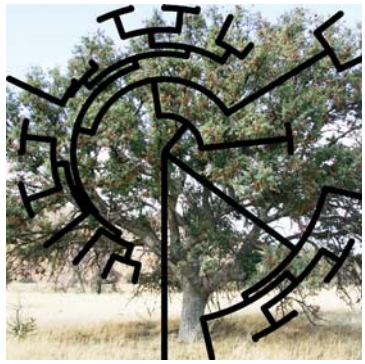

Using Phylogenomics to Infer the Evolutionary History of Oaks

Andrew L. Hipp[1,2], Deren A. Eaton[2,3], Jeannine Cavender-Bares[4], Rick Nipper[5], Paul S. Manos[6]

1. The Morton Arboretum
4100 Illinois Route 53
Lisle, IL 60532-1293, USA
Phone: +1 630 725-2094
ahipp@mortonarb.org

2. The Field Museum
Department of Botany
1400 S. Lake Shore Drive
Chicago, IL 60605-2496, USA

3. University of Chicago
Committee on Evolutionary Biology
1025 E. 57th Street
Chicago, IL 60637, USA

4. University of Minnesota
College of Biological Sciences
123 Snyder Hall
1475 Gortner Ave
Saint Paul, MN 55108, USA

5. Floragenex, Inc.
44 West Broadway
Eugene, Oregon, 97401, USA

6. Duke University
Department of Biology
Box 90338
Durham, NC 27708, USA

ABSTRACT

One of the most basic questions about oaks has long vexed botanists, systematists, and oak enthusiasts of all stripes: **what is the shape and timing of the oak tree of life?** In this paper, we present new data from a genomic study of 19 oak individuals representing a broad swath of the oaks of the Americas. The paper explains how we are using next-generation sequencing methods to analyze millions of base pairs of DNA data for all individuals studied, and what we have learned to date from this work about the oak tree of life.

Keywords: phylogenetics, *Quercus,* restriction-site associated DNA (RAD) tags, systematics

Introduction

What do we need phylogenies for? Why is it important to understand the relationships among living creatures? From a classification standpoint, we study phylogeny to make meaningful classifications and to identify species boundaries. As Charles Darwin noted, our taxonomic system is hierarchical not just because humans think hierarchically, but because our taxonomic hierarchies reflect the branching structure of the tree of life. Characteristics of the groups that we have named evolve along the tree of life, and the branching structure of that tree of life forms nested groups of organisms. Many of these hierarchical levels have formally named ranks (e.g., domains, kingdoms, phyla, classes, orders, families, and genera). There are nonhierarchical components to the tree of life: we know, for example, that Eukaryotes enclose mitochondria that are descendants of proteobacteria, and plants enclose plastids whose ancestors are cyanobacteria. But a large component of the history of life has a hierarchical structure that we convey as the tree of life.

Beyond making meaningful classifications, why else should we study phylogeny? The relationships among populations within species provide us with information about what species there are in the world, a question of paramount importance to ecologists, restorationists, horticulturalists, naturalists, lawmakers and judges, and anyone else interested or at least concerned with cataloguing biodiversity. Phylogenies are also important for predicting ecological interactions and making management decisions. Just as the botanists in Darwin's day were using the ratio of species to genera to characterize endemism and biodiversity, today's botanists use phylogenetic diversity to prioritize areas for conservation[1-3] and to track the biogeographic history and genetic connections among sensitive geographic areas or endemic taxa [4-6] Phylogenies are used to study the formation and ongoing evolution of biotic communities.[7-12] and to study species interactions integrated over time periods that are difficult to study experimentally. Phylogenies are used to study colonization history by invasive exotic species,[13, 14] providing important perspectives on the biological determinants of invasiveness. Phylogenies aren't just for systematists.

Yet for organisms in which ecological and morphological differences persist even in the face of interspecific gene flow,[15] phylogeny estimation can be problematic. [16-18] This is a pronounced problem in many forest trees, in which interfertility, high rates of outcrossing, large effective population sizes, and long generation times[19-21] make estimating phylogeny and patterns of trait evolution challenging. Oaks (*Quercus* L. : F*agaceae* Dumort.) are notable for the difficulties they pose to systematists. Renowned as a "worst case scenario for the biological species concept"[22] due to apparent local interspecific gene flow,[15, 23-32] widespread oak species nonetheless exhibit genetic coherence across broad geographic ranges.[33-35] We know relatively little about the phylogeny of this important genus. Traditional molecular approaches relying on chloroplast DNA[30, 36] or a small number of nuclear genes[37-39] often provide reliable information about broad-scale phylogenetic patterns ("what are the subgenera of oaks?"), but they typically fail to give answers for fine-scale phylogenetics ("what is the closest relative of *Quercus alba* L.?").

