bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1 Development of molecular markers for determining continental origin of wood from white
2 oaks (*Quercus* L. sect. *Quercus*)
3
4
5
6 H. Schroeder[1], R. Cronn[2], Y. Yanbaev[3], T. Jennings[4], M. Mader[1], B. Degen[1], B. Kersten[1]
7
8 [1]Thuenen-Insitute of Forest Genetics, Sieker Landstrasse 2, 22927 Grosshansdorf, Germany
9 [2]US Forest Service, Pacific Northwest Research Station, 3200 SW Jefferson Way
10 Corvallis, OR 97331, USA
11 [3]Bashkir State University, 450076 Ufa, Russia
12 [4]Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331, USA
13
14
15 Keywords
16 Next Generation Sequencing, chloroplast, molecular markers, illegal logging, timber trading
17
18
19 Corresponding author:
20 Hilke Schroeder, Sieker Landstrasse 2, 22927 Grosshansdorf, Germany, fax:
21 +494102696200, email: hilke.schroeder@ti.bund.de
22
23 Running title
24 Molecular markers for white oak determination

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

25 **Abstract**

26 To detect and avoid illegal logging of valuable tree species, identification methods for the
27 origin of timber are necessary. We used next-generation sequencing to identify chloroplast
28 genome regions that differentiate the origin of white oaks from the three continents; Asia,
29 Europe, and North America. By using the chloroplast genome of Asian *Q. mongolica* as a
30 reference, we identified 861 variant sites (672 single nucleotide polymorphisms (SNPs); 189
31 insertion/deletion (indel) polymorphism) from representative species of three continents (*Q.*
32 *mongolica* from Asia; *Q. petraea* and *Q. robur* from Europe; *Q. alba* from North America),
33 and we identified additional chloroplast polymorphisms in pools of 20 individuals each from
34 *Q. mongolica* (789 variant sites) and *Q. robur* (346 variant sites). Genome sequences were
35 screened for insertion/deletion (indel) polymorphisms to develop markers that identify
36 continental origin of oak species, and that can be easily evaluated using a variety of
37 detection methods. We identified five indel and one SNP that reliably identify continent-of-
38 origin, based on evaluations of up to 1078 individuals representing 13 white oak species and
39 three continents. Due to the size of length polymorphisms revealed, this marker set can be
40 visualized using capillary electrophoresis or high resolution gel (acrylamide or agarose)
41 electrophoresis. With these markers, we provide the wood trading market with an instrument
42 to comply with the U.S. and European laws that require timber companies to avoid the trade
43 of illegally harvested timber.

44

45

46 **Introduction**

47 Illegal logging is a serious issue not only for tropical rainforests and tropical trees, but it is
48 also a concern for tree species in temperate latitude forests. White oaks from the genus
49 *Quercus* sect. *Quercus* (Fagaceae) provide a relevant example of illegal logging in a
50 temperate zone tree, and they highlight the challenge facing importers and regulatory
51 agencies responsible for validating the taxonomic and geographic sources of timber

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

52 products. White oaks account for a significant percentage of the hardwood flooring and
53 furniture trade in Europe and the USA, and they represent one of the most important
54 hardwoods in terms of logs and lumber exports from these regions. The most important trade
55 woods of white oaks derive from the European species *Quercus robur* L. and *Q. petraea*
56 (Mattuschka) Liebl., the CITES Appendix III-protected *Q. mongolica* Fisch. Ex Ledeb. native
57 to East-Asia, and North American oaks, such as *Q. alba* L. and *Q. macrocarpa* Michx.
58 (Cassens2007). Non-governmental organizations such as the Environmental Investigation
59 Agency (http://eia-global.org/news-media/liquidating-the-forests) have documented increases
60 in the rate of illegal logging for white oak wood, especially in the Russian Far East region.
61 These activities increase the likelihood that international wood trading companies will market
62 illegally harvested wood, an activity that is banned by the U.S. Lacey Act amendment of
63 2008 and the European Union timber regulation of 2010. Violation of these regulations can
64 result in fines, forfeiture of wood, and additional payments, as was recently demonstrated
65 with improperly documented shipments of white oak flooring in the United States (US
66 Department of Justice, 2015). Under these laws, timber companies are responsible for
67 avoiding the trade of illegally harvested timber, and they are obligated to declare the species
68 name and geographic origin of traded timber in order to reduce the risk that traded timber
69 originated from illegal logging (Dormontt *et al.* 2015).
70
71 The increased attention to illegal logging has led to an increased demand for methods that
72 can be used to provide precise species identification and geographic origin verification.
73 Wood anatomical methods are widely used for tree species identification (Dormontt *et al.*
74 2015), but these methods cannot discriminate white oak species, nor identify geographic
75 origin of oaks generally. Over the last decade, worldwide programs have been established
76 using the potential of DNA as universal tool for identifying organisms (Barcode of Life
77 www.barcodeoflife.org, Hollingsworth *et al.* 2009). In plants, the success of barcoding is
78 highly dependent on several factors, including magnitude of primary divergence, frequency of

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

79 secondary contact, and mutation rate of the DNA region (Hollingsworth *et al.* 2011), so the
80 choice of suitable barcode regions in plants can be difficult (Chase *et al.* 2005; Newmaster *et*
81 *al.* 2006; Kress & Erickson 2008). Barcoding efforts in plants have focused on chloroplast
82 genomes due their simple pattern of (typically) uniparental inheritance, low effective
83 population size, and useful variation at the scale of geography and taxonomy across a wide
84 range of species (e.g. Kress & Erickson 2007; Taberlet *et al.* 2007; Lahaye *et al.* 2008;
85 Janzen *et al.* 2009; Huang *et al.* 2015). With advances in next-generation sequencing,
86 chloroplast genomes are affordable to sequence in their entirety by 'skimming' methods
87 (Straub *et al.* 2012), and whole genome analysis can reveal substantial variation, even in
88 unexpected genomic regions that are included in traditional barcoding efforts (e.g., Parks *et*
89 *al.* 2009, 2012). DNA barcoding already has proven to be appropriate for revealing illegal
90 trading (e.g. Goncalves *et al.* 2015; Pappalardo & Ferrito 2015), and it is increasingly used to
91 identify plant species in commercial trade (e.g., Handy *et al.* 2011).
92
93 The aim of this study is to use chloroplast-genome scale information to develop a cost-
94 efficient, easy-to-use assay that allows the identification of the geographic origin of white oak
95 wood products to hemisphere (Old World vs. New World) and continent (Asia; Europe; North
96 America) to support regulatory and commercial efforts to detect illegal logging of *Q.*
97 *mongolica*.
98
99 **Material and Methods**
100
101 **Plant material**
102 For next-generation sequencing, we sequenced a single oak individual from four species and
103 three continents to produce chloroplast genome references; included are *Q. mongolica* from
104 Asia (sample QUMO5_CH_1; China), *Q. petraea* from Europe (sample QUPE2_PO_1;
105 Poland), *Q. robur* from Europe (sample QURO2_SVT6; Germany), and *Q. alba* from North

