High-throughput sequencing of 5S-IGS in oaks - exploring intragenomic variation and algorithms to recognize target species in pure and mixed samples.

Roberta Piredda[1], Guido Grimm[2], Ernst-Detlef Schulze[3], Thomas Denk[4], and Marco Simeone[5]

[1]Stazione Zoologica Anton Dohrn
[2]University of Vienna
[3]Max Planck Institute for Biogeochemistry
[4]Swedish Museum of Natural History
[5]Università della Tuscia

April 28, 2020

Abstract

Measuring biological diversity is a crucial but difficult undertaking, as exemplified in oaks where complex morphological, ecological, biogeographic and genetic differentiation patterns collide with traditional taxonomy that measures biodiversity in number of species (or higher taxa). In this pilot study, we generated High-Throughput Sequencing (HTS) amplicon data of the intergenic spacer of the 5S nuclear ribosomal DNA cistron (5S-IGS) in oaks, using six mock samples that differ in geographic origin, species composition, and pool complexity. The potential of the marker for automated geno-taxonomy applications was assessed using a reference dataset of 1770 5S-IGS cloned sequences, covering the entire taxonomic breadth and distribution range of western Eurasian Quercus, and applying similarity (BLAST) and evolutionary approaches (ML trees and EPA). Both methods performed equally well, with correct identification of species in sections Ilex and Cerris in the pure and mixed samples and main genotypes shared by species of sect. Quercus. Application of different cut-off thresholds revealed that medium-high abundance sequences (>10 or 25) suffice for a net species identification of samples containing one or few individuals. Lower thresholds identify phylogenetic correspondence with all target species in highly mixed samples (analogue to environmental bulk samples) and include rare variants pointing towards reticulation, incomplete lineage sorting, pseudogenic 5S units, and in-situ (natural) contamination. Our pipeline is highly promising for future assessments of intra-specific and inter-population diversity, and of the genetic resources of natural ecosystems, which are fundamental to empower fast and solid biodiversity conservation programs worldwide.

Introduction

Oaks (*Quercus* L., Fagaceae) are anemophilous trees and shrubs of the Northern Hemisphere and are of high ecological, scientific and economic importance (Menitsky 2005). They occur in a wide range of habitats and climates ranging from semi-arid Mediterranean areas and subtropical rainforests to boreal (cold temperate) continental regions (Kremer and Hipp 2019). Oak taxonomy is still in flow; only recently the genus has been divided into two monophyletic subgenera and eight sections accommodating the 300–600 reported oak species (Denk et al. 2017). Discrepancies in species counts are due to the considerable morphological variation and phenotypic plasticity exhibited by oaks, the result of the complex interplay of strong ecological adaptation, convergence of morphological traits, hybridization, introgression, and reticulate evolution (Burger 1975; Van Valen 1976; Cavender-Bares et al. 2015; Simeone et al. 2016; McVay et al. 2017a, b; Hipp et al. 2019a). Consequently, oaks are a big challenge for modern taxonomists (Denk et al. 2017) and provide an ideal

model for the development of a holistic taxonomy that accounts for complex ecological, biogeographic and evolutionary processes (Kremer et al. 2012), thereby improving our understanding of Earth's biodiversity. At the same time, key questions about oak diversification and evolutionary history, especially in the Old World, are just about to be answered (Denk et al. 2017; Hipp et al. 2019; Yan et al. 2019; Jiang et al. 2019).

After three decades of research (since Whittemore and Schaal 1991), it has been widely demonstrated that oak plastid genomes are decoupled from species identity (Simeone et al. 2016, 2018; Pham et al. 2017; Vitelli et al. 2017) and that nuclear genes and regions with sufficient levels of variation for solid species delineation and phylogenetic inferences are still unavailable (Oh and Manos 2008; Hubert et al. 2014). Based on extensive phylogenomic data, Hipp (2018a) concluded that oaks can be defined as phylogenomic mosaics, meaning that every individual genome actually represents an assembly of different histories reflecting as well selection (ecological adaptation) and origin, or multiple origins (divergence, reticulation, and lineage sorting). Species-level phylogenetic reconstructions are affected by such genomic blending. Cohesive intra-specific gene flow is counterbalanced by local gene flow among individuals and populations in different parts of a species' range, introgression, sorting of ancestral traits and divergence (Eaton et al. 2015; McVay et al. 2017b; Hipp et al. 2018b, 2019b; Crowl et al. 2020; Leroy et al. 2020).

However, detection and distinction of horizontal gene transfer, introgression and random sorting of ancestral genetic variation can be difficult. The labor and costs associated with the necessary comprehensive datasets (both at the inter- and intraspecific level) will likely prevent a wide application of phylogenomic tools beyond species trees and selected case studies (McVay et al. 2017; Hipp et al. 2019; Jiang et al. 2019). Among nuclear DNA target regions of potentially high taxonomic resolution, the spacer regions of the ribosomal DNA (35S and 5S nrDNA) are a quite popular choice for taxonomic studies. These regions have been used for phylogenetic inferences across a broad range of evolutionary lineages, have a high copy repeat number, and universal or specific PCR primers are available allowing efficient amplification in a wide range of taxa (Volkov et al. 2001, 2003; Alvarez and Wendel 2003). In particular, the internal transcribed spacers (ITS1, ITS2) of the 18S-5.8S-25S cistron (35S rDNA) still are the most widely used nuclear markers for phylogeny and systematics, including molecular taxonomy such as DNA barcoding of complex plant groups (Hollingsworth et al. 2011). However, in many groups the ITS1 and ITS2 can be nearly invariable. In western Eurasian oaks, for instance, the nuclear ribosomal 5S DNA intergenic spacer resolved species relationships to a much higher degree than ITS data (Denk and Grimm 2010).

The 5S intergenic spacer

The 5S rRNA genes form gene arrays analogous to the 35S rDNA, localized on one or a few chromosomes, usually afar from the nucleolus organizer regions (NORs) of the 35S rDNA (*http://www.plantrdnadatabase.com/*; see Ribeiro et al. 2011 for selected oak and beech species). In higher plants, the number of repeats per genome ranges from hundreds to thousands (Cloix et al. 2000), and the 5S rRNA coding regions are separated by a generally variable, non-transcribed intergenic spacer (5S-IGS) with high phylogenetic resolution (e.g., Forest et al. 2005; Blattner et al. 2009; Grimm and Denk 2010; Garcia and Kovařik 2013; Mlinarec et al. 2016). Its inter-individual and intra-genomic variability is probably the main reason why it is little used. It cannot be directly sequenced but requires cloning and methodological frameworks that can make use of the often higher intra-genomic than inter-individual diversity (Göker and Grimm 2008; Potts et al. 2014; for an application using 5S-IGS data see e.g. Simeone et al. 2018). Therefore, large numbers of 5S-IGS variants should ideally be cloned and sequenced to identify intra- and inter-array polymorphism, and phylogenetic signals useful for inferring evolutionary patterns (Eidesen et al. 2007). In previous studies, Denk and Grimm (2010) and Simeone et al. (2018) produced a data set with more than 1000 Sanger-sequenced 5S-IGS clones (covering 30 species, nearly 300 individuals) and demonstrated the potential of this marker for the circumscription of most of the investigated oak species, to recognize hybridization, and to infer reticulation/introgression events.

High-Throughput Sequencing

High-Throughput Sequencing (HTS) has enabled sequencing of thousands to millions of sequences at one time and has changed our view on Earth's biodiversity at all organismal levels (Deiner et al. 2017). Because of constantly declining costs, HTS technologies put forth new tools to address questions previously tackled through labor-intensive (and often less efficient) cloning steps, and amplicon sequencing of loci with high information content (target sequencing) is probably the most straightforward application (Ekblom and Galindo 2010). A further key contribution of HTS lies in the possibility of exploring datasets that cover wide taxonomic and/or geographic breadths. Good inter- and intra-specific taxon sampling is usually required to address the processes underlying speciation, diversification, distribution and species assembly, especially when taxonomic uncertainties and high diversity are involved. Hence, the utilization of HTS is especially cost-effective as many individuals can be combined (multiplexed) in the same sequencing run and rare variants can be readily detected (Babik et al. 2009; Glenn 2011).