In this paper, we present preliminary results on the phylogeny of oaks using a new method of phylogenetic reconstruction, sequenced restriction associated DNA (RAD-Seq)[40]. This method provides a much-needed tool for surveying the genome of organisms like oaks, in which we need to sample broadly across the genome without having a

sequenced reference genome as a roadmap. It also provides us with sequence data that we can use to relate our phylogeny to ongoing genomic work in oaks, potentially, in the near future, allowing us to figure out what genes move among oak species and lineages. [41, 42] This level of detailed genomic inquiry was previously unavailable to us. Our study opens new doors to understanding how oaks have diversified and what constitutes an oak species or lineage. This paper focuses on explaining the methods we are using and their interpretation. Technical details of analysis will be left for a paper currently being submitted to a separate journal for publication.

Methods

Sampling

The target of this study is a clade of predominantly American oaks, comprising *Quercus* sections *Quercus*, *Lobatae* Loudon, and *Protobalanus* (Trelease) A. Camus[37]. These sections are the white oaks, red or black oaks, and intermediate or golden oaks respectively. We selected 19 species from this clade and one member of section *Cerris* Dumort. to serve as an outgroup, the species we used to identify the root of the New World oak clade. The root is the oldest point on the phylogenetic tree, the putative ancestor of all species on the tree, and is typically identified using a more distantly related species, or outgroup. Samples were all drawn from previously collected material, many obtained from the seed exchange and field trip of the 2006 IOS Conference in Dallas, and reared in the greenhouse at the University of Minnesota.

DNA extraction

DNA was extracted from fresh material using the standard DNeasy plant extraction protocol (DNeasy, Qiagen, Valencia, CA), with modifications that we have used for previous studies in oaks.[34, 43] In this method, fresh or frozen leaf tissue is ground thoroughly by hand in liquid nitrogen to a fine powder, using a mortar and pestle. The resulting leaf tissue powder is then incubated at 65°C in a solution of sodium dodecyl sulfate, a component of many detergents. The detergent digests cell walls and membranes without damaging the DNA inside the cells. RNase was included in this step to digest RNA, as our sequences of interest are all in the genomic DNA. Unwanted cell structures (proteins) and secondary compounds are all precipitated out of solution at near-freezing temperatures for 10 minutes, then centrifuged to separate these from the buffer that contains the DNA. This buffer is then sucked out using a pipette and applied to a filter that is centrifuged at high speed to separate the DNA (which passes through the filter) from cell debris. The cell debris remains on the filter, which is thrown away, and the DNA is precipitated in a salt and ethanol solution. The DNA precipitate is washed several times on a second filter before being suspended in a standard DNA buffer. This DNA extraction can then be frozen and used for years or decades of molecular study.

RAD sequencing

Restriction associated DNA sequencing (RAD-Seq) was conducted at Floragenex following the methods of Baird, Etter[40]. RAD sequencing allows us to sample DNA sequence data from across the entire genome of an organism at fairly low cost, by subsampling just those regions of the genome that lie adjacent to a restriction cut site. Restriction cut sites are defined sequences, typically 4 to 8 base pairs long—i.e., 4–8 nucleotides in length, measuring just one strand of the double-stranded DNA molecule—

that a suite of bacterial/archaebacterial restriction enzymes will cut. More than 3,000 restriction enzymes are known[44], each of which cuts at a particular sequence, the restriction site. These enzymes act as a defensive mechanism for the bacteria against viruses, but they also serve well in the laboratory.

1. Restriction digestion

The first step in RAD sequencing is cutting the DNA using a restriction enzyme. In our oak work, we use *Pst*I, a restriction enzyme derived from the bacterium *Providencia stuartii* Ewing that cuts only at the 6-base-pair sequence

5'… CTGCA|G …3'
3'… G|ACGTC …5'

Like many other restriction enzymes, *Pst*I leaves a ragged end when it cuts, meaning that the cuts in each strand of the double-stranded DNA molecule (cuts indicated above by the '|' symbol) do not exactly line up with each other. This is essential to the next step of the process, adapter ligation, for the ragged end serves as a sort of "sticky end" to which the adapter will fit like a puzzle piece. Assuming a GC content of 40%, a genome size of 500 million bases (both of which are typical of oaks), and a completely random draw of the four nucleotides that make up DNA (Adenine, Cytosine, Guanine, and Thymine), we expect about 72,000 *Pst*I cut sites in the oak genome.