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

106 America (sample QUAL_VT_1; USA). To develop a panel of polymorphisms for Asia and
107 Europe, we used next-generation sequencing to screen two pooled DNA samples that
108 included 20 individual specimens of *Q. mongolica* or European *Q. robur*, respectively. Each
109 of the *Q. robur/Q. mongolica* specimens were sampled from 10 geographically-widespread
110 populations. To develop a panel of polymorphisms for North American white oaks, we
111 sequenced chloroplast genomes from additional specimens representing the following
112 species: *Q. alba*, *Q. bicolor* Willd., *Q. garryana* Douglas ex. Hook., *Q. lyrata* Walter, *Q.*
113 *macrocarpa* Michx., *Q. michauxii* Nutt., *Q. prinoides* Willd., and *Q. stellata* Wangenh.
114
115 For marker validation, DNAs from 13 *Quercus* species were screened. *Q. mongolica* and *Q.*
116 *dentata* Thunb. represented Asian oaks (Far East Russia) with 200 and 10 specimens,
117 respectively. *Q. robur*, *Q. petraea* and *Q. pubescens* WillD. represented European oaks, with
118 360, 210, and 200 specimens, respectively. Finally, eight white oak species were screened
119 from North America (*Q. alba*, *Q. bicolor*, *Q. garryana*, *Q. lyrata*, *Q. macrocarpa*, *Q. michauxii*,
120 *Q. prinoides*, *Q. stellata*), with between 5 and 25 specimens per species.
121
122 **Next-generation sequencing analyses**
123 Aliquots of DNA (~0.5 - 1 μg) were sheared to a median length of ~300 bp and converted into
124 sequencing libraries using Illumina TruSeq v.2 kits at the USDA Forest Service (Corvallis,
125 OR; individual samples, each indexed with dual-index adapters) or GATC Biotech AG
126 (Konstanz, Germany; pooled samples). Sequencing was performed using three approaches:
127 (A) using the Illumina MiSeq with 2x150 bp paired-end reads for individual *de novo* genome
128 reference assemblies; (B) using the Illumina MiSeq with 2x300 bp paired-end reads for
129 pooled samples and reference-guided mapping; and (C) using the Illumina HiSeq with 100
130 bp single-end reads for individual North American species and reference-guided mapping. All
131 reactions used version 3 sequencing chemistry. Information on raw clusters, sequence yield,

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

132 and approximate target sequence (chloroplast genome) coverage depth is provided in Table
133 1.
134
135 *De novo reference construction* – Raw read quality filtering was accomplished using
136 Trimmomatic v0.30 (Bolger *et al.* 2014), and we removed reads with a mean Phred score
137 less than 33. Reads were digitally normalized to a coverage of 20 using, khmer v0.7.1
138 (Crusoe *et al.* 2014), and kmer-filtered sequences were assembled using Velvet Optimizer
139 v2.2.5 (https://github.com/Victorian-Bioinformatics-Consortium/VelvetOptimiser.git) and
140 Velvet v1.2.10 (Zerbino & Birney 2008). K-mer lengths ranging from 21 to 121 were
141 evaluated, with a final k-mer length of 121 selected for assembly. *De novo* contigs ≥ 100 bp
142 in length were screened for homology to the *Quercus rubra* chloroplast genome (NCBI
143 NC_020152) using BLAT (Kent 2002). Contigs showing high similarity were retained for
144 reference assembly and ordered against the *Q. rubra* chloroplast genome. Reference
145 sequences were constructed to include the large single-copy (LSC) region, one of two
146 inverted repeats (IR), and the small single-copy (SSC) region, equivalent to positions 1 –
147 135,502 from *Q. rubra* NC_020152.
148
149 *Reference-guided identification of SNP and indel variation* – Reference guided read mapping
150 and polymorphism detection was performed using CLC Genomics Workbench version 7.5.1
151 (CLC-bio, a Qiagen company; Aarhus, Denmark). The reference chloroplast sequence of the
152 *Q. mongolica* individual QUMO5_CH_1 generated by *de novo* reference construction (see
153 above) was used as reference for read mapping. The trimmed Illumina data of the two pools
154 (*Q. mongolica*, *Q. robur*) and the trimmed HiSeq data of representative North American
155 individuals were mapped to the reference scaffold using a length fraction of 0.9 and a
156 similarity fraction of 0.94. Variants detected by CLC Genomics Workbench included SNPs
157 and small indels, and these were exported to tab-delimited files and processed using an in-
158 house script (*Variant Tools*, see below) to identify species-specific polymorphism.

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

159

160 **Post processing of identified SNPs and indels**

161 To merge the SNP and indel tables and find common variants present in two or more
162 individuals/pools, we developed *Variant Tools,* a command line program implemented in
163 Ruby. This program merges individual sample SNP and indel tables (CSV format) produced
164 by CLC GWB to create a multi-individual SNP and indel matrix. Required input options
165 include the reference sequence (fasta format), an input directory containing the variant CSV
166 tables, and an option specifying the input data type (SNP or indel). The reference fasta can
167 contain one reference sequence or multiple reference contigs. Optionally, coverage tables
168 (produced by CLC from read mappings to a reference) of every individual can be included in
169 the analysis by specifying a directory containing coverage files (CSV format). Furthermore
170 several filtering options are available to reduce the output according to user-provided
171 thresholds. The output from *Variant Tools* is stored in a CSV file and contains several data
172 columns: the reference sequence name, the reference position, the variant length, the calling
173 type (SNP, MNP, deletion or insertion), the reference base(s), the alternative base(s) for
174 every individual, the coverage for every individual at the reference position, different
175 summary statistics, and sequences flanking the called variant.

176

177 The flanking sequences are calculated based on two given distance thresholds. An upper
178 and a lower threshold define minimum distances in base pairs between two called variants
179 on the genomic scale. If a variant occurs within the genomic range of the lower threshold no
180 flanking sequence is created. If a variant resides within the genomic range of the upper
181 threshold the length of the flanking sequence created is defined by the lower threshold. If no
182 variant is found within the range of both thresholds a flanking sequence with the length of the
183 upper threshold is calculated. The default thresholds are 75 bp and 50 bp.

184

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

185 The *Variant Tools* create different summary statistics while the variant matrix is generated.
186 *The number of individual alleles deviating from the reference* is a count for all found variants
187 in all individuals at a specific genomic position. *The number of alleles matching the reference*
188 *with minimal coverage* is a count for all positions in all individuals where no variant has been
189 called and that are supported by a minimum coverage. The threshold for the minimum
190 coverage is specified by the user. The default threshold is set to a minimum coverage of 8.
191 *Critical forward reverse balance* is an indicator for systematic sequencing errors and
192 describes how many forward and reverse reads are supporting the called variant. The value
193 is averaged over all individuals showing the variant.
194 The *Variant Tools* are open source software under ongoing development. They are available
195 under the terms of the ICS license and can be obtained from
196 https://github.com/ThuenenFG/varianttools.
197

198 **DNA extraction, PCR, restriction, and genotyping**

199 *Leaves*: One cm^2 of a single leaf was ground to powder in liquid nitrogen. Total DNA was
200 extracted, following a modified ATMAB protocol by Dumolin *et al.* (1995). PCR reactions for
201 leaf-derived DNA contained ~30 ng template DNA, 10x PCR buffer, 1.5 or 1.75 mM $MgCl_2$,
202 200 µM dNTPs, 0.4 unit AmpliTaq Gold DNA polymerase (ThermoFisher Scientific,
203 Darmstadt, Germany), and 0.05 to 0.13 µM of each primer in a total volume of 15 µl. PCR
204 was carried out in a Sensoquest Thermocycler (Göttingen, Germany) with a pre-denaturation
205 step at 94°C for 10 min, followed by 25 to 30 cycles of 94°C for 45 sec (30 sec *trn*CD),
206 suitable annealing temperature for each primer combination (between 52°C and 57°C) for 45
207 sec (30 sec *trn*CD), 72°C for 45 sec (1 min *trn*LF) and a final elongation at 72°C for 10 min.
208 PCR amplification products were checked relative to a 100 bp ladder (Life Technologies,
209 Martinsried, Germany) on a 1 % agarose gel stained with Roti-Safe GelStain (Carl Roth
210 GmbH & Co. KG, Karlsruhe, Germany); afterwards, PCR products were run on an ABI3730