Studies characterizing the abundance and patterns of intragenomic nrDNA polymorphisms in different organisms are increasing (Stage and Eickbush 2007; Ganley and Kobayashi 2007; Bik et al. 2013; Straub et al. 2012; Mahelka et al. 2013; Wang et al. 2016; Symonová 2019). Determination of the full sequence of the 35S cistrons was successful in different plant groups (e.g. Malè et al. 2014; Turner et al. 2016; Ji et al. 2019), although so far with little use to open phylogenetic questions. This is partly due to limited sampling (usually a single individual per species or higher taxonomic units). Simon et al. (2012) used a deep sequencing approach to detect intragenomic ITS polymorphisms among populations of*Arabidopsis* . However, phylogenetic studies investigating intragenomic nrDNA polymorphism patterns across many species within the same genus are still scarce (e.g. Song et al. 2012; Weitemeier et al. 2015), and the full extent of the divergence of the 5S-IGS intra-genomic variants in plants has not yet been adequately explored (cf. Galián et al. 2014).

In this pilot study, we generated amplicon data of the intergenic spacer of the 5S nuclear ribosomal DNA cistron (5S-IGS) using High-Throughput Sequencing (HTS) from six geographic samples of different composition: pure samples, including only material of a single target species, and mixed samples, including all species found at a certain place. The investigated species cover all common lineages of western Eurasian oaks (sects *Cerris* , *Ilex* and *Quercus*). Amplicon data were analyzed using our clone-sequence data as reference, and the taxonomic resolution of the 5S-IGS region was assessed comparing the performance of similarity algorithm (Basic Local Alignment Search Tool—*BLAST*) and evolutionary approaches (Maximum Likelihood trees—*ML* ; Evolutionary Placement Algorithm—*EPA*).

To our knowledge, this work is the first to thoroughly analyze intragenomic variation of the nuclear ribosomal 5S region in plants (see also Heitkam et al. 2015, who identified and developed probes for 5S genes while screening Illumina-generated sequences for repetitive genome elements). The potential applications of our approach are manifold and span from the delineation of oak species, the assessment of intra- and interspecies diversity, the detection of hybridization/introgression patterns, the identification of cryptic lineages, to gaining genetic insights into the structure, assembly, function and evolution of oak communities.

Materials and Methods

Sampling and lab procedures

Forty-one oak trees belonging to 13 species (three *Quercus*sections) were morphologically identified and collected in the wild. DNA extractions were performed from silica-gel dried leaves with the DNeasy plant minikit (QIAGEN). DNAs were quantified with a NanoDrop spectrophotometer (TermoFisher Scientific). We prepared six artificial samples, consisting of pure and mixed species (Tab. 1), by pooling equal amounts of DNA of every individual up to 20 total ng. Paired-end Illumina sequencing (2×300 bp sequencing) was performed by a commercial provider (LGC Genomics GmbH). The nuclear ribosomal 5S intergenic spacer (5S-IGS) was amplified with the plant-specific primer pair CGTGTTTGGGCGAGAGTAGT (forward) and CTGCGGAGTTCTGATGG (reverse), developed in collaboration with Dr. Berthold Fartmann (LGC Genomics). Raw sequences were deposited in the Sequence Read Archive under BioProject PRJNA611057.

Bioinformatics analyses

We assembled a reference dataset consisting of 1770 5S-IGS sequences covering the taxonomic breadth and distribution range of western Eurasian *Quercus* by combining data from previous works (Denk and Grimm 2010; Simeone et al. 2018). The total dataset was dereplicated to identify and remove identical sequences using mothur v.1.33.0 (Schloss et al. 2009). Unique (nonredundant) sequences were aligned using mafft (Katoh and Standley 2013) and the alignment was manually checked using SeaView v.4.0 (Gouy et al. 2010). Sequence length and percentage of GC content were calculated within jemboss 1.5 (Carver and Bleasby 2003) and plotted using ggplot2 R package (Wickham 2016). A maximum likelihood tree was inferred using RAxML v.8.2.11 (Stamatakis 2014), with the GTR+GAMMA model of nucleotide substitution. The tree was visualized in iTOL (*www.itol.embl.de*) (Letunic and Bork 2019). The database including unique 5S-IGS representative sequences (unaligned fasta file), the corresponding taxonomy file (in Mothur compatible format), and an Excel file (providing correspondence between clone codes, GenBank accessions and geographic origin) is available on FigShare (https://doi.org/10.6084/m9.figshare.12016272.v1).

Illumina paired-end reads were processed using mothur v.1.33.0 (Schloss et al. 2009). Contigs between read pairs were assembled and differences in base calls in the overlapping region were solved using the ΔQ parameter as described in Kozich et al. (2013). Primer sequences were removed, and no ambiguous bases were allowed. The remaining reads were dereplicated and screened for chimeras using UCHIME in *de-novo* mode (Edgar et al. 2011) within mothur. Taxonomic assignment of the dereplicated reads was performed using standalone BLAST in BLAST+ suite (Altschul et al. 1990; Camacho et al. 2009) against the reference dataset. Reads assigned with a query coverage <220 bp and with similarity <90% were removed.

Length and percentage of GC content of the cleaned dataset were calculated within jemboss 1.5 (Carver and Bleasby 2003). Scatter plots between GC content and reads abundance were visualized with ggplot2 R package (Wickham 2016). Within each sample, we generated four subsets based on the abundance of reads (subset reads [?]2; subset reads [?]5; subset reads [?]10; subset reads [?]25). This procedure produced 24 subsets (4 subsets x 6 samples); each of them was aligned with the reference data and used to infer the corresponding tree with RAxML v.8.2.11 (Stamatakis 2014) and the GTR+GAMMA model, then visualized in iTOL to inspect coherence of clades.

Reads with total abundance [?]25 in the total dataset (six samples) were also assigned using Evolutionary Placement Algorithm (EPA; Berger et al 2011) as implemented in RAxML onto the phylogenetic tree inferred from the *Quercus* reference data. For computing the phylogenetic placements, an alignment comprising both the reference and the query reads was performed within mothur (align.seqs). EPA outputs showed multiple possible placement positions with different likelihood weights (probabilities) in all branches of the tree. Placement results (tree.jplace standard placement format) for the total *Quercus*reference dataset and each *Quercus* clade were visualized using iTOL. The relative performances of the taxonomic assignments obtained with BLAST (best hit) and EPA (tree branch placement) were visualized using semicircle donut plots generated with geom_arc_bar() function in ggforce R package (https://ggforce.data-imaginist.com/).

Results

Reference dataset: cloned 5S IGS data covering all four sections of Quercus in western Eurasia

The dereplication step (identification and removal of identical sequences) reduced the total *Quercus* reference dataset (1770 sequences obtained via PCR, cloning and Sanger-sequencing; Denk and Grimm 2010; Simeone et al. 2018) to 1160 representative sequences. Most identical sequences (442) occurred within single individuals or within species. Twenty-two variants, 163 identical sequences, were shared by members of different species, generating ambiguities in the taxonomic assignment. These sequences were flagged as "ambiguous"; they comprised two variants shared by *Q. baloot* with *Q. ilex*and *Q. coccifera* (section *Ilex*), eight variants shared between two or three different members of section *Cerris*(*Q. crenata, Q. suber, Q. cerris, Q. trojana, Q. look, Q. brantii*), and thirteen variants shared by several members of section*Quercus* (details shown in

Supplementary file S1). *Quercus pontica* , the western Eurasian species of disjunct and relict two-species sect. *Ponticae* , is characterized by species-unique 5S-IGS variants.

The length range of the reference sequences was 258–407 bp (289–407 bp in section *Ilex* , 297–397 bp in section *Cerris,* 258–296 bp in section *Quercus* , and 291 bp in *Q. pontica*). Some species exhibited identical length variants (*Q. canariensis* ,*Q. faginea* , *Q. pyrenaica, Q. boissieri, Q. infectoria, Q. pontica*). All the remaining species revealed intragenomic and/or intraspecific variation, with units differing by <10 bp (e.g.,*Q. alnifolia, Q. afares, Q. suber* , *Q. frainetto*), or by 20–30 bp (e.g., *Q. aucheri, Q. coccifera, Q. ilex, Q. brantii, Q. crenata, Q. look, Q. robur, Q. petraea, Q. pubescens*). Two distinct classes of length variants, differing by 40–90 bp, were found in*Q. floribunda, Q. ilex, Q. cerris, Q. trojana, Q. ithaburensis, Q. macrolepis* , and *Q. libani* . Besides unique, unrepresentative variants of *Q. floribunda* and *Q. ilex* , the short variants identified both inter-individual (*Q. libani* : 2 individuals), and intra-individual variation (*Q. cerris* : 3 individuals; *Q. ithaburensis, Q. macrolepis* and *Q. trojana:* 1 individual; cf. Denk and Grimm 2010; Simeone et al. 2018). GC content (P_{GC}) of the reference sequences ranged between 44.1 to 56.9% (median= 53.61%, mean = 53.66 %; SD = 1.25%) (Fig. 1A). Lowest GC contents ($P_{CG} < 49\%$) were found in few sequences of *Q. trojana, Q. cerris, Q. ilex* and *Q. suber* , exhibiting known pseudogenous tendency (Denk and Grimm 2010; Simeone et al. 2018). Only one species, *Q. alnifolia,* a Cypriot endemic of section *Ilex* , was entirely below the mean range of the genus. Outlier sequence variants (1st and last three percentiles) in length and CG content relatively to each species' median were labelled in the downstream analyses. Length and CG content of the total dataset and each species are shown in Fig. 1A, B and Supplementary file S1.