2. Adapter ligation and shearing

The ragged ends left by *Pst*I serve as a platform to which we ligate a manufactured double-stranded DNA called an adapter. This first adapter ('P1') includes a nucleotide sequence needed for DNA amplification, a sequence required for DNA sequencing on the Illumina sequencer (see 3, Illumina sequencing, below) and a "barcode," a unique combination of 5 nucleotides that identify the individual sequenced. Sequences are then randomly sheared to get fragments of varying lengths. After shearing, a second adapter ('P2') is ligated to the fragments, with a PCR amplification site embedded in it. We now have a pool of DNA fragments with a 'P1' site at one end and a 'P2' site at the other. Because the shearing process leaves some fragments without a restriction site, we also have a pool of fragments with a P2 at each end, which we don't want to sequence. The P2 adapters are therefore built in such a way that only fragments with a restriction site at one end (and thus a P1 at one end and P2 at the other) will be duplicated in a subsequent round of DNA amplification. In this final stage of DNA preparation for sequencing, fragments with a restriction site are enriched by a factor of roughly 130,000. Thus, while a few sequences may show up that are not associated with a DNA restriction site, they should be very few indeed. At this point, our DNA is referred as a RAD library.

3. Illumina sequencing

There are several methods of massively parallel or "next-generation" sequencing. The method we are using generates, as of spring 2012, sequence reads of approximately 85 nucleotides in length on an Illumina/Solexa Genome Analyzer IIx. The data we present in this paper also includes sequencing runs from spring 2010 that returned 60 nucleotides of data per sequence from 19 individuals, of which we replicated seven individuals in 2012. In brief (see more complete introduction to next-generation sequencing technologies in[45]) the RAD library is applied to a glass plate called a flow cell, in which individual DNA strands from the RAD library bind to separate sequencing sites on the plate. Each

DNA strand is then duplicated (amplified) on the flow cell surface to form a colony of thousands of identical DNA strands. In the final step, each DNA strand in each of these colonies is amplified in a stepwise process, during which a mix of fluorescently labeled is washed over the place, the next nucleotide in the sequence is added, a photo is snapped, then the plate is readied for addition of the next nucleotide. Because each nucleotide is labeled with a different color, the sequence of colors photographed for each colony spells out the DNA sequence for that colony. These sequences are stored by a computer hooked up to the sequencer, error-checked, and returned to us as a batch of sequences.

4. Preliminary data analysis

Processed data were returned from sequencing in the Illumina 1.3+ variant of the FASTQ format,[46] with Phred quality scores for all bases.[47] Quality, read lengths, and base composition of FASTQ data were assessed in R v. 2.15.2[48] using the ShortRead package.[49]

Creating a DNA data matrix

Data were analyzed following a custom pipeline that approximately follows the method of Catchen, Amores[50]. In this method, sequences are clustered first by individual, and highly similar sequences are clustered into "stacks," from which heterozygote base pair positions (i.e., positions of the genome in which the mother oak and father oak contributed different bases) are distinguished from sequencing errors. Each stack is referred to here as a locus, a term commonly used to reference a region of the genome not assumed to be a gene. For each individual, each sequence stack is summarized into a consensus sequence, and these consensus sequences are then clustered among individuals to generate a data matrix for each locus. Not every individual has a sequence in every locus. Nonetheless, loci are concatenated to make a data matrix with missing data (a hole-ly data matrix).

Several parameters must be specified in creating this data matrix, including the quality of sequencing reads, percent similarity required to cluster sequences, the depth of sequence stacks needed to make a locus, the minimum number of individuals per locus, and various properties of data variability. A range of parameter values was investigated for this study, and the basic topology recovered varied only in (1) the amount of statistical support of the placements recovered, and (2) the precise placement of *Quercus robur* L. Details of these analyses will be investigated in a future study. The analysis pipeline we utilize is coded in Python and utilizes UCLUST [51] and MUSCLE[52, 53] for clustering and multiple alignment respectively. Details of the pipeline and methods of assessing rates of error and heterozygosity are being published elsewhere (by DE). For this paper, we present analysis of clustering results in which a minimum of four individuals were required for each locus, and report on phylogenetic analyses conducted both with and without *Q. robur* L.