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

211 capillary sequencer. Fragment analysis was performed using GeneMarker™ software v.
212 2.4.0 (Softgenetics, State College, PA, USA).
213
214 *Timber*: For genotyping analysis of timber-derived DNA, mostly a special DNA extraction
215 protocol has been developed and patented (Lowe *et al*. 2015) based on the CTAB method.
216 Exceptionally the innuPREP Plant DNA Kit from Analytik Jena (Germany) was used. Due to
217 the small amount of total DNA extracted from timber, the DNA quantity wasn't measured;
218 rather, a standard dilution of 1:10 (DNA:water) was used for all PCR reactions. PCR
219 conditions were similar to those used for leaf material, but with slight modifications
220 (described in Results).
221
222 For one marker (*trn*DT), a restriction enzyme digest was used to reveal a SNP polymorphism
223 in the amplified fragment. The restriction digestion reaction contained 10µl PCR product, 2µl
224 10x CutSmart® buffer, 0.5µl enzyme (FastDigest® *Hinf*I, New England Biolabs, Ipswich, MA)
225 in a final volume of 20µl. The reaction lasted 15 min at 37°C followed by an inactivation at
226 80°C for 20 min. Restriction products were either visualized relative to a 50 bp ladder (Life
227 Technologies, Germany, Martinsried) using an 8 % polyacrylamide gel stained with ethidium
228 bromide, or using an ABI3730 capillary sequencer.
229
230 **Probabilities for fixation of gene markers**
231 The Thünen-Institute of Forest Genetics possesses a large collection of reference samples
232 that contains oak species from Europe (*Quercus robur*, *Q. petraea*), North America (*Q. alba*,
233 *Q. macrocarpa,* and others) and Asia (*Q. mongolica, Q. dentata*). Based on the numbers in
234 this collection of white oaks from different continents, we computed the maximal potential
235 frequency of variants that were not observed using 95% confidence intervals (Newcombe
236 1998). This can be described as a method to determine the risk (potential error rate) that a
237 genetic variant (allele/haplotype) assumed to be exclusive to one continent is found in one or

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

238 more individuals originating from another continent. The calculations were carried out using
239 the online forma at http://vassarstats.net/prop1.html based on 962 European, 325 Asian and
240 61 American white oak individuals for the gene markers *psa*I-*ycf*4, *psb*E-*pet*L, *trn*LF, and
241 *trn*CD. For the gene marker *trn*DT the sample sizes were 115 European, 425 Asian and 19
242 American white oak individuals.
243
244 **Results**
245 **Next-generation sequencing, reference genome assembly, and identification of cpDNA**
246 **length variants in white oaks**
247 Next-generation sequencing of four indexed oak exemplars (*Q. alba, Q. mongolica*, *Q.*
248 *petraea. Q. robur*) yielded between 1.3 and 1.9 Gbp per individual (Table 1). *De novo* contig
249 assembly with Velvet produced between 872 and 7,085 contigs, and the longest contigs from
250 each assembly were from the chloroplast genome. Our best *de novo* chloroplast assembly
251 derived from *Q. mongolica* QUMO5_CHI_1, which was represented by a single large contig
252 totalling 134.6 kb, and it spanned the three main chloroplast genome regions (large single
253 copy, inverted repeat, small single copy) in their entirety. Alignment of these contigs against
254 the published *Q. rubra* chloroplast genome yielded an alignment of 136,475 nucleotides
255 (alignment excludes one inverted repeat). Parsimony analysis of ordered *de novo* contigs
256 and *Q. rubra* as an outgroup produced a topology similar to the topology previously
257 established with chloroplast DNA restriction site analysis (Manos *et al.* 1999; Fig. 1), with the
258 Old World white oaks *Q. mongolica*, *Q. petraea* and *Q. robur* resolving as a sister group to
259 the New World *Q. alba*. Within this white oak chloroplast genome reference data set, we
260 identified 672 single nucleotide variants and 189 indels (Fig. 1).
261
262 Next-generation sequencing of the *Q. mongolica* and *Q. robur* DNA pool produced 17.4 and
263 28.4 million paired-end 300 bp sequences (10.4 and 17.1 Gbp of sequence data
264 respectively; Table 1). Mapping of paired-end reads from the *Q. mongolica* and *Q. robur*

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

265 pools against the *Q. mongolica* QUMO5_CHI_1 chloroplast genome reference revealed 346
266 variant positions for the *Q. robur* pool, and 789 variant positions for the *Q. mongolica* pool
267 (Table 1). After read mapping, the two variant tables were compared to filter those variants
268 that showed fixed differences between the continents. Because the aim of this study was to
269 develop markers that differentiate between the continents, the next step was to reduce the
270 dataset to these variants appearing with a frequency between 95 and 100%. This analysis
271 left five indels and 15 SNPs. For marker development, we focused in indels due to their
272 simplicity of analysis. Checking of these indels within the mapping revealed a $(T)_N$
273 microsatellite, two indels with a difference in the fragment length of two bp, and one indel
274 with one bp difference. These were removed from the further analyses and only the two
275 longest indels in two spacer regions (*psa*I-*ycf*4, *psb*E-*pet*L), one with four and one with six bp
276 difference, remained.
277
278 Short read sequences from eight North American species were also mapped to the QUMO5
279 reference, and we specifically searched for indels differentiating North American species
280 from the reference QUMO5 with a frequency of 95 to 100%. We identified three indels that
281 consistently differentiated North America from Asia. These length polymorphisms were found
282 in three spacer regions (*trn*LF, *trn*CD, and *trn*DT) and they ranged in length from 2 bp to 8
283 bp.
284
285 **Primer design, marker validation and resequencing**
286 For the five indel-including cpDNA regions, primers were designed using the reference
287 QUMO5 to amplify fragments ranging from 110 bp to 190 bp. A preliminary validation
288 performed with three individuals each of *Q. robur, Q. petraea, Q. mongolica, Q. alba*, and *Q.*
289 *macrocarpa* revealed that all seven cpDNA regions could successfully be amplified by PCR.
290 Subsequent Sanger sequencing validated the sequence of the intervening region, the repeat
291 type, and BLASTN analysis confirmed annotations (Table 2).

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

292

293 Two indels differentiate European (*Q. petraea, Q. robur*) and Asian (*Q. mongolica*) white
294 oaks, and these are located in the *psa*I-*ycf*4 linker (4 bp difference) and the *psb*E-*pet*L linker
295 (6 bp difference; Table 2), and these are inferred as mutations (deletions, specifically) are
296 restricted to Asian white oaks. The further validation of these indels was conducted by
297 screening the amplification products of 10 additional individuals of North American species,
298 50 individuals of two European species (*Q. robur*, *Q. petraea*), and two Asian species (*Q.*
299 *mongolica*, *Q. dentata*). This validation revealed that white oaks from North America and
300 Europe showed the same fragment length (Table 3), and confirmed that the Asian species
301 shared deletions that resulted in shorter, diagnostic fragment lengths.
302 Three indels differentiate Old World (*Q. petraea, Q. robur*, *Q. mongolica*) and North
303 American white oaks, and these are located in the *trn*L-*trn*F (*trn*LF) linker (5 bp difference),
304 the *trn*C-*pet*N linker of the broader *trn*C-*trn*D (*trn*CD) linker region (8 bp) and in the *trn*E-*trn*T
305 linker of the broader *trn*D-*trn*T (*trn*DT) linker region (2 bp; Table 2). These indels were
306 validated with the same individuals of all above mentioned species. The validation revealed
307 that white oaks from Asia and Europe showed identical fragment lengths (Table 3), and that
308 the North American species shared mutations that resulted in diagnostic fragment lengths.
309