HTS dataset

The total cleaned HTS dataset (six samples) comprised 208,720 reads. The number of reads retained in each pre-processing step, their length, GC content and distribution in the six samples are reported in Supplementary file S2.

The HTS dataset showed a length range of 254-399 bp, and a P_{GC} range of 42.6–64.5%, thus matching those of the reference dataset. Extreme values were exclusive to singleton reads (abundance = 1). The application of different abundance thresholds ([?]2, up to [?]25) did not affect the length range and the occurrence of two clear classes of variants (Supplementary file S2). Reads with abundance [?]2 showed P_{GC} = 48.1–57.1%; excluding those with lower occurrence (abundance <5) had little effect. Filtering for reads with higher abundance increasingly eliminated lower GC values and further approached the mean range of the reference dataset; reads with abundance [?]25 showed a P_{GC} of 52.5–56.4%. Therefore, high abundance cut-offs (i.e., 10 or 25) represent all the major 5S-IGS variants of the species included in the investigated samples and remove rare variants. These latter may include spurious data, pseudogenic variants, potential outlier of biological significance, and inherent biases of the HTS methodology leading to scarce amplification of certain individuals. As such, they should be carefully evaluated.

Reference tree and basic geno-taxonomic capacity of 5S-IGS sequences

The final multiple sequence alignment of the cloned reference data produced a matrix with 492 characters. Figure 2A–D shows the ML tree of the reference dataset, rooted between members of sections*Ponticae* and *Quercus* (subgenus *Quercus*) and sections *Cerris* and *Ilex* (subgenus *Cerris* ; following Denk et al. 2017; Hipp et al. 2019). The main lineages (sections) are clearly differentiated, with *Q. pontica* (sect.*Ponticae*) being embedded in the section *Quercus* subtree (Fig. 2A). Regarding their geno-taxonomic association, i.e. the ability of genetic data to recognize a (morphologically defined) target species, it is important to note that even though species do not form exclusive clades (Fig. 2B-D), except for *Q. alnifolia* (sect. *Ilex*),*Q. afares* (sect. *Cerris*) and *Q. pontica* (sect.*Ponticae*), conspecific 5S-IGS sequences always cluster within certain subtrees. In sect. *Ilex* (Fig. 2B), determination of non- or near-identical sequences via tree-inference is possible down to the species level: *Q. alnifolia* 5S-IGS variants are unique; common variants of *Q. coccifera* and *Q. aucheri* intermix (and may include occasional *Q. ilex* clones) and are separated from a large clade including most *Q. ilex* sequences (cf. Denk and Grimm 2010). In *section* Cerris (Fig. 2C), distinct

subtrees collect accessions of the central-western Mediterranean *Q. suber-Q. crenata* lineage, the eastern Mediterranean subsection *Macrolepides* (*Q. brantii, Q. macrolepis, Q. ithaburensis*), the Anatolian *Q. libani -trojana* lineage, and the mixed 'oriental' and 'occidental' lineages comprising the remaining species (*Q. afares* , *Q. castaneifolia, Q. cerris, Q. look, Q. euboica, Q, trojana* ; cf. Simeone et al. 2018). In section *Quercus* (Fig. 2D), no obvious structure is visible, association of sequences to discrete species is largely impossible. Three distinct subtrees collected consistently only sequences of *Q. pyrenaica-Q. canariensis*(genotype Q1), *Q. robur-Q. dalechampii* (genotype Q2), and*Q. boissieri* (Q3 genotypes). Sequences of *Q. pontica* form an exclusive subtree, corresponding to sect. *Ponticae* . A visual representation of the identified diagnostic groups of variants is presented in Supplementary file S3. The NEWICK format of the RAxML reference tree can be downloaded at https://doi.org/10.6084/m9.figshare.12016317.v1.

Geno-taxonomic composition of samples, cut-off effects

Genetic assessment via BLAST and per-sample ML phylogenetic inferences provided taxonomic patterns congruent with the species composition of each sample (Tab. 1, Supplementary files S2, S4).

Sample D5, pure *Q. afares* —Of the 31,620 HTS sequences retrieved in this sample, 31,121 (98.4%) were assigned by BLAST to *Q. afares* and a few hundreds to members of the *Cerris* crown (*Q. castaneifolia* , *Q. cerris* , *Q. trojana*) or to the same section (*Q. suber* , *Q. brantii*). Negligible amounts of sequences were assigned to a further species of the same section (*Q. macrolepis;* three sequences), or to different sections (18 sequences assigned to members of section *Quercus* and*Ilex*). Using an abundance cut-off [?]25, all HTS sequences grouped within the *Q. afares* reference ML subtree. With cut-off [?]10, the HTS sequences expanded the *Q. afares* -comprising larger subtree of sect. *Cerris* and brushed against a minor clade with *Q. cerris* , *Q. trojana* , *Q. brantii* (Type 4b, collecting likely ancestral, underived sequences; Fig. S3-2 in Supplementary file S3); with cut-off [?]5, the *Q. afares* subtree was inflated further and included a single reference sequence of *Q. trojana* . The main*Cerris* lineages could hardly be differentiated with cut-off [?] 2. A dozen HTS reads joined sequences of the oriental *Q. trojana-Q. libani* , and a few other were highly differentiated and formed a separate cluster (possible pseudogenes).

Sample G2, pure *Q. ilex* —Of the 13,091 HTS sequences produced, nearly all (13,067; 99.8%) were assigned by BLAST to *Q. ilex* , and negligible amounts were assigned to *Q. coccifera* or members of different sections. In the ML analysis, all HTS sequences with abundance [?]5 grouped within a large *Q. ilex* subtree of the reference data; with cut-off [?]2, the HTS sequences grouped across different *Q. ilex* subtrees, and only few sequences placed together with *Q. aucheri* .

Sample H1, pure *Q. faginea* —Of the 58,003 HTS sequences produced, 33,546 (57.8%) were assigned by BLAST to *Q. canariensis* , an Iberian-North African sister species of *Q. faginea* , and 23,770 (41%) were variously assigned to either ambiguous or specific sequences of section *Quercus* , including *Q. faginea, Q. pyrenaica, Q. petraea* , *Q. pubescens* , and (secondarily) *Q. frainetto, Q. vulcanica,* and *Q. robur* . With ML, most HTS sequences with abundance [?]25 and [?]10 grouped within the *Q. pyrenaica-Q. canariensis* subtree referring to the*canariensis-pyrenaica* -unique type Q1, and all other sequences were scattered across the undiagnostic Q0 subtrees, often, but not necessarily, grouping with *Q. faginea* sequences. With the decreasing abundance thresholds (cut-offs = 5 and 2), nearly the entire section *Quercus* was covered by HTS sequences (except the eastern Mediterranean Q3 type; Fig. S3-4 in Supplementary file S3).

Sample F2, mixed, one species per section—Of the 41,527 HTS sequences produced, 1,307, 6,399 and 9,454 (i.e., total of 48.5%) were assigned by BLAST to the target species *Q. suber, Q. ilex* and *Q. canariensis* , respectively. All other sequences were assigned to the*Q. suber-crenata* shared types (1,297), *Q. coccifera* (36), sister of *Q. ilex* , and to ambiguous or specific sequences of section *Quercus* (mainly *Q. faginea, Q. pyrenaica, Q. petraea, pubescens* , and *Q. frainetto*). With ML, dispersion of HTS sequences onto the reference tree increased with decreasing abundance thresholds; however, a *Q. ilex* subtree was always identified, and placements outside the *Q. suber-Q. crenata*subtree were only recorded with cut-off = 2 . Conversely, most HTS sequences centered around *Q. faginea, Q. pyrenaica* and *Q. canariensis* references

(ubiquitous, undiagnostic type Q0 and *canariensis-pyrenaica*- specific type Q1; Fig. S3-4 in Supplementary file S3) only when a cut-off = 25 was applied.