Phylogenetic analysis

To assess phylogenetic relationships, we used maximum likelihood as implemented in RAxML v7.2.6,[54] which is optimized for large phylogenomic datasets. The maximum likelihood method as applied to phylogenetic inference is described in a thorough treatment by Felsenstein[55]. In short, every phylogenetic tree is a hypothesis that confers a probability on our data matrix. In general, a dataset in which species A and B share lots of mutations that no other species share will tend to be most probable if species A and B are truly sister to each other. The maximum likelihood method in phylogenetics searches the space of phylogenetic hypotheses for trees that maximize the probability of the observed data under a quantitative model of nucleotide evolution. This is the standard

method of phylogenetic inference for DNA sequence data. Analyses were conducted using the 'GTRGAMMA' general time reversible model of nucleotide evolution. Under this model, all possible transitions among nucleotides (A, G, C, T) are allowed to differ in rate, but those rates are assumed to be symmetrical (e.g., the probability of an A to C mutation is assumed to be the same as the probability of a C to A mutation). These rates of mutation are not assumed to be constant along the entire length of the sequences we analyze, but to vary among site to site according to a gamma (Γ) distribution. Branch support was assessed using 200 nonparametric bootstrap replicates, which are simply reanalyses of the dataset, resampling nucleotides at random.

Results

RAD sequences

For the year 2010 (initial) run, individuals yielded 177,168 to 725,871 sequences (mean = 558,006, sd = 136,157) of 60 base pairs each (Figure 1, in red). For the 2012 (replicate) run, each individual yielded between 743,556 and 4,539,385 sequences (mean = 3,056,861, sd = 1,369,094) of 95 base pairs each (Figure 1, in black). This is a 5.5-fold increase in number of sequences yielded between 2010 and 2012. After removing the 5 base pairs left over from the *Pst*I cut and the 5-base-pair barcode from each sequence, and ignoring decreases in quality toward the ends of the reads, this is a 9.3-fold increase in total sequence data per individual between 2010 and 2012.

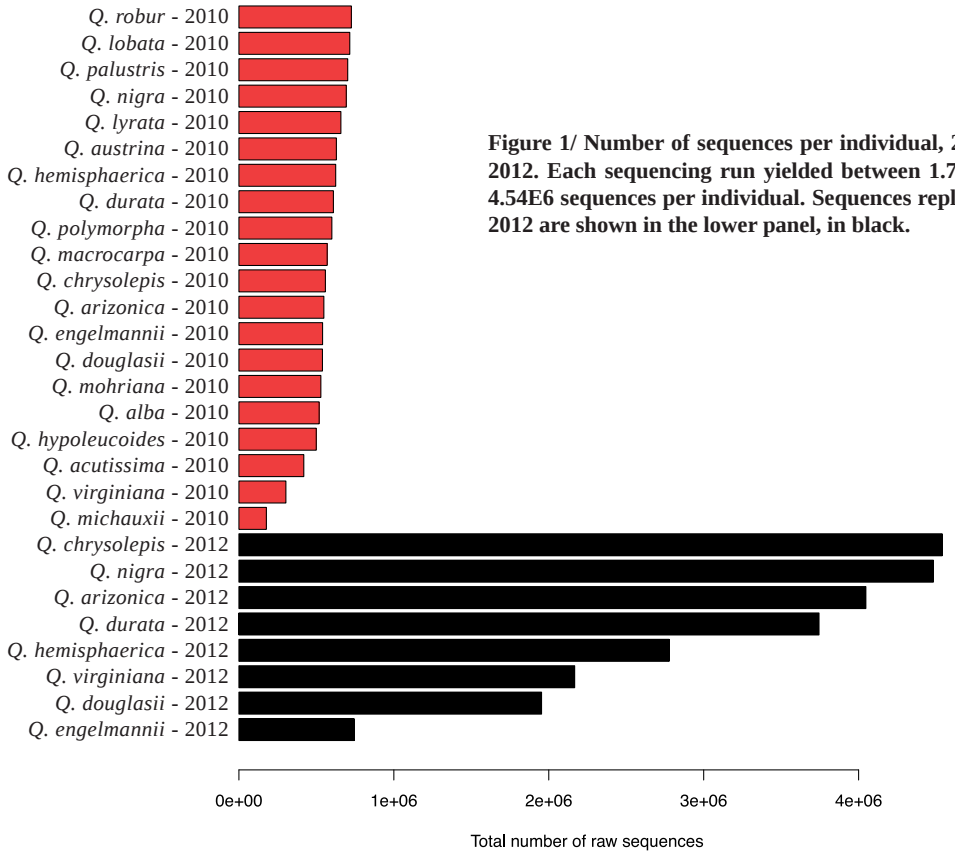

Figure 1/ Number of sequences per individual, 2010 and 2012. Each sequencing run yielded between 1.77E5 and 4.54E6 sequences per individual. Sequences replicated in 2012 are shown in the lower panel, in black.