310 Resequencing (Sanger) of the *trn*DT region identified a single nucleotide polymorphism
311 (SNP) differentiating Asia from Europe and North America. The SNP lies within a *Hinf*I
312 restriction site, and restriction digestion of this region was predicted to yield three fragments
313 in Asian white oaks and two fragments in European and North American species (all oaks
314 share one *Hinf*I site). By labeling the forward and reverse amplification primer, restriction
315 digestion of the PCR fragment with *Hinf*I allows the visualization of two of the three
316 fragments, one of which is diagnostic for Asian white oaks due to its truncated length. Since
317 this single region offers the possibility to differentiate all three continents with one assay, we
318 decided to include this SNP into the marker set. In total, the five markers have been

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

319 evaluated with samples sizes ranging from 516 to 1078, with samples representing 13 oak
320 species from the three continents (Table 3).
321
322 The nucleotide variation shown to be characteristic for Asian white oaks (*Q. mongolica* and
323 *Q. dentata*) are also present in the complete chloroplast genome of one individual of *Q.*
324 *aliena*, another Asian white oak species (GenBank accession KP301144.1; Lu *et al.* 2015).
325 We computed the probabilities of fixation for the gene markers. Since we require a minimum
326 of two independent markers for continent assignment, we performed the calculations for the
327 risk of not identifying a rare variant in the reference samples based on a combination of two
328 markers. By this means the risk of not identifying a rare variant in the reference samples was
329 calculated to be less than 0.022 % for Europe, 0.0051 % for Asia and less than 0.098 % for
330 America. Thus, it is extremely unlikely that studied gene markers are not fixed to just one
331 variant in the different groups.
332
333 **Marker set design and optimization for timber**
334 All above described analyses were performed using single PCR reactions and fragment
335 analyses to optimize each marker separately. Subsequently, the markers were successfully
336 multiplexed for fragment analysis using the fluorescence labeling as given in Table 2 (Fig. 2).
337 For the development of the markers and multiplexing, DNA from fresh leaves was used. The
338 protocol was later optimized for DNA from timber. From our experience, DNA from timber is
339 more sensitive to all PCR parameters, thus, all markers were singly tested with DNA from
340 timber and the PCR was optimized for the DNA from timber. Differences in the PCR
341 conditions used for the two different materials are given in Table 4. Due to the sensitivity of
342 timber DNA in PCR, multiplexing of the PCRs of the five markers is not advisable, but
343 multiplexing of the markers on the sequencer worked as well with timber DNA as with DNA
344 from leaves. The only difference is that the PCR product from leaf DNA is diluted 1:50 and
345 the PCR product from timber DNA 1:10 for use on the sequencer.

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

346

347 Our analyses used a capillary sequencer to visualize length polymorphisms of these
348 fragments. However, due to the large size differences of these indels, all markers can be
349 distinguished on a polyacrylamide gel, even for differences as small as two base pairs, as
350 shown for the fragment *trn*DT (Fig. 3). In this way, polymorphisms can be screened in
351 laboratories where no sequencer is available.

352

353 The functionality of the optimized PCR protocols and the multiplexed sequencer runs has
354 been tested by means of orders worked on in the "Thünen Centre of Competence on the
355 Origin of Timber". Timber from these orders included highly processed wood as flooring or
356 parquet, as well as treated solid wood samples as different parts of furniture, barrels or
357 boards, and unused solid wood as firewood. In total, over 80 processed timber samples and
358 130 treated solid wood samples have been evaluated (data not shown). Based on our
359 experiences so far we have a success rate of sufficient DNA amplification for our gene
360 markers for 58 % for solid wood samples.

361

362 **Discussion**

363 A set of five chloroplast markers have been developed and optimized to analyze DNA from
364 timber to identify the continental origin of white oak wood. Small fragment sizes (< 200 bp)
365 were chosen because genotyping success with DNA from timber is highest when fragments
366 under 200 bp are targeted. This has recently also been shown for DNA from old and dried
367 insect specimens of museum collections when using mitochondrial barcoding regions
368 (Mitchell 2015). For the identification of haplotypes within oak species from wood samples,
369 Deguilloux *et al.* (2003) similarly developed chloroplast microsatellite and SNP markers that
370 targeted small DNA fragments.

371

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

372 The sequencing data revealed no specific indels to differentiate oak species within the
373 classical barcoding regions *mat*K, *rbc*L or the linker *trn*H-*psb*A (e.g., Chase *et al.* 2005; 2007;
374 Lahaye *et al.* 2008; Hollingsworth 2011; Tripathi *et al.* 2013). Recently, the barcode regions
375 *mat*K and *trn*H-*psb*A were evaluated for their power to discriminate select species from oak
376 sections *Cerris*, *Heterobalanus* (= "Group Ilex"), *Lobatae*, and *Quercus* (Simeone *et al.*
377 2013). In this study, the *mat*K region proved to have too low resolution for the differentiation
378 within the genus; interestingly, for *trn*H-*psb*A, the variability was too high to identify fixed
379 interspecific differences (Simeone *et al.* 2013).
380
381 The intergenic linkers *trn*LF, *trn*CD and *trn*DT we found valuable in this work have been
382 widely tested in population and evolutionary genetic studies of plants, and they show wide
383 variation in their ability to discriminate species and lineages (Shaw *et al.* 2005). For example,
384 the intergenic linker *trn*LF proved to be not variable enough for overall barcoding approaches
385 (Chase *et al.* 2005). For the trial of phylogenetic reconstructions this *trn*LF linker lacked
386 variation in closely related species (Dong *et al.* 2012). Similarly, differentiation within the
387 genus *Populus* failed using *trn*LF (Schroeder *et al.* 2012). Nevertheless, there are other
388 examples for successful use of this marker in molecular systematics (Martin-Bravo *et al.*
389 2007; Pirie *et al.* 2007 and citations therein) and for unraveling of the phylogeny of different
390 plant species (Drábková *et al.* 2004). Similarly, *trn*CD and *trn*DT have been used in
391 comprehensive studies of chloroplast DNA diversity in European white oaks (Bordács *et al.*
392 2002; Cottrell *et al.* 2002; Csaikl *et al.* 2002; Fineschi *et al.* 2002; Jensen *et al.* 2002; König
393 *et al.* 2002; Olalde *et al.* 2002; Petit *et al.* 2002a,b,c). In Japan, four oak species have been
394 differentiated using *trn*DT among other chloroplast markers (Kanno *et al.* 2004).Hence, as for
395 many regions within the chloroplast, the applicability of these spacers to questions of species
396 identity depends on the specific phylogenetic and geographic context where they are used.
397

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

398 Forensic applications of molecular markers are already established with regard to illegal
399 wildlife trade (e.g. parrots: Goncalves *et al.* 2015; sea turtles: Rehman *et al.* 2015) or for
400 identification of products made of endangered animal species (e.g. 'whale meat': Baker *et al.*
401 2010; horn: Yan *et al.* 2013). The barcode of wildlife project
402 (http://www.barcodeofwildlife.org/) has been originated especially for this purpose. Further on
403 control of the seafood market in different countries is well supported by usage of barcoding
404 markers (e.g. Pappalardo & Ferrito 2015; Carvalho *et al.* 2015). For identification of illegal
405 logging of tropic tree species the use of molecular markers is already widespread (e.g.
406 Höltken *et al.* 2012, Nithaniyal *et al.* 2014, Hartvig *et al.* 2015), but the methods are less
407 established for tree species from temperate zones. Thus, the presented markers should be
408 applied to give commercial vendors of white oak wood the possibility to exercise 'due
409 diligence' when placing timber on the European market and the public authorities to control
410 timber imports should questions emerge on the correct declaration of wood.
411
412 **Acknowledgements**
413 We are grateful for the financial support by German Federal Ministry of Food and Agriculture
414 (BMEL), the Deutsche Bundesstiftung Umwelt (DBU) and the USDA Forest Service
415 International Programs Office. We thank Lasse Schindler for developing of the timber specific
416 DNA extraction protocol, and Aki M. Höltken for initial discussions. We are grateful to Vivian
417 Kuhlenkamp, Laura Schulz, Susanne Jelkmann and Ann-Christine Bergmann for technical
418 assistance. We also thank collaborators and institutions that provided samples of North
419 American oaks, including the Morton Arboretum, Lisle, IL (Andrew Hipp, Marlene Hahn),
420 North Carolina State University, Raleigh, NC (Paul Manos), and the USDA Forest Service
421 (Paul Berrang, Andy Bower; Britton Flash; Steven Forry; Ben Gombash; Mark Twery).
422 Finally, we thank Mark Dasenko, Matthew Peterson, Chris Sullivan (Oregon State University
423 Center for Genome Research and Biocomputing) for assistance with sequencing, and Brad