Sample E5, mixed, one species per section—Of the 26,352 HTS sequences produced, only 1,140, 1,838 and 302 (i.e., in total 12.4%) were assigned by BLAST to the target species *Q. coccifera, Q. suber,*and *Q. canariensis* , respectively. Mirroring sample F2, all other sequences were assigned to the *Q. suber-crenata* shared types, *Q. ilex* , and to ambiguous or specific sequences of section *Quercus* (mostly *Q. faginea, Q. pyrenaica, Q. petraea, Q. pubescens* , and *Q. frainetto*). ML tree inferences also mirror those of sample F2: the *Q. coccifera* and *Q. suber-Q. crenata* subtrees are clearly identified with all abundance thresholds, together with the increasing dispersion along the sect. *Quercus*lineage. Only with cut-off = 2, few HTS sequences were placed in a *Q. ilex-Q. aucheri* subtree and together with the likely ancestral *Q. cerris-Q. trojana-Q. suber* sequence clade.

Sample E4, mixed, 8 species, all three sections—Of the 38,127 sequences produced in this sample, 10,699 (28.1%) were assigned by BLAST to the target species included: *Q. coccifera* (28), *Q. cerris* (335), *Q. trojana* (479), *Q. macrolepis* (4,107), *Q. frainetto* (2,549), *Q. petraea* (352), *Q. infectoria* (332), and *Q. pubescens* (2,517). All other sequences were generally assigned to members of subsection *Macrolepides*(sect. *Cerris*), and to sect. *Quercus* . With ML, the dispersion of HTS reads increased with the reduced cut-offs, just like the less-complex mixed samples. Reflecting the species composition of the sample, a minor part of the HTS sequences always placed within the *Macrolepides* and the 'occidental' *Cerris* clades, whereas most sequences were dispersed across the section *Quercus* subtree. Only with cut-off = 2, a *Q. coccifera*- exclusive clade was identified.

Figure 3A, B reports the ML trees of samples E5 (cut-off = 5), and E4 (cut-off = 2), depicting the different levels of information that can be obtained in relation to abundance cut-offs and/or complexity of samples. The NEWICK format of the 24 RAxML trees can be downloaded at https://doi.org/10.6084/m9.figshare.12016317.v1.

Automated species recognition using BLAST and EPA

Figure 4 summarizes the results of the taxonomic assignment of HTS sequences with abundance >25 obtained for each sample using the BLAST and EPA identification approaches (details provided in Supplementary file S5). Both approaches determined correct, unequivocal assignations of the species of sections *Ilex* and *Cerris*included in the pure (D5 and G2) and in the mixed samples (E4, E5, F2). For samples including material covering sect. *Quercus,* BLAST and EPA results differ (below section-level) because the former is overly specific when assigning a sequence to a reference. Our reference data do not include 5S-IGS variants unique to *Q. frainetto, Q. infectoria,* or *Q. petraea-pubescens* (identified in samples E4, E5, F2; see *Reference dataset*). In sample H1 (pure *Q. faginea* , ecomorphotype '*Q. lusitanica* '), all genotype-Q1 5S-IGS variants (Fig. S3-4 in Supplementary file S3) uniquely shared by *Q. canariensis-Q. pyrenaica* in the reference data are identified as *Q. canariensis* by BLAST, but either as *Q. canariensis*or *Q. canariensis-pyrenaica* by EPA.

Fig. 5 shows two examples of the EPA assignations of samples containing members of section *Quercus* on the reference tree. The HTS sequences of sample F2 (containing *Q. ilex* , *Q. suber* , and *Q. canariensis*) unambiguously group on *Q. ilex* and *Q. suber* branches; a *Q. canariensis-pyrenaica* minor cluster is also identified, together with *Q. faginea* and other *Quercus* section subclades, often in basal positions. The HTS sequences of sample H1 (pure *Q. faginea*) aggregate nearly on the same sect. *Quercus* subclades of sample F2, with a larger occurrence of the most derived *Q. faginea* types, and a second cluster, including *Q. canariensis* and *Q. petraea* . The general ability of EPA in the identification process in the six tubes with cut-off [?]25 is shown in Fig. 6.

Discussion

Diversity of the 5S-IGS in Quercus

Observed intra-, inter-specific and intra-individual polymorphism of the intergenic spacer of the 5S cistron in *Quercus* (in terms of unit length, GC content, and nucleotide sequence) is consistent with many other plant groups (Negi et al. 2002; Fulnečk et al. 2002; Forest et al. 2005; Grimm and Denk 2010; Mahelka et al. 2013; Mlinarec et al. 2016; Tynkevich and Volkov 2019). The intra-individual and/or intra-specific occurrence of two (or more) distinct 5S array types (e.g. short and long units in members of sect. *Cerris* , mainly *Q. libani*) is also consistent with the cited studies. The evolutionary significance of such intra-genomic polymorphisms is at odds with the generally acknowledged model of concerted evolution of rDNA, and its presence across many living organisms has led to the definition of the 'birth-and-death' model of evolution (Nei and Rooney 2005; Rooney and Ward 2005). Nevertheless, the occurrence of single rDNA loci in an individual, and the location of rDNA arrays far from the telomere may potentially facilitate concerted evolution (Eickbush and Eickbush 2007). In *Quercus* , one single 5S rDNA locus was physically mapped at a pericentromeric region (Ribeiro et al. 2011), and several species show only one single length and/or main sequence variant; a combined effect of both concerted and birth-and-death evolution models is therefore possible (Galian et al. 2014).

New gene variants can be generated by gene duplication, or gained by auto- or allopolyploidization (rDNA homoeologues; cf. Cronn et al. 2002), hybridization (crossing of evolutionary lineages) and unilateral gene flow (introgression). As they diverge, some may become non-functional pseudogenes, characterized by increased substitution rates and reduced GC content, and may eventually be eliminated (Volkov et al. 2007; see also Dzialuk et al. 2007 and Ribeiro et al. 2011, for the detection of triploids in natural oak populations). Contrary to the 35S (45S) rDNA cistron (encoding for the 18S, 5.8S and 25S rRNA genes), little is known about the GC content in functional 5S repeat units and its non-coding intergenic spacers (Symonová 2019). In this study, outlier sequence variants in CG content, relatively to each species' mean (generally lower than 50%, with the exception of *Q. alnifolia*) were coincident with long terminal branches on the reference tree, and/or incoherent groupings (Supplementary file S3), thus indicating potential pseudogeny. This number was, however, negligible (8 sequences in the reference dataset and 248 in the HTS dataset, corresponding to 0.004 and 0.002%, respectively). Conversely, the short sequence variants exhibited by e.g. *Q. libani* are inconspicuous (Tynkevich and Volkov 2019); these and other length outliers (Supplementary file S1) from known hybrids or species able to hybridize (e.g., *Q. petraea-Q.robur-Q. pubescens*) were discussed in Denk and Grimm (2010) and Simeone et al. (2018).

In general, the 5S-IGS sequences of endemic species appeared more homogeneous (e.g., *Q. alnifolia* , *Q. pontica* , *Q. afares*). Inter-specifically shared variants included sister species (e.g., *Q. suber* - *Q. crenata* , *Q. cerris* - *Q. look*), members of the same phylogenetic series (e.g., *Q. cerris* - *Q. trojana, Q. faginea* - *Q. canariensis*), or interfertile species (e.g.,*Q. cerris* - *Q. suber, Q. petraea* - *Q. pubescens, Q. petraea* - *Q. robur*). One ambiguous variant of section *Ilex* was generated by a*Q. baloot* outlier with reduced GC content (potential pseudogene). Other variants were shared among more distantly related species, but always within the same section, and could be explained by reticulation or retained ancient polymorphism. Species of section *Quercus* were the most difficult to resolve. Their 5S-IGS variants are more homogeneous in structure and sequence than those of the two other sections. This likely reflects the origin of the three sections on topographically different continents (Greenland/America, no west-east crossing mountain ranges vs. Asia, complex tectonic history), the more ancient origin of sections *Ilex* and *Cerris* , and the relatively recent expansion of white oaks into the Euro-Mediterranean region, involving bottlenecks and rapid radiation (Hipp et al. 2019b). At the same time, the hybridization rate in sect. *Quercus* is so extensive that the whole clade can be considered a syngameon (Hipp et al. 2019a). Besides a complex molecular evolution of the 5S cistron, further ecological events such as retention of ancestral traits, hybridization, species/population isolation, and drift may have played key roles in shaping the diversity patterns and the partial homogenization observed in the reference dataset (cf. Denk and Grimm 2010; Simeone et al. 2018). Nevertheless, the 5S-IGS data

confirmed its utility to resolve species or closely related species-groups in sections *Cerris, Ilex, Ponticae* and (to a lower degree) *Quercus* .