DNA data matrix

Within individuals, the average number of sequences used to estimate the consensus sequence for each locus was 12.8 ± 0.8 (standard deviation) for 2010 data, 38.1 ± 13.2 for 2012 data. A total of 61,054 loci were inferred with a minimum of 4 individuals per locus, 24,778 with a minimum of 10 individuals per loci. By comparison, the longest previous DNA-based dataset utilized in oak phylogenetic inference[43] utilized 2,932 AFLP bands, each of which reflects the evolution of 16 to 18 base pairs constituting the recognition sites flanking that band, a total of ca. 47,000 base pairs of data. All pairs of technical replicates share fewer than 52% of the loci found in the union set of loci for the pair. Locus coverage in the 2012 sequencing runs was 19% to 127% greater than the 2010 sequencing runs for the same individuals.

Phylogeny

Analysis of the aligned data matrix recovers section *Lobatae* as sister to sections *Quercus* and *Protobalanus*, and all three of these as monophyletic insofar as we have sampled them (Figure 2). It also places the live oaks of the *Virentes* group sister to the remainder of section *Quercus*. All of these relationships are recovered with 100% bootstrap support. This topology has also been recovered in previous phylogenetic studies on oaks based on DNA sequences[37] and AFLP data,[43] but with lower statistical support.

Phylogenetic analysis of the sequence data, treating missing loci as missing characters, places all 2012 technical replicates sister to their 2010 counterparts, with terminal branch lengths substantially shorter than the subtending internode (Figure 3; p. 68). This suggests that missing data have negligible effect on species placement on the tree, and that data are readily combined across sequencing runs. This is a substantial improvement over AFLP

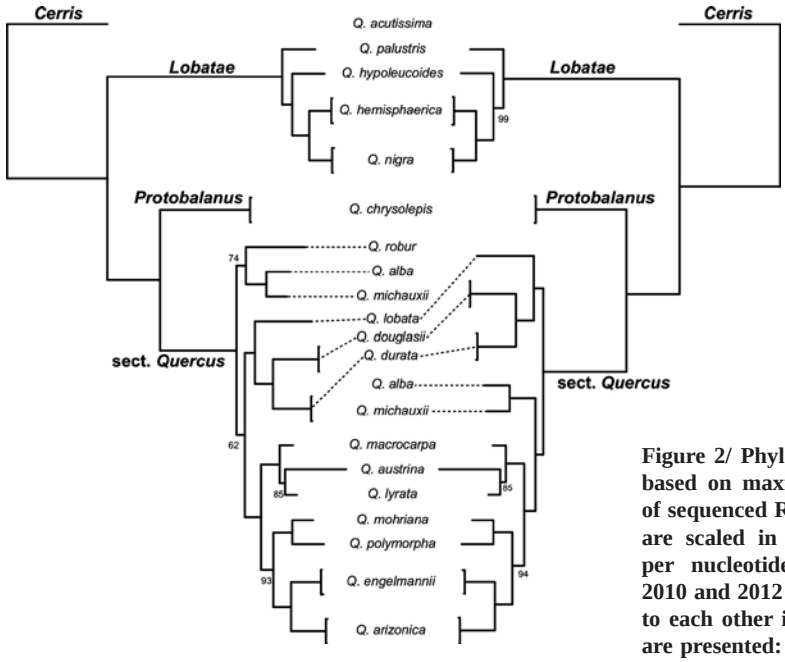

Figure 2/ Phylogenetic trees of *Quercus* based on maximum likelihood analysis of sequenced RAD data. Branch lengths are scaled in number of substitutions per nucleotide. Individuals from the 2010 and 2012 sequencing runs fall next to each other in all cases. Two analyses are presented: one in which all taxa are included (left panel), and one in which *Quercus robur* is excluded (right panel).

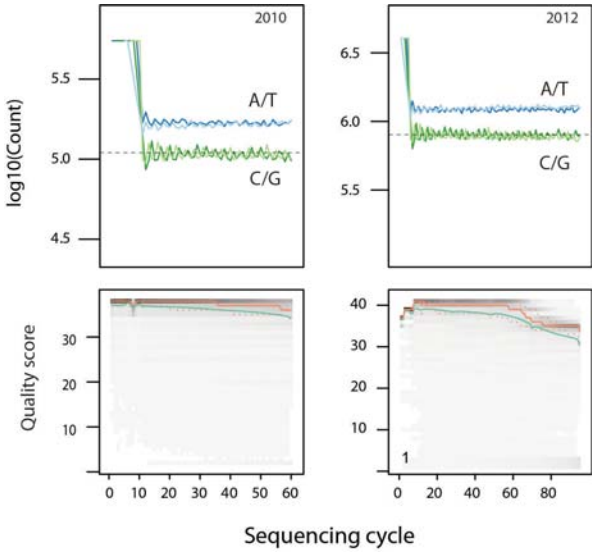

Figure 3/ Quality and base-pair composition of two representative sequencing runs. The same individual *Quercus alba* L. was sequenced in both 2010 (left panels) and 2012 (right panels), using the same extraction and RAD library preparation. DNA quality is reported using Phred quality scores, which have been in use for reporting DNA sequencing quality since 1998. Quality decreases toward the end of each read, and 2012 sequences exhibit lower quality commensurate with their longer sequencing reads. Percent C plus percent G also appears to decrease very slightly toward the ends of the reads in 2010 and 2012. Read quality and percent CG include the first 10 bases of each sequencing read, which comprise the individual-specific DNA barcode and five bases of the *Pst*I restriction cut site. These 10 bases are stripped off prior to all other analyses.

data, in which combining data across separate analyses is time-consuming, requiring rescoring of the entire data matrix, and often presents technical challenges.