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

424 Kinder, Alex Moad, and Shelley Gardner (US Forest Service) for project and planning
425 assistance.

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

426 **References**

427 Baker CS, Steel D, Choi Y *et al.* (2010) Genetic evidence of illegal trade in protected whales
428 links Japan with the US and South Korea. *Biology Letters*, **6**, 647-650.
429 Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence
430 data. *Bioinformatics*, **30 (15)**, 2114-2120.
431 Bordács S, Popescuc F, Sladed D *et al.* (2002) Chloroplast DNA variation of white oaks in
432 northern Balkans and in the Carpathian Basin. *Forest Ecology and Management*, **156**,
433 197-209.
434 Cassens D (2007) *White oak*. In: Hardwood lumber and veneer series, Purdue Extension,
435 FNR-292-W.
436 Carvalho DC, Palhares RM, Goncalves Drummond M, Frigo TB (2015) DNA Barcoding
437 identification of commercialized seafood in South Brazil: A governmental regulatory
438 forensic program. *Food Control*, **50**, 784-788.
439 Chase MW, Salamin N, Wilkinson M *et al.* (2005) Land plants and DNA barcodes: short-term
440 and long-term goals. *Philosophical Transactions of the Royal Society of London, Series B*,
441 **360**, 1889-1895.
442 Chase MW, Cowan RS, Hollingsworth PM *et al.* (2007) A proposal for a standardised
443 protocol to barcode all land plants. *Taxon*, **56 (2)**, 295-299.
444 Cottrell JE, Munro RC, Tabbener HE *et al.* (2002) Distribution of chloroplast DNA variation in
445 British oaks (*Quercus robur* and *Q. petraea*): the influence of postglacial colonisation and
446 human management. *Forest Ecology and Management*, **156**, 181-195.
447 Crusoe MR, Alameldin HF, Awad S *et al.* (2015) The khmer software package: enabling
448 efficient nucleotide sequence analysis. *F1000 Research*, **4**, 900.
449 Csaikl UM, Burg K, Fineschi S, König AO, Mátyás C, Petit RJ (2002) Chloroplast DNA
450 variation of white oaks in the alpine region. *Forest Ecology and Management*, **156**, 131-
451 145.
452 Drábková L, Kirschner J, Vlcek C, Paces V (2004) *Trn*L–*trn*F intergenic spacer and *trn*L
453 intron define major clades within *Luzula* and *Juncus* (Juncaceae): Importance of structural
454 mutations. *Journal of Molecular Evolution*, **59**, 1-10.
455 Deguilloux MF, Pemonge MH, Bertel L, Kremer A, Petit RJ (2003) Checking the geographical
456 origin of oak wood: molecular and statistical tools. *Molecular Ecology*, **12**, 1629-1636.
457 Dong W Liu J, Yu J, Wang L, Zhou S (2012) Highly variable chloroplast markers for
458 evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLoS One*, **7**
459 **(4)**, e35071.
460 Dormontt EE, Boner M, Braun B *et al.* (2015) Forensic timber identification: It's time to
461 integrate disciplines to combat illegal logging. *Biological Conservation,* **191**, 790-798.
462 Dumolin S, Demesure B, Petit RJ (1995) Inheritance of chloroplast and mitochondrial
463 genomes in pedunculate oak investigated with an efficient PCR method. *Theoretical and*
464 *Applied Genetics,* **91**, 1253-1256.
465 Fineschi S, Taurchini D, Grossoni P, Petit RJ, Vendramin GG (2002) Chloroplast DNA
466 variation of white oaks in Italy. *Forest Ecology and Management*, **156**, 103-114.
467 Goncalves PFM, Oliveira-Marques AR, Matsumoto TE, Miyaki CY (2015) DNA Barcoding
468 identifies illegal parrot trade. *Journal of Heredity*, **106**, 560-564.
469 Handy SM, Parks MB, Deeds JR *et al.* (2011) Use of the chloroplast gene *ycf1* for the
470 genetic differentiation of pine nuts obtained from consumers experiencing dysgeusia.
471 *Journal of Agricultural and Food Chemistry*, **59,** 10995-11002.
472 Hartvig I, Czako M, Kjær ED, Nielsen LR, Theilade I (2015) The use of DNA barcoding in
473 identification and conservation of rosewood (*Dalbergia* spp.). *PLoS One*, **10 (9)**,
474 e0138231.
475 Höltken AM, Schröder H, Wischnewski N, Degen B, Magel E, Fladung M (2012)
476 Development of DNA-based methods to identify CITES-protected timber species: A case
477 study in the Meliaceae family. *Holzforschung*, **66**, 97-104.

