	26.1	113976	118571	4595	*Swa*I	117302	2	*rps12*	*ndhF* exon
12	26.1	113976	118571	4595	*Pac*I	114509	2	intergenic *trnR-ycf1*	-
13	29	127696	131625	3929	*Eag*I	129513	2	*rps12*	*ndhH* exon
14	29	127696	131625	3929	*Bgl*II	130647	2 or 3	*rps12*	intergenic *ycf1-rps15*
15	30.1	133645	134411	766	*Dra*I	133915	2	*rps12*	*ycf1* exon

[1] NCBI GenBank accession number JX970937.1. [2] Combinations of these CAPS together with the cpSSR markers were sufficient for the identification of all 10 haplotypes (see also Table 5). [3] Three fragments were observed in samples from FC-B.

Table 5

Restriction fragments and cpSSR alleles associated with the 10 haplotypes

CAPS+ cpSSRs	Haplotype CAPS	cpSSRs	1.1 *HinP1*I	1.2 *Aci*I	3.2 *Bmg*BI	4.1* *Hae*III	5.2 *Aci*I	8 *Hha*I	8.1 *Cla*I	8.2 *Mse*I	11 *Eco*RV	17.1 *Apa*I	26.1 *Swa*I	26.1 *Pac*I	29 *Eag*I	29* *Bgl*II	30.1 *Dra*I	ccmp2	ccmp4	ucd4	udt1	udt4
1	1	A	2	1	2	2	1	1	2	1	1	1	2	2	1	2	1	228	116	99	86	145
2	2	A	2	1	2	2	1	1	2	1	1	2	2	2	1	2	1	228	116	99	86	145
3	3	A	2	1	2	2	2	1	2	1	1	1	2	2	1	2	1	228	116	99	86	145
4	5	B	2	1	2	1	1	1	1	2	1	1	2	2	2	3	1	227	115	98	85	146
5	3	C	2	1	2	2	2	1	2	1	1	1	2	2	1	2	1	228	116	99	87	145
6	4	G	1	2	1	2	2	2	2	1	2	1	1	1	1	2	2	226	117	97	85	146
7	4	H	1	2	1	2	2	2	2	1	2	1	1	1	1	2	2	226	117	97	84	146
8	6	K	2	1	2	3	2	1	2	1	1	1	2	2	1	2	2	228	116	99	85	145
9	3	K	2	1	2	2	2	1	2	1	1	1	2	2	1	2	1	228	116	99	85	145
10	7	L	1	2	2	2	2	2	2	1	1	1	1	1	1	2	2	226	118	98	85	146

Note. [1] Numbers indicate numbers of fragments observed after digestion by restriction enzymes. [2] Described in Weising and Gardner (1999) Deguilloux et al. (2003). 4.1-*Hae*III and 29-*Bgl*II had more than one restriction site and were genotyped as two loci to allow analysis with the Arlequin software. CAPS markers 1.1, 4.1, 5.2, 11, 17.1 were sufficient to identify all seven CAPS-based haplotypes. CAPS markers 4.1, 5.2 and 17.1 together with cpSSR markers were sufficient to identify all 10 haplotypes

As expected, the seven CAPS haplotypes included the main haplotypes previously identified with cpSSR markers (Table 5; Pettenkofer et al. 2019). However, together with cpSSR markers a higher haplotype resolution can be obtained. Some CAPS markers identify the same haplotypes due to the tight linkage and linkage disequilibrium. Therefore, it is important to notice that only three CAPS markers together with five cpSSR markers are needed to detect all 10 haplotypes found in our sample (see Table 5). The five CAPS markers 1.1, 4.1, 5.2, 11 and 17.1 together with *udt1* showed the highest genotyping ability and cost efficiency since they also resolved all 10 cpDNA haplotypes. The successful application of all primer-enzyme combinations also in a population of *Q. ellipsoidalis* showed the transferability of these markers. The newly developed CAPS markers in combination with cpSSR *udt1* can thus be used as a cost-effective and easy way of determining the chloroplast variation in populations of these two oak species and potentially other closely related species. This opens new possibilities for the study of postglacial migration routes and for tracing seeding material or identifying the origin of established stands in case reference information for assignment is available. The obtained genome sequences of trees representing four different regions can also be used to develop additional cpSSR, indel and CAPS markers, and thus contribute to future studies of genetic diversity and structure in oak populations.