Discussion

Despite our very sparse sampling (18 out of more than 250 in the New World oak clade), two phylogenetic results stand out in this study within section *Quercus* (the white oaks). First, this study supports previous findings[37, 39, 43] that the Eurasian white oaks of section *Quercus* are embedded within the otherwise New World clade sampled here. The position we find, however, is novel: whereas we find *Quercus robur* to be embedded within or sister to one of the Eastern North American clades, previous study using AFLP data has suggested that the Eurasian white oaks of are sister to the non-*Virentes* members of section *Quercus* from North America[43], and a study utilizing nuclear ribosomal DNA sequences suggested a relationship between the Western North American *Q. sadleriana* R.Br.ter. and *Q. pontica* K. Koch of the Western Caucasus Mountains[39]. Our placement of *Q. robur* is far from conclusive, however, as suggested by the fact that analysis with *Q. robur* appears to drag the Eastern North American *Q. alba* and *Q. michauxii* Nutt. to a position sister to the remainder of the white oaks, but with low statistical support for that placement (62%, nonparametric bootstrap). Moreover, the placement of *Q. robur* is highly sensitive to clustering parameters (alternative analyses not shown here), suggestive that different partitions of the RAD sequence dataset may encode different placements of *Q. robur*, due either to hybridization or to rapid diversification at the base of the section.[56-59] Resolving the position of the Old World white oaks will certainly require additional sampling of both Eurasian and American species of section *Quercus* and thorough analysis of a large number of nuclear loci.

Second, our data separate the New World white oaks into small geographic clades, with the Eastern North American taxa non-monophyletic. Prior morphological studies of the genus (e.g.,[60, 61]) have suggested some morphological groupings within subgenus

Quercus as a step toward an infrasectional classification of the genus. The results here, in which, for example, the Californian white oaks (*Q. lobata* Née, *Q. douglasii* Hook. & Arn., *Q. durata* Jeps.) form a clade distinct from the predominantly southwestern North American/Mexican oaks sampled (*Q. mohriana* Buckley ex Rydb., *Q. polymorpha* Schldtl. & Cham., *Q. engelmannii* Greene, *Q. arizonica* Sargent), suggest that such a classification may be within reach, and that that classification may rest strongly on biogeography.

Finally, our study demonstrates the utility of RAD data for reconstructing phylogenetic relationships in a problematic group, spanning roughly 40 million years of evolutionary history. As part of our currently funded work, we are sampling roughly 150 species of the New World oak clade (sections *Lobatae*, *Protobalanus*, and *Quercus*), using the same methods described in this paper. Our expectation is that within the coming few years, we will have a handle on the shape and timing of a large portion of the oak tree of life, which has for so long proved elusive.

1/ *Quercus alba*.

Acknowledgements

This work was supported by NSF Awards #1146488 to AH, #1146380 to JCB, and #1146102 to PM. Kari Koehler (UM, Minneapolis, Minnesota, U.S.A.) managed all plants in the greenhouse and performed all DNA extractions for this study. Tressa Atwood (Floragenex, Portland, Oregon, U.S.A.) executed all labwork and worked closely with participating labs (AH, JCB) to ensure quality of extractions and data. Antoine Kremer

(INRA, UMR BioGeCo, Cestas, France) reviewed this manuscript and hosted, with The International Oak Society, the meeting at which this work was presented. Béatrice Chassé graciously extended the invitation to AH to present this work, coordinated many of the aspects of the meeting and proceedings, and provided editorial feedback on the manuscript.

Author contributions

AH, PM, and JCB conceived the study and provided DNA extractions for sequencing. RN coordinated the RAD library preparation and sequencing. DE developed the analysis pipeline. DE and AH analyzed the data. All authors contributed to writing and editing of the paper.

Photographers. Title page: Andrew Hipp (trees of life). Photo 1: Philippe de Spoelberch.

References

1. D.P. Faith. 2008. Threatened species and the potential loss of phylogenetic diversity: conservation scenarios based on estimated extinction probabilities and phylogenetic risk analysis. *Conservation Biology* 22(6): 146–470.