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

478 Hollingsworth PM, Forrest LL, Spouge JL (2009). A DNA barcode for landplants.
479 *Proceedings of the National Academy of Science USA,* **106**, 12794-12797.
480 Hollingsworth PM (2011) Refining the DNA barcode for landplants. *Proceedings of the*
481 *National Academy of Science USA,* **108 (49)**, 19451-19452.
482 Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode.
483 *PLoS ONE,* **6 (5)**, e19254.
484 Huang X-C, Xi X-Q, Conran JG, Li J (2015) Application of DNA barcodes in Asian tropical
485 trees – a case study from Xishuangbanna nature reserve, Southwest China. *PLOS one*,
486 **10 (6)**, e0129295.
487 Janzen DH, CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proceedings*
488 *of the National Academy of Science USA,* **106 (31)**, 12794-12797.
489 Jensen JS, Gillies A, Csaikl U *et al.* (2002). Chloroplast DNA variation within the Nordic
490 countries. *Forest Ecology and Management*, **156**, 167-180.
491 Kanno M, Yokoyama J, Suyama Y, Ohyama M, Itoh T, Suzuki M (2004) Geographical
492 distribution of two haplotypes of chloroplast DNA in four oak species (*Quercus*) in Japan.
493 *Journal of Plant Research*, **117**, 311-317.
494 Kent WJ (2002) BLAT - the BLAST-like alignment tool. *Genome Research,* **12**, 656–664.
495 König AO, Ziegenhagen B, van Damb BC *et al.* (2002) Chloroplast DNA variation of oaks in
496 western Central Europe and genetic consequences of human influences. *Forest Ecology*
497 *and Management*, **156**, 147-166.
498 Kress WJ, Erickson DL (2007) A two-locus DNA barcode for plants: the coding region rbcL
499 gene complements the non-coding *trn*H-*psb*A spacer regions. *PLOS one*, **6**, e508.
500 Kress WJ, Erickson DL (2008) DNA barcodes: Genes, Genomics and bioinformatics.
501 *Proceedings of the National Academy of Science USA,* **105 (8)**, 2761-2762.
502 Lahaye R, Van der Bank M, Bogarin D *et al.* (2008) DNA barcoding the floras of biodiversity
503 hotspots. *Proceedings of the National Academy of Science USA,* **105 (8)**, 2923-2928.
504 Lowe AJ, Jardine DI, Cross HB, Degen B, Schindler L, Hoeltken AM (2015) A method of
505 extracting plant nucleic acids from lignified plant tissue. Patent WO/2015/070279 filled
506 2014-11-14 and issued 2015-05-21,
507 Lu S, Hou M, Du FK, Li J, Yin K (2015) Complete chloroplast genome of the Oriental white
508 oak: *Quercus aliena* Blume. *Mitochondrial DNA*, **26**, 1-3.
509 Manos PS, Doyle JJ, Nixon KC (1999) Phylogeny, biogeography, and processes of
510 molecular differentiation in *Quercus* subgenus *Quercus* (Fagaceae). *Molecular*
511 *Phylogenetics and Evolution*, **12 (3)**, 333-349.
512 Martin-Bravo S, Meimberg M, Luceño M *et al.* (2007) Molecular systematics and
513 biogeography of Resedaceae based on ITS and *trn*L-F sequences. *Molecular*
514 *Phylogenetics and Evolution*, **44**, 1105-1120.
515 Mitchell A (2015) Collection in collections: a PCR strategy and primer set for DNA barcoding
516 for decades-old dried museum specimens. *Molecular Ecology Resources*, **15**, 1102-1111.
517 Newcombe RG (1998) Two-Sided Confidence Intervals for the Single Proportion:
518 Comparison of Seven Methods. *Statistics in Medicine*, **17**, 857-872.
519 Newmaster SG, Fazekas AJ, Ragupathy S (2006) DNA barcoding in the land plants:
520 valuation of *rbcL* in a multigene tiered approach. *Canadian Journal of Botany*, **84**, 335–
521 341.
522 Nithaniyal S, Newmaster SG, Ragupathy S, Krishnamoorthy D, Vassou SL, Parani M (2014)
523 DNA barcode authentication of wood samples of threatened and commercial timber trees
524 within the tropical dry evergreen forest of India. *PLoS One*, **9 (9)**, e107669.
525 Olalde M, Herrán A, Espinel S, Goicoechea PG (2002) White oaks phylogeography in the
526 Iberian Peninsula. *Forest Ecology and Management*, **156**, 89-102.
527 Pappalardo AM, Ferrito V (2015) DNA barcoding species identification unveils mislabeling of
528 processed flatfish products in southern Italy markets. *Fisheries Research*, **164**, 153-158.
529 Parks M, Cronn R, Liston A (2009) Increasing phylogenetic resolution at low taxonomic
530 levels using massively parallel sequencing of chloroplast genomes. *BMC Biology*, **7**, 84.

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

531 Parks M, Cronn R, Liston A (2012) Separating the wheat from the chaff: mitigating the effects
532 of noise in a plastome phylogenomic data set from *Pinus* L. (Pinaceae). *BMC Evolutionary*
533 *Biology,* **12**, 100.
534 Petit RJ, Latouche-Hallé C, Pemonge MH, Kremer A (2002a) Chloroplast DNA variation of
535 oaks in France and the influence of forest fragmentation on genetic diversity. *Forest*
536 *Ecology and Management*, **156**, 115-129.
537 Petit RJ, Csaikl UM, Bordács S *et al.* (2002b) Chloroplast DNA variation in European white
538 oaks phylogeography and patterns of diversity based on data from over 2600 populations.
539 *Forest Ecology and Management*, **156**, 5-26.
540 Petit RJ, Brewer S, Bordács S *et al.* (2002c) Identification of refugia and post-glacial
541 colonisation routes of European white oaks based on chloroplast DNA and fossil pollen
542 evidence. *Forest Ecology and Management*, **156**, 49-74.
543 Pirie MD, Balca MP, Vargas Z, Botermans M, Bakker FT, Chatrou LW (2007) Ancient
544 paralogy in the cpDNA *trn*L-F region in Annonaceae: Implications for plant molecular
545 systematics. *American Journal of Botany*, **94 (6)**, 1003-1016.
546 Rehman A, Jafar S, Raja NA, Mahar J (2015) Use of DNA Barcoding to control the Illegal
547 Wildlife Trade: A CITES case report from Pakistan. *Journal of Bioresource Management*,
548 **2 (2)**, 19-22.
549 Schroeder H, Hoeltken AM, Fladung M (2012) Differentiation of *Populus* species using
550 chloroplast SNP-markers - essential for comprehensible and reliable poplar breeding.
551 *Plant Biology*, **14**, 374-381.
552 Shaw J, Lickey EB, Beck JT *et al.* (2005) The tortoise and the hare II: relative utility of 21
553 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of*
554 *Botany*, **92 (1)**, 142-166.
555 Simeone MC, Piredda R, Papini A, Vessella F, Schirone B (2013) Application of plastid and
556 nuclear markers to DNA barcoding of Euro-Mediterranean oaks (*Quercus*, Fagaceae):
557 problems, prospects and phylogenetic implications. *Botanical Journal of the Linnean*
558 *Society,* **172 (4)**, 478-499.
559 Straub SC, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A (2012) Navigating the tip
560 of the genomic iceberg: Next-generation sequencing for plant systematics. *American*
561 *Journal of Botany*, **99 (2)**, 349-64.
562 Taberlet P, Coissac E, Pompanon F *et al.* (2007) Power and limitations of the chloroplast
563 *trn*L (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, **35**, e14.
564 Tripathi AM, Tyagi A, Kumar A *et al.* (2013) The Internal Transcribed Spacer (ITS) region
565 and *trn*H-*psb*A are suitable candidate loci for DNA barcoding of tropical trees species from
566 India. *PLoS One*, **8 (2)**, e57934.
567 US Department of Justice (2015) Statement of facts, United States of American v. Lumber
568 Liquidators. Case document 2:15-cr-00126-RAJ-LRL. Available online:
569 http://www.justice.gov/opa/file/787141/download.
570 Yan D, Luo JY, Han YM *et al.* (2013) Forensic DNA barcoding and Bio-Response studies of
571 animal horn products used in traditional medicine. *PLoS One*, **8 (2)**, e55854.
572 Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de
573 Bruijn graphs. *Genome Research,* **18 (5)**, 821-829.
574
575
576

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

577 **Data Accessibility**
578 DNA sequences from Sanger sequencing of the indel containing fragments derived from this
579 study have been deposited in GenBank (Accession numbers: KU201020-KU201034).
580 Next-generation sequences supporting this study are available from the NCBI GenBank as
581 BioProject PRJNA269970 (SRA Accessions SRS954648, SRS954649, and SRS954650).
582 SRA accession numbers of the next-generation sequencing data of the two DNA-pools are
583 coming soon. Draft chloroplast genome references will be submitted to Dryad soon.
584 Genotype calls for samples will be available on the EVOLTREE website
585 (http://www.evoltree.eu/index.php/e-recources/databases) soon.
586 The open source software *Variant Tools* is available on
587 https://github.com/ThuenenFG/varianttools.
588
589 **Author Contributions**
590 Y.Y. coordinated sample collections in Russia and China, Y.Y. and H.S. coordinated sample
591 collections in Europe, and R.C. coordinated sample collections in North America. H.S. and
592 B.K. performed the analysis of sequence pools, identified variants, and designed indel
593 screens. H.S. coordinated the Sanger sequencing and the genotyping. R.C. and T.J.
594 constructed libraries for North American species, constructed de novo genome references,
595 and assisted with variant analysis. B.D. did statistical calculations and together with R.C.
596 initiated the project. M.M. developed the Variant Tools. H.S., B.K., B.D. and R.C. were
597 involved in overall coordination. H.S. wrote the manuscript and all authors contributed to and
598 approved the final manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

599 **Tables and Figures** (with captions)

600

601 **Table 1:** Results of next-generation sequencing for oak reference assembly and
602 polymorphism screening. Indexed individuals of oaks were sequenced using 150 bp
603 paired-end reads and evaluated using *de novo* assembly (reference assembly). Pooled
604 individuals were sequenced using 300 bp paired-end reads and evaluated using
605 reference-guided assembly (polymorphism screen).