Uncertainty of the HTS data

In HTS amplicon sequencing (e.g. DNA metabarcoding), abundance of reads does not generally reflect taxa abundance (or biomass) in the sample matrix or bulk communities (but see Keller et al. 2015; Hirai et al. 2015; Piredda et al. 2017). A recent review on meta-analyses suggested that the methodology possesses some quantitative ability, but with a large degree of uncertainty (Lamb et al. 2019). Factors affecting the quantitative performance seem to be not related to the different sequencing platforms, whereas primer affinity, PCR biases, natural variation in copy number (and variation in biomass), are the most common explanations proposed (Lamb et al. 2019; Kelly et al. 2019). In our study, we applied a conservative approach where the performances of the HTS results and the original species composition of samples were considered only as incidence-data (presence/absence of taxa) (Hajibabaei et al. 2011; Porter et al. 2018). In relation to the potential sources of misidentifications (i.e., false-presence and false-absence), BLAST and EPA performed equally well. We did not score species present in the original DNA sample but not collected by the HTS data (false-absence). However, cases of false-presence (species revealed by the HTS data but not present in the original DNA sample) were reported both by BLAST and EPA, when reads with cut-off abundances generally lower than 10 were included in the analyses. These outcomes are well known in metabarcoding studies and confirm that a filtering of low abundance reads can generally avoid an inflation of diversity of samples with false-present taxa. Explaining the possible sources of false-presence is a hard task. In a qualitative and quantitative HTS assessment of artificial pollen mixtures, Bell et al. (2019) detected false presences also in negative controls, mostly corresponding to species included in their sample mixtures. Such false presences can be related to cross-contamination, flow-cell chimeras, or sequencing errors in the index sequences and usually consist of very low abundance reads. Exploring different abundance thresholds to control contamination is therefore an important step. However, we note that the false-presence signals retrieved in our data correlate not only with the used cut-offs but also with the taxonomic complexity of the samples. In the pure and mixed samples including members of sect. Quercus, for instance, only sequences with abundance >25 allowed exclusive identification of the correct species or species groups, whereas sequences with lower abundances were only assignable to the entire section. In contrast, the pure and mixed samples with members of sect. Ilex and Cerris scored sequences assignable to different (sister or close) species only with cut-offs <5. Interestingly, the pure samples referring to members of different sections never revealed sequences assignable to members of absent sections. These low-abundance sequences can all be therefore interpreted as ancestral, underived 5S IGS variants (within each section), or possible pseudogenes, and not exclusively as contaminations. Conversely, in the more complex sample E4 (including 16 individuals of eight total species), the three target species of sect. Cerris became all detectable with cut-off = 10, whereas Q. coccifera could be identified only with cut-off = 2. Therefore, the effect of different abundance cut-offs cannot be standardized, but rather adjusted based on the complexity of samples and the stated aims of a study, trading-off comprehensiveness vs. identification reliability. In sections Ilex and Cerris, utilization of medium-high abundance sequences (>10 or 25) can be more desirable for a net identification of species in samples containing one or few individuals. Data can be used to identify ambiguous specimens (e.g. hybrids, allochtonous species, morphologically deviating individuals), to delineate intra-specific variation and estimate inter-population diversity. Lower abundances (<10) can be more useful to reveal significant phylogenetic signals such as ancestry, introgression, pseudogeny. In more complex samples, such as environmental bulk samples, lower abundances (<10) may allow detection of less frequent species, although resulting in a reduced taxonomic resolution. Concerning the species of sect. Quercus, 'per-se' impossible to differentiate, abundances lower than 25 appear of no use, unless for complex samples containing members of the other sections. A valuable development of our approach would be a thorough inspection of all the abundance cut-offs in pure samples of any given species, to evaluate and assess the consistency of the retrievable signal in relation to contamination, possible pseudogenes, and potential (or the designed) applications.

General geno-taxonomic potential

BLAST and EPA approaches provided congruent identifications at high thresholds (abundance cut-off [?]25). Both methods largely agreed in identifying HTS sequences of sections *Ilex* and *Cerris* in the pure and mixed samples and, given the absence of sub-sectional resolution of the backbone tree (see Fig. 2) and the reference dataset, neither method could unambiguously assign sequences to the target species of sect. *Quercus* . Nevertheless, concordant species groups were always detected by the two methods, with respect to the maximum possible phylogenetic and geno-taxonomic resolution within the section. EPA always identified the correct target species or species lineage of each sample; BLAST often retrieved one or several of the possible species, thus over-taxonomizing the query and including species absent in the sample. This is likely due to the common share of 5S-IGS variants among oak species, especially in section *Quercus* (cf. Denk and Grimm 2010), and to the biased species representation in the reference dataset, where some species were under-represented compared to others such as *Q. canariensis* . We expect that both EPA and BLAST will be co-informative when more 5S-IGS references will be newly made available and more data accumulate. EPA will likely outperform BLAST in case of new, distinct variants in the data that are absent in the reference. In the future, samples displaying incongruence in the BLAST vs. EPA results will require additional nucleome- and plastome-based analyses, ideally accompanied by morpho-ecological re-examinations. With this workflow, our method has the potential as a fast identification tool to evaluate the possible occurrence of peculiar (e.g. alien) genotypes, rare lineages, or hidden hybridization/introgression events. Keeping in mind the overall species identification capacity of the 5S-IGS DNA region, the combination of BLAST and EPA automated taxomomy on the HTS data could be also useful to survey the overall genetic richness of local oak communities, as well as filtering sequences of certain species or species-groups from mixed samples for large-scale phylogenetic and biogeographic analyses.

Limitations and prospects: the case of Q. canariensis and Q. faginea

In the reference dataset, the best-sampled, species of section *Quercus* is *Q. canariensis* (Denk and Grimm 2010), one species of subsection *Galliferae* (Tschan and Denk 2012). Denk and Grimm (2010) highlighted the high variation occurring in *Q. canariensis* at both the ITS and the 5S IGS markers, compared to other species of the same section, and the large extent of identical variants shared with *Q. faginea* . In this work, we identified 5S-IGS variants apparently unique to *Q. canariensis* and *Q. pyrenaica* (genotype 'Q1'; Fig. S3-4 in Supplementary file S3) that attract hits in all HTS samples comprising material of either *Q. faginea* (pure sample H1), or *Q. canariensis* (sample E5 and F2) (Fig. 4). This finding could indicate that the investigated *Q. faginea* population is in fact a *Q. canariensis* relict, morphologically mimicking the particular steppe-, dry-climate adapted morphotype of *Q. faginea* locally known as '*Q. lusitanica* '. According to Tschan and Denk (2012), the exclusive presence of the *Q. lusitanica* morphotype in highly disturbed habitats, where it typically attains a shrubby habitus, and its morphology (based on a large dataset of micro- and macroscopic characters), point towards its consideration as an extreme adaptation of *Q. faginea* . The two species are also not readily diagnosable morphologically. It is therefore conceivable that particularly severe ecological conditions trigger *Q. canariensis* approaching morphologically *Q. faginea/lusitanica* , especially in the contact zone of both species. Asymmetrical introgression could enforce this pattern (e.g. Neophytou et al. 2010). However, it may also indicate that the reference data is unrepresentative for the genetic diversity in *Q. faginea.* In any case, it remains to explain the connection with *Q. pyrenaica,* a sympatric (but not belonging to *Galliferae*) species, only found in (south-)western Europe and (mainly) Northwestern Africa. It could be due to kinship with the only other fully mesic Iberian member of section *Quercus* (*Q. canariensis*) included in the reference dataset, thus indicating *in situ* differentiation of a regional mesic lineage. In the RAD-seq work of Hipp et al. (2019b), *Q. faginea* , *Q. canariensis* , *Q. pyrenaica* and *Q. lusitanica* , all represented by 1–2 individuals, were distributed on all four subclades produced in the section, thus indicating an early split of these western Mediterranean species. Lepais et al. (2013) and Leroy et al. (2017) showed that *Q. pyrenaica* seems to be not fully incorporated in the *Q. petraea-robur-pubescens* syngameon and concluded that *Q. pyrenaica* may represent a more anciently diverged species. The possible identification of ancestral sectional traits is therefore plausible. Interestingly, we identified a geographical partitioning in the HTS reads assigned to *Galliferae*(and *Q. pyrenaica*) in the three samples comprising western Mediterranean species of

this subsection (H1, F2, E5), and sample E4, where the only eastern Mediterranean species of this subsection (*Q. infectoria*) was mixed with several other members of the section. The *canariensis-pyrenaica* shared 5S-IGS variants (and the higher intra-genomic 5S-IGS diversity) could then be a witness of this antiquity: obtained before the *Galliferae* were isolated from the *Q. petraea-robur* species complex and retained in the two least evolved, and/or earliest diverged, species of each lineage as intra-genomic variation while being lost in all other species due to concerted evolution. Future studies with denser samplings including a single taxon per sample may therefore be interesting to see whether such variants show coherent distributional patterns or relate to morphological gradients.