2. D.P. Faith and A.M. Baker. 2006. Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. *Evolutionary Bioinformatics*, 2: p. 70–77.

3. S. Knapp, I. Kuhn, O. Schweiger, and S. Klotz. 2008. Challenging urban species diversity: contrasting phylogenetic patterns across plant functional groups in Germany. *Ecology Letters* 11: 1054–1064.

4. M.H. Hoffmann and M. Röser. 2009. Taxon recruitment of the arctic flora: an analysis of phylogenies. *New Phytologist* 182(3): 774–780.

5. B.W. Van Ee, N. Jelinski, P.E. Berry, and A. Hipp. 2006. Phylogeny and biogeography of *Croton alabamensis* (*Euphorbiaceae*), a rare shrub from Texas and Alabama, using DNA sequence and AFLP data. *Molecular Ecology* 15(10): 2735–2751.

6. T.M. Harding, P.S. Soltis, and D.E. Soltis. 2000. Diversification of the North American shrub genus Ceanothus (Rhamnaceae): conflicting phylogenies from nuclear ribosomal DNA and chloroplast DNA. *Am. J. Bot.* 87(1): 108–123.

7. J. Cavender-Bares, K.H. Kozak, P.V.A. Fine, and S.W. Kembel. 2009. The merging of community ecology and phylogenetic biology. *Ecology Letters* 12(7): 693–715.

8. C.O. Webb, D.D. Ackerly, M.A. McPeek, and M.J. Donoghue. 2002. Phylogenies and community ecology. *Annual Review of Ecology and Systematics* 33: 475–505.

9. C.H. Graham and P.V.A. Fine. 2002. Phylogenetic beta diversity: linking ecological and evolutionary processes across space in time. *Ecology Letters* 11(12): 1265–1277.

10. S.M. Vamosi, S.B. Heard, J.C. Vamosi, and C.C. Webb. 2009. Emerging patterns in the comparative analysis of phylogenetic community structure. *Molecular Ecology* 18(4): 572–592.

11. P.V.A. Fine, I. Mesones, and P.D. Coley. 2004. Herbivores Promote Habitat Specialization by Trees in Amazonian Forests. *Science* 305(5684): 663–665.

12. J. Cavender-Bares, D.D. Ackerly, D.A. Baum, and P.A. Bazaaz. 2004. Phylogenetic overdispersion in Floridian oak communities. *American Naturalist* 163(6): 823–843.

13. K. Saltonstall, Cryptic invasion by a non-native genotype of the common reed, *Phragmites australis*, into North America. 2002. *Proceedings of the National Academy of Sciences of the United States of America* 99(4): 2445–2449.

14. C.E. Lee and G.W. Gelembiuk. 2008. Evolutionary origins of invasive populations. *Evolutionary Applications* 1(3): 427–448.

15. L. Van Valen. 1976. Ecological species, multispecies, and oaks. *Taxon* 25: 233-239.

16. L.S. Kubatko. 2009. Identifying Hybridization Events in the Presence of Coalescence via Model Selection. *Systematic Biology* 58(5): 478–488.

17. C. Meng and L.S. Kubatko. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theoretical Population Biology* 75(1): 35–45.

18. H. Huang, Q. He, L.S. Kubatko, and L.L. Knowles. 2010. Sources of Error Inherent in Species-Tree Estimation: Impact of Mutational and Coalescent Effects on Accuracy and Implications for Choosing among Different Methods. *Systematic Biology* 59(5): 573–583.

19. J.L. Hamrick. 2004. Response of forest trees to global environmental changes. *Forest Ecology and Management* 197(1-3): 323–335.

20. O. Savolainen, T. Pyhajarvi, and T. Knurr. 2007. Gene flow and local adaptation in trees. *Annual Review of Ecology Evolution and Systematics.* 38: 595–619.

21. S.M. Hoban, T.S. McCleary, S.E. Schlarbaum, and J. Romero-Severson. 2009. Geographically extensive hybridization between the forest trees American butternut and Japanese walnut. *Biology Letters* 5(3): 324–327.

22. J.A. Coyne and H.A. Orr., *Speciation* (Sunderland, MA: Sinauer Associates, 2004).

23. R. Petit, C. Bodénès, A. Ducousso, G. Roussel, and A. Kremer. 2004. Hybridization as a mechanism of invasion in oaks. *New Phytologist* 161: 151–164.

24. C. Burgarella, Z. Lorenzo, R. Jabbour-Zahab, R. Lumaret, E. Guichoux, R.-J. Petit, Á. Soto, and L. Gil. 2009. Detection of hybrids in nature: application to oaks (*Quercus suber* and *Q. ilex*). *Heredity* 102(5): 442–452.