Acknowledgments

References

Alexander LW, Woeste KE (2014) Pyrosequencing of the northern red oak (*Quercus rubra* L.) chloroplast genome reveals high quality polymorphisms for population management. Tree Genetics & Genomes 10(4): 803–812. https://doi.org/10.1007/s11295-013-0681-1

Birchenko I, Feng Y, Romero-Severson J (2009) Biogeographical distribution of chloroplast diversity in northern red oak (*Quercus rubra* L.). The American Midland Naturalist 161(1): 134–145. https://doi.org/10.1674/0003-0031-161.1.134

Borkowski DS, McCleary T, McAllister M, Romero-Severson J (2014) Primers for 52 polymorphic regions in the *Quercus rubra* chloroplast, 47 of which amplify across 11 tracheophyte clades. Tree Genetics & Genomes 10(4): 885–893. https://doi.org/10.1007/s11295-014-0729-x

Deguilloux M-F, Dumolin-Lapègue S, Gielly D, Grivet D, Petit RJ (2003) A set of primers for the amplification of chloroplast microsatellites in *Quercus*. Molecular Ecology Notes 3(1): 24–27. https://doi.org/10.1046/j.1471-8286.2003.00339.x

Desmarais KM (1998) Northern red oak regeneration: biology and silviculture, University of New Hampshire, Department of Natural Resources, Durham, NH.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5. A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources 10(3): 564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

Finkeldey R, Gailing O (2013) Genetics of chloroplasts. In: Maloy S, Hughes K (eds) Brenner's Encyclopedia of Genetics. Elsevier, pp 525–527

Gailing O, Wachter H, Heyder J, Rogge M, Finkeldey R (2009) Chloroplast DNA analyses of very old, presumably autochthonous *Quercus robur* L. stands in North Rhine-Westphalia. Forst- und Jagdzeitung 180: 221–227

Grivet D, Deguilloux M-F, Petit RJ, Sork VL (2006) Contrasting patterns of historical colonization in white oaks (*Quercus spp*.) in California and Europe. Molecular Ecology 15(13): 4085–4093. https://doi.org/10.1111/j.1365-294X.2006.03083.x

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41: 95–98

Kaundun SS, Matsumoto S (2003) Development of CAPS markers based on three key genes of the phenylpropanoid pathway in tea, *Camellia sinensis* (L.) O. Kuntze, and differentiation between *assamica* and *sinensis* varieties. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik 106(3): 375–383. https://doi.org/10.1007/s00122-002-0999-9

Kibbe WA (2007) OligoCalc. An online oligonucleotide properties calculator. Nucleic Acids Research 35(Web Server issue): W43-6

Liesebach M, Schneck V (2011) Entwicklung von amerikanischen und europäischen Herkünften der Roteiche in Deutschland. Development of American and European provenances of northern red oak in Germany. Forstarchiv 82(4): 125–133

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 Genetics & Genomes 9(3): 707–722. https://doi.org/10.1007/s11295-012-0586-4

Lind-Riehl JF, Sullivan AR, Gailing O (2014) Evidence for selection on a CONSTANS-like gene between two red oak species. Annals of Botany 113(6): 967–975. https://doi.org/10.1093/aob/mcu019

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(2): 513–524. https://doi.org/10.1111/j.1365-294X.2005.02400.x

Nagel R-V (2015) Roteiche (*Quercus rubra* L.). In: Vor T, Spellmann H, Bolte A et al (eds) Potenziale und Risiken eingeführter Baumarten. Göttingen University Press, Göttingen: 219-267

Nguyen VB, Giang VNL, Waminal NE, Park H-S, Kim N-H, Jang W, Lee J, Yang T-J (2018) Comprehensive comparative analysis of chloroplast genomes from seven *Panax* species and development of an authentication system based on species-unique single nucleotide polymorphism markers. Journal of Ginseng Research 44: 135-144. https://doi.org/10.1016/j.jgr.2018.06.003

Petit RJ, Brewer S, Bordács S, Burg K, Cheddadi R, Coart E, Cottrell J, Csaikl UM, van Dam B, Deans JD, Espinel S, Fineschi S, Finkeldey R, Glaz I, Goicoechea PG, Jensen JS, König AO, Lowe AJ, Madsen SF, Mátyás G, Munro RC, Popescu F, Slade D, Tabbener H, Vries SGM de, Ziegenhagen B, Beaulieu J-L de, Kremer A (2002) Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. Forest Ecology and Management 156(1-3): 49–74. https://doi.org/10.1016/S0378-1127(01)00634-X

Petit RJ, Kremer A, Wagner DB (1993) Geographic structure of chloroplast DNA polymorphisms in European oaks. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik 87(1-2): 122–128. https://doi.org/10.1007/BF00223755

Pettenkofer T, Burkardt K, Ammer C, Vor T, Finkeldey R, Müller M, Krutovsky K, Vornam B, Leinemann L, Gailing O (2019) Genetic diversity and differentiation of introduced red oak (*Quercus rubra*) in Germany in comparison to reference native North American populations. European Journal of Forest Research 138(2): 275-285. https://doi.org/10.1007/s10342-019-01167-5

Weising K, Gardner RC (1999) A set of conserved PCR Primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. Genome 42(1): 9–19. https://doi.org/10.1139/g98-104

Zhang R, Hipp AL, Gailing O (2015) Sharing of chloroplast haplotypes among red oak species suggests interspecific gene flow between neighboring populations. Botany 93(10): 691–700. https://doi.org/10.1139/cjb-2014-0261

Supporting information

All SNPs and indels identified in the *Quercus rubra* chloroplast genome, as well as tested CAPS markers are listed in the supplemental Microsoft Excel file.