606

	Reference Assembly				Polymorphism Screen	
	QUMO5	QUPE2	QURO2[1]	QUAL	QUMON Pool	QUROB Pool
Sequences	6,183,339	5,281,517	6,676,925	4,280,049	17,370,324	28,438,226
Bases sequenced (Gbp)	1.747	1.502	0.977	1.207	10.42	17.06
Proportion of reads mapping to cp genome	3.3	1.9	7.6	0.4	0.95	2.25
De novo contigs (length sum, Mbp)	2,085	872	80	7,085	--	--
Longest contig	135,603	60,701	27,043	57,523	--	--
Length chloroplast contigs	4	12	19	20	--	--
Variants relative to QUMO5	0	126	124	572	789	346

607

608 [1] *Q. robur* was sequenced using 100 bp single-end reads.

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

609 **Table 2:** List of primers for the amplification and resequencing of the newly developed
610 markers. Fluorescent-labeling of the primers is given in column "sequences": FAM = blue,
611 VIC = green, PET = red. In the last column, the accession numbers of the related markers
612 for the three species *Q. robur*, *Q. mongolica* and *Q. alba* are given. "Length" means
613 sequence length.
614

Marker-name	Primer	Sequence 5'-3'	length (bp)	Annea-ling (°C)	Acces. no.
*psa*I-*ycf*4	psaI ycf4	CGTGTGTAAACATGATATATGAG_FAM GAGTAATTCATCGAATTGGTTAG	174-178	55	KU2010 20-22
*psb*E-*pet*L	psbE petL	AAGGAATTGGTTAGTTGTCCAG_VIC TTACATATCTTAAATTAGAGAGCC	179-185	55	KU2010 23-25
*trn*LF	trnL1 trnF2	CAATACATATCATTTCTTGTACTG_PET TAGATAACTTGAGTTTATGTCAATT	130-135	53	KU2010 26-28
*trn*CD	trnC5 trnD5	TTGGATAGACGAACGGGGAAT_FAM TATCATATTAAATTGATTGCCGG	115-123	57	KU2010 29-31
*trn*DT	trnD3 trnT4	GGATAGGGATCAACAAGTTATTG_PET CAAGACCGACCCTAATTGAAT_PET	187-189	52	KU2010 32-34

615

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

616 **Table 3:** Number of individuals per species and continent tested with the five markers,
617 and fragment length based on sequencing for each marker and species.
618

continent	species	[1] No. individuals	*psa*I-*ycf*4	*psb*E-*pet*L	*trn*LF	*trn*CD	[2] *trn*DT
Europe	*Q. robur*	531 (103)	178	185	135	123	86/101
	Q. petraea	273 (12)	178	185	135	123	86/101
	Q. pubescens	158 (0)	178	185	135	123	
USA	*Q. alba*	15 (4)	178	185	130	115	88/101
	Q. macrocarpa	12 (7)	178	185	130	115	88/101
	Q. bicolor	7 (4)	178	185	130	115	88/101
	Q. garryana	4 (0)	178	185	130	115	
	Q. lyrata	4 (2)	178	185	130	115	88/101
	Q. michauxii	5 (1)	178	185	130	115	88/101
	Q. stellata	8 (1)	178	185	130	115	88/101
	Q. prinoides	6 (0)	178	185	130	115	
Asia	*Q. mongolica*	316 (420)	174	179	135	123	86/71 [30]
	Q. dentata	9 (5)	174	179	135	123	86/71 [30]

619
620 [1] The number of individuals tested includes numbers for all loci except *trn*DT, which is
621 given in parentheses.
622 [2] Fragment lengths for *trn*DT show the lengths after restriction digestion with *Hinf*I. Asian
623 species contain an additional internal restriction site that yields one additional unlabeled
624 fragment after *Hinf*I digestion; these are shown in brackets.

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

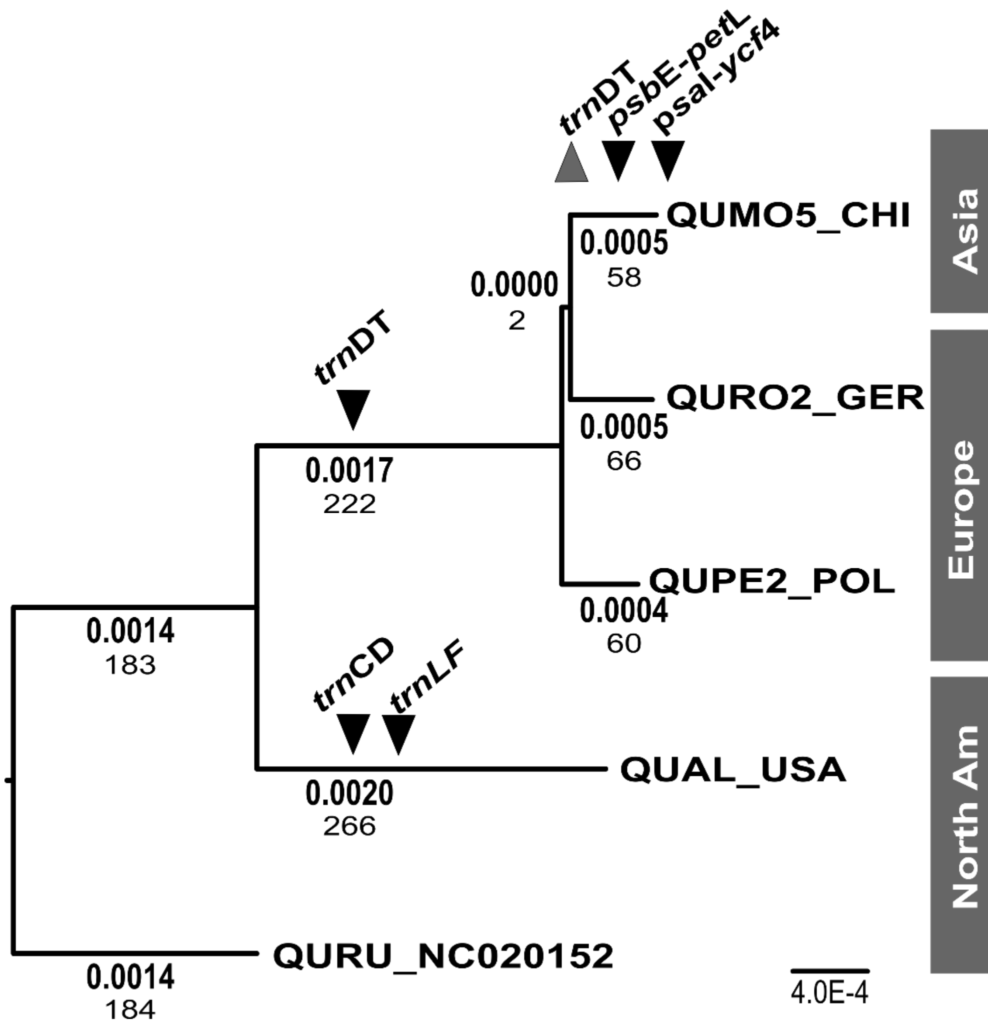
625 **Table 4:** PCR conditions compared for leaf and timber. Only the differences are shown, all
626 other parameters are as given in material and methods.
627