Conclusion

The presented pipeline showed a good match with a complementary dataset (Sanger sequenced cloned fragments) and high potential for geno-taxonomy and evolutive inferences in oaks. Pure samples can be used to identify unknown individuals and inspect their genetic background. Mixed samples demonstrated a good ability to reveal presence/absence of taxa or specific lineages without missing any target in the sample. In addition, the pipeline has the potentiality to identify processes such as hybridization and introgression. At the same time, complex patterns resulting from multiplexed samples can be compared among different areas to evaluate genetic connectivity, discover local hotspots of genetic diversity, map the overall per-section diversity and possibly retrieve specific phylogeographic patterns. The only limitation is that the 5S IGS marker has different geno-taxonomic resolution in oak lineages. This problem applies mainly to members of section *Quercus* , but it could be partly overcome by expanding the benchmark reference.

Gaining better knowledge and evidence of the genetic resources of natural ecosystems, evolutionary dynamics and community assemblages is fundamental to improve our understanding of the natural processes shaping past and future biodiversity trends. HTS approaches are an innovative and important tool at this regard. They can provide new fundamental insights on species identity, history and natural evolution of communities. Adequately generating and interpreting data remain challenging but can be of utmost importance to reverse biodiversity decline, one of the fundamental targets of the 2030 Agenda for Sustainable Development (*www.sustainabledevelopment.un.org*). In oaks, the quest for a speedy and reliable marker for species identification and detection of introgression is on the way (Lepoittevin et al. 2015; Fitzek et al. 2018). However, transferability of the recently developed SNP toolkits to all taxa distributed across western Eurasia and other parts of the world, and even among distant populations within the same species, remains to be addressed. Our method is more flexible, as it can be used independently of the set-up of large genomic species-specific SNP surveys or genotype-by-sequencing approaches that for the western Eurasian oaks rely on just two species of one section (*Q. robur* and *Q. petraea*).

Acknowledgements

GWG has been funded by the Austrian Science Fund FWF (Project no. M-1751-B16).

EDS and GWG gratefully acknowledge the support of the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig funded by the German Research Foundation (FZT 118), which financed the sequencing.

Cited literature

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990). Basic local alignments earch tool. Journal of Molecular Biology, 215, 403-410.

Alvarez, I., Wendel, J. F. (2003). Ribosomal ITS sequences and plant phylogenetic inference. Molecular Phylogenetics and Evolution, 29, 417-434.

Babik, W., Taberlet, P., Ejsmond, M. J. A.N., Radwan, J. (2009). New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. Molecular Ecology Resources, 9, 713–719.

Bailey, C. D., Carr, T. G., Harris, S. A., Hughes, C. E. (2003). Characterization of angiosperm nrDNA polymorphism, paralogy, and pseudogenes. Molecular Phylogenetics and Evolution, 29, 435-455.

Bell, K. L., Burgess, K. S., Botsch, J. C., Dobbs, E. K., Read, T. D., Brosi, B. J. (2019). Quantitative and qualitative assessment of pollen DNA metabarcoding using constructed species mixtures. Molecular ecology, 28, 431-455.

Berger, S. A., Krompass, D., Stamatakis, A. (2011). Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. Systematic biology, 60, 291-302.

Bik, H. M., Fournier, D., Sung, W., Bergeron, R. D., Thomas, W. K. (2013). Intra-genomic variation in the ribosomal repeats of nematodes. PLoS ONE, 8:e78230

Blattner, F. E. (2009). Progress in phylogenetic analysis and a new infrageneric classification of the barley genus Hordeum (Poaceae; Triticeae). Breeding Science, 59, 471–480.

Burger, W. C. (1975). The species concept in Quercus. Taxon, 24, 45–50.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T. L. (2009). BLAST+: Architecture and applications. BMC Bioinformatics, 10, 421.

Carver, T., Bleasby, A. (2003). The design of Jemboss: a graphical user interface to EMBOSS. Bioinformatics, 19, 1837-1843.

Cavender-Bares, J., Gonzalez-Rodriguez, A., Eaton, D. A. R., Hipp, A. L., Beulke, A., Manos, P. S. (2015). Phylogeny and biogeography of the American live oaks (Quercus subsection Virentes): A genomic and population genetics approach. Molecular Ecology, 24, 3668-3687.

Cavender-Bares J. (2018). Sympatric parallel diversification of major oak clades in the Americas and the origins of Mexican oak diversity. New Phytologist, 217, 439-452.

Cloix, C., Tutois, S., Mathieu, O., Cuvillier, C., Espagnol, M. C., Picard, G., Tourmente, S. (2000). Analysis of 5S rDNA arrays in Arabidopsis thaliana: physical mapping and chromosome-specific polymorphisms, Genome Research, 10, 679–690.

Cronn, P., Cedroni, M., Haselkorn, T., Grover, C., Wendel, J. F. (2002). PCR-mediated recombination in amplification products derived from polyploid cotton. Theoretical and Applied Genetics, 104, 482-489.

Crowl, A. A., Manos, P. S., McVay, J. D., Lemmon, A. R., Lemmon, E. M., Hipp, A. L. (2020). Uncovering the genomic signature of ancient introgression between white oak lineages (Quercus). New Phytologist. doi:10.1111/nph.15842

Deiner, K., Bik, H. M., Machler, E., Seymour, M., Lacoursiere-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Molecular Ecology, 26, 5872-5895.

Denk, T., Grimm, G. W. (2010). The oaks of western Eurasia: traditional classifications and evidence from two nuclear markers. Taxon, 59, 351-366.

Denk, T., Grimm, G. W., Manos, P.S., Deng, M., Hipp, A.L. (2017) An Updated Infrageneric Classification of the Oaks: Review of Previous Taxonomic Schemes and Synthesis of Evolutionary Patterns. In Gil-Pelegrin, E., Peguero-Pina, J., Sancho-Knapik, D. (Eds). Oaks Physiological Ecology. Exploring the Functional Diversity of Genus Quercus L.. Springer, Cham.

Dzialuk, A., Chybicki, I., Welc, M., Soliwiniska, E., Burczyk, J. (2007). Presence of triploids among oak species. Annals of Botany, 99, 959–964.

Eaton, D. A. R., Hipp, A. L., Gonzalez-Rodriguez, A., Cavender-Bares, J. (2015). Historical introgression among the American live oaks and the comparative nature of tests for introgression. Evolution, 69, 2587-2601.

Edgar, R. C., Haas, B. J, Clemente, J. C. Quince, C., Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics, 27, 2194-2200.

Eickbush, T. H, Eickbush, D. G. (2007). Finely orchestrated movements: evolution of the ribosomal RNA genes. Genetics, 175, 477–485.

Eidesen, P. B., Alsos, I. G., Popp, M., Stensrud, O., Suda, J., Brochmann, C. (2007). Nuclear vs. plastid data: complex Pleistocene history of a circumpolar key species. Molecular Ecology, 16, 3902-3925.

Ekblom, R., Galindo, J. (2010). Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity, 107, 1-15.

Fitzek, E., Delcamp, A., Guichoux, E., Hahn, M., Lobdell, M., Hipp, A. L. (2018). A nuclear DNA barcode for eastern North American oaks and application to a study of hybridization in an arboretum setting. Ecology and Evolution, 8, 5837-5851.

Forest, F., Savolainen, V., Chase, M. W., Lupia, R., Bruneau, A., Crane, P. R. (2005). Teasing Apart Molecular- Versus Fossil-based Error Estimates when Dating Phylogenetic Trees: A Case Study in the Birch Family (Betulaceae). Systematic Botany, 30, 118-133.

Fulneček, J., Lim, K. Y., Leitch, A. R., Kovarík, A., Matyásek, R. (2002). Evolution and structure of 5S rDNA loci in allotetraploid Nicotiana tabacum and its putative parental species. Heredity, 88, 19-25.

Galián, J. A., Rosato, M., Rossellò, J. A. (2014). Partial sequence homogenization in the 5S multigene families may generate sequence chimeras and spurious results in phylogenetic reconstructions. Systematic Biology, 63, 219–230.