	Leaf				Timber			
	PCR cycles	Conc. $MgCl_2$	Enhan-cer	Conc. Primer	PCR cycles	Conc. $MgCl_2$	Enhan-cer	Conc. Primer
psaI-ycf4	30	1.75 mM	no	0.07 µM	40	2.0 mM	no	0.1 µM
psbE-petL	30	1.75 mM	no	0.05 µM	40	2.0 mM	no	0.05 µM
trnLF	30	1.75 mM	yes	0.2 µM	40	2.0 mM	yes	0.2 µM
trnCD	25-30	1.5 mM	yes	0.13 µM	35	2.5 mM	yes	0.3 µM
trnDT	25	1.75 mM	yes	0.1 µM	40	2.0 mM	yes	0.1 µM

628

bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

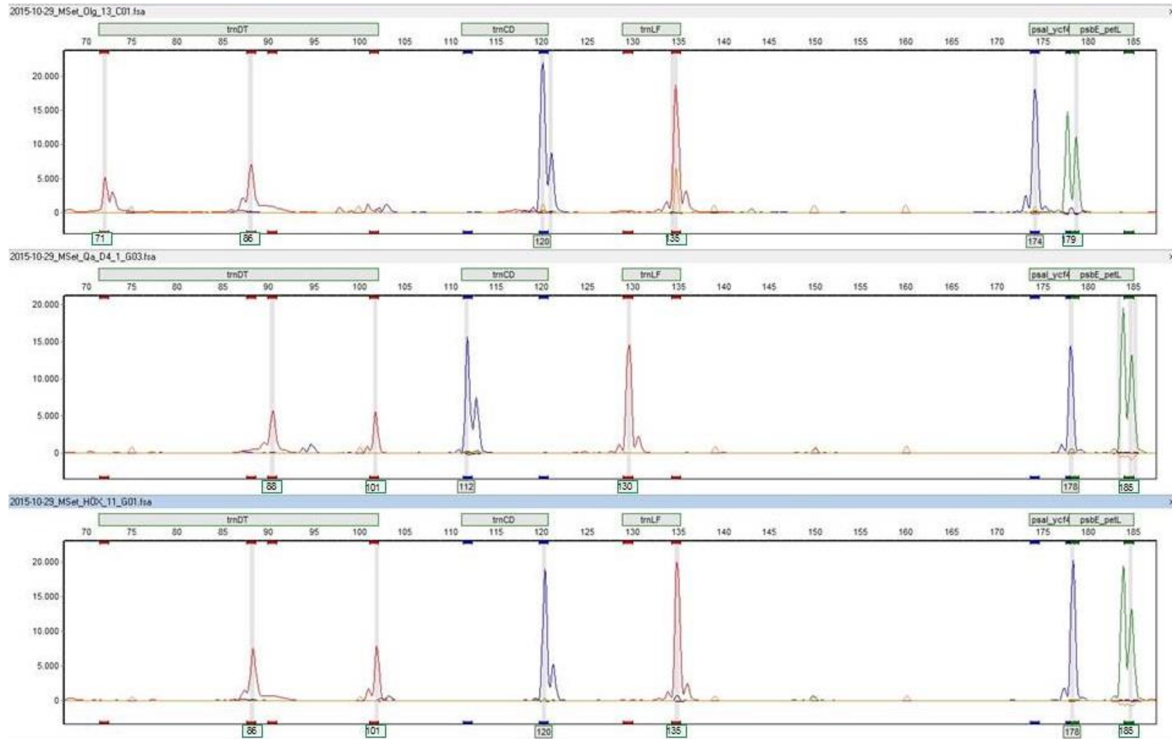
629 **Fig. 1:** Phylogenetic relationship among chloroplast genomes of white oak species
630 representing Old World and New World lineages. The best maximum likelihood tree is
631 shown for four white oak chloroplast genomes (*Q. mongolica*; *Q. robur*; *Q. petraea*; *Q.*
632 *alba*) and one outgroup genome (*Q. rubra*). Inferred branch lengths in maximum likelihood
633 substitutions are shown in bold; the number of substitutions inferred from parsimony
634 shows these values. The phylogenetic resolution of informative indel markers are shown
635 in black inverted triangles, and the resolution of the diagnostic PCR-RFLP marker is
636 shown as a grey triangle.
637



638

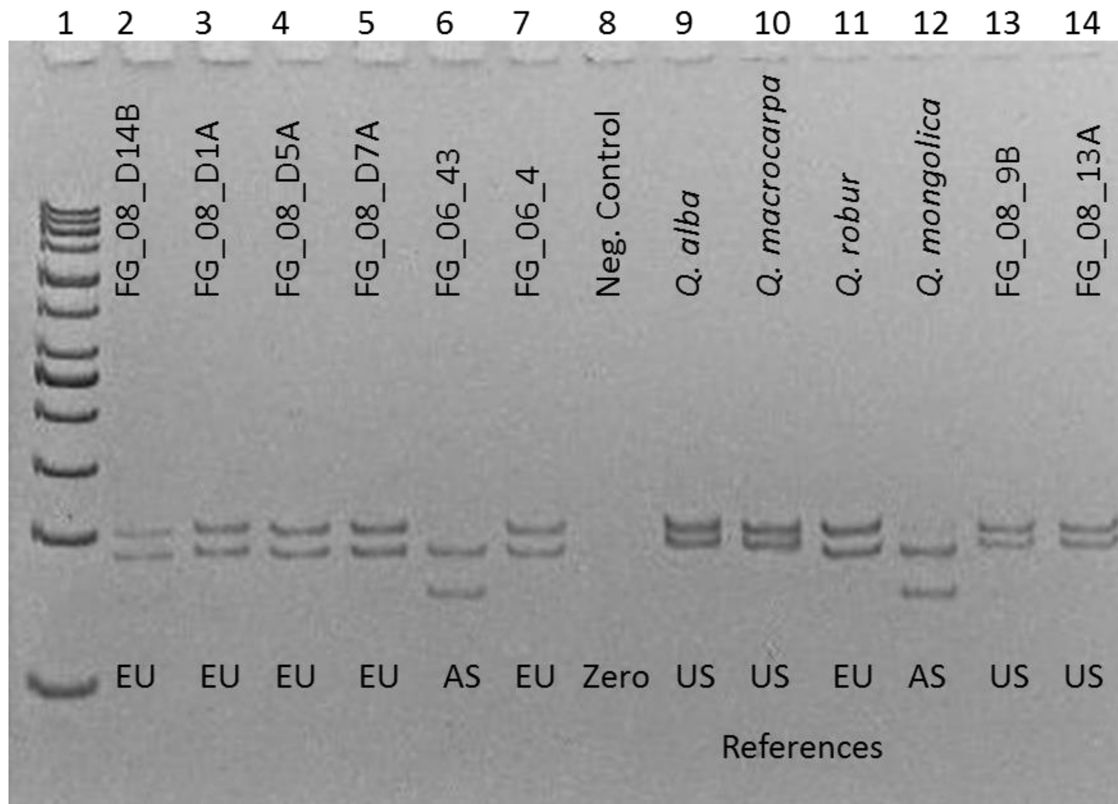
bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

639 **Fig. 2:** Fragment patterns of the five markers for individuals from Asia (top), North
640 America (middle) and Europe (bottom). The sequence sizes for each peak as given in
641 Table 3 are shown beneath the peaks. The first blue peaks appear smaller (112, 120)
642 than the sequenced length (115, 123) given in Table 3. The color code of the peaks is as
643 described in Table 2.
644



bioRxiv preprint doi: https://doi.org/10.1101/038562; this version posted February 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

645 **Fig. 3:** Marker *trn*DT visualized on a polyacrylamide gel. Lane 1: 50 bp ladder, lane 8:
646 zero control, lane 2-7 and 13-14: analysis of wood-derived DNA, its location is inferred
647 from genotypes, lane 9-12: references from North America (US), Europe (EU) or Asia
648 (AS), respectively.



649