Ganley, A. R. D,, Kobayashi, T. (2007). Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. Genome Research, 17, 184-191.

Garcia, S., Kovařik, A. (2013). Dancing together and separate again: gymnosperms exhibit frequent changes of fundamental 5S and 35S rRNA gene (rDNA) organization. Heredity, 111, 23-33.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. Molecular Ecology Resources, 11, 759-769.

Gouy, M., Guindon, S., Gascuel, O. (2010). SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular biology and evolution, 27, 221-224.

Grimm, G. W., Denk, T. (2010). The reticulate origin of modern plane trees (Platanus, Platanaceae) - a nuclear marker puzzle. Taxon, 59, 134-147.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A., & Baird, D. J. (2011). Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS ONE, 6, e17497.

Heitkam, T., Petrasch, S., Zakrzewski, F., Kögler, A., Wenke, T., Wanke, S., Schmidt, T. (2015). Next-generation sequencing reveals differentially amplified tandem repeats as a major genome component of Northern Europe's oldest Camellia japonica. Chromosome Research, 23, 791.

Hipp, A. L. (2018a). Pharaoh's Dance: the oak genomic mosaic. PeerJ, 6, e27405v1.

Hipp, A. L., Manos, P. S., González-Rodríguez, A., Hahn, M., Kaproth, M., McVay, J. D., Valencia-Avalos, S., Cavender-Bares, J. (2018b). Sympatric parallel diversification of major oak clades in the Americas and the origins of Mexican species diversity. New Phytologist, 217, 439-452.

Hipp, A. L., Manos, P. S., Hahn, M., Avishai, M., Bodénès, C., Cavender-Bares, J., Crowl, A., Deng, M., Denk, T., Fitz-Gibbon, S., Gailing, O., Socorro González-Elizondo, M., González-Rodríguez, A., Grimm, G. W., Jiang, X.-L., Kremer, A., Lesur, I., McVay, J. D., Plomion, C., Rodriguez-Correa, H., Schulze, E.-D.,

Simeone, M. C., Sork, V. L. and Valencia-Avalos, S. (2019b). Genomic landscape of the global oak phylogeny. New Phytologist. doi: 10.1111/nph.16162.

Hipp, A. L., Whittemore, A.T., Garner, M., Hahn, M., Fitzek, E., Guichoux, E., Cavender-Bares, J., Gugger, P. F., Manos, P.S., Pearse, I. S., Cannon, C. H. (2019a). Genomic identity of white oak species in an eastern North Amarican syngameon. Annals of the Missouri Botanical Garden, 104, 455-477.

Hirai, J., Kuriyama, M., Ichikawa, T., Hidaka, K., Tsuda, A. (2015). A metagenetic approach for revealing community structure of marine planktonic copepods. Molecular ecology resources, 15, 68-80.

Hollingsworth, P. M., Graham, S. W., Little, D. P. (2011). Choosing and using a plant DNA barcode. PLoS ONE, 6, e19254.

Hubert, F., Grimm, G. W., Jousselin, E., Berry, V., Franc, A., Kremer, A. (2014). Multiple nuclear genes stabilize the phylogenetic backbone of the genus Quercus. Systematics and Biodiversity, 12, 405-423.

Ji, Y., Liu, C., Yang, Z., Yang, L., He, Z., Wang, H., Yang, J., Yi, T. (2019). Testing and using complete plastomes and ribosomal DNA sequences as the next generation DNA barcodes in Panax (Araliaceae). Molecular Ecology Resources, 19, 1333-1345.

Jiang, X.L., Hipp, A. L., Deng, M., Su, T., Zhou, Z.-K., Yan, M.-X. (2019). East Asian origins of European holly oaks (Quercus section Ilex Loudon) via the Tibet-Himalaya. Journal of Biogeography, 46, 2203-2214.

Katoh, K., Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution, 30, 772-780.

Keller, A., Danner, N., Grimmer, G., Ankenbrand, M. V. D., Von Der Ohe, K., Von Der Ohe, W., Rost, S., Hartel, S., Steffan-Dewenter, I. (2015). Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. Plant Biology, 17, 558-566.

Kelly, R. P., Shelton, A. O., Gallego, R. (2019). Understanding PCR processes to draw meaningful conclusions from environmental DNA studies. Scientific reports, 9, 1-14.

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., Schloss, P. D. (2013). Development of a dual index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Applied Environmental Microbiology, 79, 5112–20.

Kremer, A., Abbott, A. G., Carlson, J. E., Manos, P. S., Plomion, C., Sisco, P., Staton, M. E., Ueno, S., Vendramin, G. G. (2012). Genomics of Fagaceae. Tree Genetics & Genomes, 8, 583-610.

Kremer, A., Hipp, A. L. (2019). Oaks: an evolutionary success story. New Phytologist, doi:10.1111/nph.16274

Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., Taylor, M. I. (2019). How quantitative is metabarcoding: A meta-analytical approach. Molecular ecology, 28, 420-430.

Lepais, O., Roussel, G., Hubert, F., Kremer, A., Gerber, S. (2013). Strength and variability of postmating reproductive isolating barriers between four European white oak species. Tree Genetics & Genomes, 9, 841-853.

Lepoittevin, C., Bodenes, C., Chancerel, E., Villate, L., Lang, T., Lesur, I., Boury, C., Ehrenmann, F., Zelenica, D., Boland, A., Besse, C., Garnier-Gere, P., Plomion, C., Kremer, A. (2015). Single-nucleotide polymorphism discovery and validation in high-density SNP array for genetic analysis in European white oaks. Molecular Ecology Resources, 15, 1446-1459.

Leroy, T., Rougemont, Q., Dupouey, J.-L., Bodenes, C., Lalanne, C., Belser, C., Labadie, K., Provost, G. L., Aury, J.-M., Kremer, A., Plomion, C. (2020). Massive postglacial gene flow between European white oaks uncovered genes underlying species barriers. New Phytologist, doi:10.1111/nph.16039.

Leroy, T., Roux, C., Villate, L., Bodenes, C., Romiguier, J., Paiva, J. A. P., Dossat, C., Aury, J.-M., Plomion, C., Kremer, A. (2017). Extensive recent secondary contacts between four European white oak species. New Phytologist, 214, 865-878.

Letunic, I., Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic acids research, 47, W256-W259.

Mahelka, V., Kopecky, D., Baum, B. R. (2013). Contrasting Patterns of Evolution of 45S and 5S rDNA Families Uncover New Aspects in the Genome Constitution of the Agronomically Important Grass Thinopyrum intermedium (Triticeae). Molecular Biology and Evolution, 30, 2065-2086.

McVay, J. D., Hauser, D., Hipp, A. L., Manos, P. S. (2017b). Phylogenomics reveals a complex evolutionary history of lobed-leaf white oaks in western North America. Genome, 60, 733-742.

McVay, J. D., Hipp, A. L., Manos, P. S. (2017a). A genetic legacy of introgression confounds phylogeny and biogeography in oaks. Proceedings of the Royal Society B, 284, 20170300.

Menitsky, Y. L. (2005). Oaks of Asia. Science Publishers, Enfield, New Hampshire, USA

Mlinarec, J., Franjević, D., Bočkor, L., Besendorfer, V. (2016). Diverse evolutionary pathways shaped 5S rDNA of species of tribe Anemoneae (Ranunculaceae) and reveal phylogenetic signal, Botanical Journal of the Linnean Society, 182, 80-99

Negi, M. S., Rajagopal, J., Chauhan, N., Cronn, R., Lakshmikumaran, M. (2002). Length and sequence heterogeneity in 5S rDNA of Populus deltoides. Genome, 45, 1181-1188.

Nei, M., Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. Annual Review of Genetics, 39, 121-152.

Neophytou, C., Aravanopoulos, F. A., Fink, S., Dounavi, A. (2010). Detecting interspecific and geographic differentiation pattern in two interfertile oak species (Quercus petraea (Matt.) Liebl. and Quercus robur L.). Forest Ecology and Management, 259, 2026-2035.

Oh, S.-H., Manos, P. S. (2008). Molecular phylogenetics and cupule evolution in Fagaceae as inferred from nuclear CRABS CLAW sequences. Taxon, 57, 434–451.

Pham, K. K., Hipp, A. L., Manos, P. S., Cronn, R. C. (2017). A time and a place for everything: phylogenetic history and geography as joint predictors of oak plastome phylogeny. Genome, 60, 720-732.

Piredda, R., Tomasino, M. P., D'erchia, A. M., Manzari, C., Pesole, G., Montresor, M., Kooistra, W. H., Sarno, D., Zingone, A. (2017). Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean Long Term Ecological Research site. FEMS microbiology ecology, 93, fiw200.

Porter, T. M., Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. Molecular ecology, 27, 313-338.

Potts, A. J., Hedderson, T. A., Grimm, G. W. (2014). Constructing Phylogenies in the Presence Of Intra-Individual Site Polymorphisms (2ISPs) with a Focus on the Nuclear Ribosomal Cistron. Systematic Biology, 63, 1-16

Ribeiro, T., Loureiro, J., Santos, C., Morais-Cecílio L. (2011). Evolution of rDNA FISH patterns in the Fagaceae. Tree Genetics & Genomes, 7, 1113-1122.

Rooney, A. P., Ward, T. J. (2005). Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. Proceedings of the National Academy of Sciences of the U.S.A., 102, 5084-5089.

Schloss, P. D., Westcott, S. L., Ryabin, T. Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J.,

Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied Environmental Microbiology, 75, 7537- 41.

Simeone, M. C., Cardoni, S., Piredda, R., Imperatori, F., Avishai, M., Grimm, G. W., Denk, T. (2018). Comparative systematics and phylogeography of Quercus Section Cerris in western Eurasia: inferences from plastid and nuclear DNA variation. PeerJ, 6, e5793.

Simeone, M. C., Grimm, G. W., Papini, A., Vessella, F., Cardoni, S., Tordoni, E., Piredda, R., Franc, A., Denk, T. (2016). Plastome data reveal multiple geographic origins of Quercus Group Ilex. PeerJ, 4, e1897.

Simon, U. K., Trajanoski, S., Kroneis, T., Sedlmayr, P., Guelly, C., Guttenberger, H. (2012). Accession-specific haplotypes of the internal transcribed spacer region in Arabidopsis thaliana - a means for barcoding populations. Molecular Biology and Evolution, 29, 2231–2239.

Song, J., Shi, L., Li, D., Sun, Y., Niu, Y., Chen, Z., Luo, H., Pang, X., Sun, Z., Liu, C., Lv, A., Deng, Y., Larson-Rabin, Z., Wilkinson, M., Chen, S. (2012). Extensive pyrosequencing reveals frequent intra-genomic variations of internal transcribed spacer regions of nuclear ribosomal DNA. PLoS ONE, 7, e43971.

Stage, D. E., Eickbush, T. H. (2007). Sequence variation within the rRNA gene loci of 12 Drosophila species. Genome Research, 17, 1888-1897.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post- analysis of large phylogenies. Bioinformatics, 30, 1312-1313.

Straub, S. C. K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., Liston, A. (2012). Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. American Journal of Botany, 99, 349-364.

Symonová, R. (2019). Integrative rDNAomics - Importance of the Oldest Repetitive Fraction of the Eukaryote Genome. Genes, 10, 345.

Tschan, G. F., Denk, T. (2012). Trichome types, foliar indumentum and epicuticular wax in the Mediterranean gall oaks, Quercus subsection Galliferae (Fagaceae): implications for taxonomy, ecology and evolution. Botanical Journal of the Linnean Society, 169, 611–644.

Turner, B., Paun, O., Munzinger, J., Chase, M. W., Samuel, R. (2016). Sequencing of whole plastid genomes and nuclear ribosomal DNA of Diospyros species (Ebenaceae) endemic to New Caledonia: many species, little divergence. Annals of Botany, 117, 1175–1185.

Tynkevich, Y. O., Volkov, R. A. (2019). 5S Ribosomal DNA of distantly related Quercus species: molecular organization and taxonomic application. Cytology and Genetics, 53, 459-466.

Van Valen, L. (1976). Ecological species, multispecies, and oaks. Taxon, 25, 233-239.

Volkov, R. A., Komarova, N. Y., Panchuk, I. I., Hemleben, V. (2003). Molecular evolution of rDNA external transcribed spacer and phylogeny of sect. Petota (genus Solanum). Molecular Phylogenetics and Evolution, 29, 187-202.

Volkov, R. A., Zanke, C., Panchuk, I., Hemleben, V. (2001). Molecular evolution of 5S rDNA of Solanum species (sect. Petota): Application for molecular phylogeny and breeding. Theoretical and Applied Genetics, 103, 1273-1282.

Wang, W., Ma, L., Becher, H., Garcia, S., Kovarikova, A., Leitch, I. J., Leitch, A. R., Kovarik, A. (2016). Astonishing 35S rDNA diversity in the gymnosperm species Cycas revoluta Thunb. Chromosoma, 125, 683-699.

Weitemier, K., Straub, S. C. K., Fishbein, M., Liston, A. (2015). Intragenomic polymorphisms among high-copy loci: a genus-wide study of nuclear ribosomal DNA in Asclepias (Apocynaceae). PeerJ, 3, e718

Whittemore, A. T., Schaal, B. A. (1991). Interspecific gene flow in sympatric oaks. Proceedings of the National Academy of Sciences of the U.S.A., 88, 2540–2544.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Yan, M., Liu, R., Li, Y., Hipp, A. L., Deng, M., Xiong, Y. (2019) Ancient events and climate adaptive capacity shaped distinct chloroplast genetic structure in the oak lineages. BMC evolutionary biology, 19, 202.

Data accessibility

All generated raw HTS sequences are deposited in the Sequence Read Archive under the BioProject PRJNA611057. A database including 1160 unique 5S-IGS sequences (unaligned fasta file), the corresponding taxonomy file (in Mothur compatible format), and an Excel file (providing correspondence between clone codes, GenBank accessions and geographic origin) can be downloaded at (https://doi.org/10.6084/m9.figshare.12016272.v1).

All RAxML inferred trees can be downloaded in NEWICK format at (https://doi.org/10.6084/m9.figshare.12016317.v1). Other relevant data are within the paper and its Supplementary Files.

Author Contributions

RP, GWG, EDS, TD, MCS designed research

RP, GWG, MCS performed research

RP, GWG, MCS contributed reagents and analytical tools

RP, GWG, MCS analyzed data

RP, GWG, EDS, TD, MCS wrote the paper

Tables, Figures and supplementary files

Table 1. Investigated dataset with details on the type of sample assembled for the analysis, the geographic origin and taxonomic identity of the samples; N = number of individuals (= DNA extracts) multiplexed (pooled) in each tube; n = number of individual DNA extracts per species.

Figure 1 A-B . Boxplots of length (A) and GC-content (B) of the 5S-IGS units in the reference *Quercus* dataset (1160 sequences) and in each species.

Figure 2. Circular RAxML tree of the reference dataset. A: full tree with sections colored; B-D: only species of each section are colored.

Figure 3. Raw RAxML trees (unrooted) including references and HTS data of samples E5 (subset cut-off = 5) and E4 (subset cut-off = 2). Black leaves represent HTS reads; reference sequences are colored based on sections.

Figure 4. Comparison between BLAST and EPA assignments of HTS sequences with abundance >25 in each sample. Colors of the pie-sectors are according to the main 5S-IGS variants identified.

Figure 5. RAxML jplace tree including references and the EPA placements (in black) for a pure (F2) and a mixed (H1) samples with abundance cut-off = 25.

Figure 6. RAxML jplace tree including trees references and the EPA placements for the total HTS six samples with abundance > 25 (in red).

Supplementary files

S1 [XLSX]—Description of the 5S-IGS reference dataset, including: unique (non-redundant) and interspecifically shared ("ambiguous") sequences; main structural characteristics (length and GC content) in the different species (with 25th and 75th percentiles); outlier sequence variants.

S2 [XLSX]—Basic description of the obtained HTS dataset, including: the number of reads retained after the preprocessing steps; length and GC content of each HTS sequence with related scatter plots; details of the distribution of the HTS sequences in the six samples; BLAST assignations.

S3 [PDF]—RAxML-inferred guide trees based on 1160 5S-IGS reference sequences, including: unrooted tree with outliers labelled; annotated subtrees for each *Quercus* section.

S4 [PDF]—Colored, annotated versions of 24 RAxML trees inferred for the reference sequences and six HTS samples with four different abundance cut-offs (2, 5, 10, 25).

S5 [XLSX] —Taxonomic assignations of the HTS sequences with total abundance >25, obtained using BLAST and EPA in each sample.

Hosted file

`Table1.docx` available at https://authorea.com/users/311548/articles/442252-high-throughput-sequencing-of-5s-igs-in-oaks-exploring-intragenomic-variation-and-algorithms-to-recognize-target-species-in-pure-and-mixed-samples