25. O. Lepais, R.-J. Petit, E. Guichoux, J.E. Lavabre, F. Alberto, A. Kremer, and S. Gerber. 2009. Species relative abundance and

direction of introgression in oaks. *Molecular Ecology* 18: 2228–2242.

26. R.-J. Petit and L. Excoffier. 2009. Gene flow and species delimitation. *Trends in Ecology & Evolution* 24(7): 386–393.

27. L. Lagache, E.K. Klein, E. Guichoux, and R.-J. Petit, Fine-scale environmental control of hybridization in oaks, *Molecular Ecology*, 2013. 22(2): 423–436.

28. W.C. Burger. 1975. The species concept in Quercus. *Taxon* 24: 45–50.

29. C. Lexer, A. Kremer, and R.J. Petit. 2006. Shared alleles in sympatric oaks: recurrent gene flow is a more parsimonious explanation than ancestral polymorphism. *Molecular Ecology* 15: 2007–2012.

30. Whittemore, A.T. and B.A. Schaal. 1991. Interspecific gene flow in sympatric oaks. *Proceedings of the National Academy of Sciences USA* 88: p. 2540–2544.

31. S. Dumolin-Lapegue, B. Demesure, S. Fineschi, V. Le Comte, and R.-J. Petit. 1997. Phylogeographic structure of white oaks throughout the European continent. *Genetics* 146: 1475–1487.

32. J.W. Hardin. 1975. Hybridization and introgression in *Quercus alba. Journal of the Arnold Arboretum* 56: 336–363.

33. A. González-Rodríguez, D.M. Arias, S. Valencia, and K. Oyama. 2004. Morphological and RAPD analysis of hybridization between *Quercus affinis* and *Q. laurina (Fagaceae)*, two Mexican red oaks. *American Journal of Botany* 91(3): 40–409.

34. A.L. Hipp and J.A. Weber. 2008. Taxonomy of Hill's Oak *(Quercus ellipsoidalis: Fagaceae):* Evidence from AFLP Dat. *Systematic Botany* 33: 148–158.

35. G. Muir, C.C. Fleming, and C. Schlötterer. 2000. Species status of hybridizing oaks. *Nature* 405: 1016.

36. R.-J. Petit, R., E. Pineau, B. Demesure, R. Bacilieri, A. Docousso, and A. Kremer. 1997. Chloroplast DNA footprints of postglacial recolonization by oaks. *Proceedings of the National Academy of Sciences USA* 94: 9996–10001.

37. P.S. Manos, J.J. Doyle, and K.C. Nixon. 1999. Phylogeny, Biogeography, and Processes of Molecular Differentiation in *Quercus* Subgenus *Quercus (Fagaceae). Molecular Phylogenetics and Evolution* 12(3): 333–349.

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

39. T. Denk and G.W. Grimm. 2010. The oaks of western Eurasia: Traditional classifications and evidence from two nuclear markers. *Taxon* 59: 351–366.

40. N.A. Baird, P.D. Etter, T.S. Atwood, M.C. Currey, A.L. Shiver, Z.A. Lewis, E.U. Selker, and W.A. Cresko. 2008. *Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers.* PLoS ONE, 3(10): e3376.

41. A. Kremer, A.G. Abbott, J.E. Carlson, P.S. Manos, C. Plomion, P. Sisco, M.E. Staton, S. Ueno, and G.G. Vendramin. 2012. Genomics of *Fagaceae. Tree Genetics & Genomes*, doi: 10.1007/s11295-012-0498-3.

42. J. Durand, C. Bodénès, E. Chancerel, J.-M. Frigerio, G. Vendramin, F. Sebastiani, A. Buonamici, O. Gailing, H.-P. Koelwijn, F. Villani, C. Mttioni, M. Cherubini, P.G. Goicoechea, A. Herrán, Z. Ikaran, C.Cabané, S. Ueno, F. Alberto, P.-Y. Dumouline, E. Guichoux, A. de Daruvar, A. Kremer, and C.Plomion. 2010. A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. *BMC Genomics* 11(1): 570.

43. I.S. Pearse, and A.L. Hipp. 2009. Phylogenetic and trait similarity to a native species predict herbivory on non-native oaks, *Proceedings of the National Academy of Sciences of the United States of America*. 106(43): 18097–18102.

44. Wikipedia, *Restriction enzyme.* ***http://en.wikipedia.org/wiki/Restriction_enzyme*** 2013. Accessed 21 January 2013.

45. E.R. Mardis. 2008. Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics* 9(1): 387–402.

46. P.J.A. Cock. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* 38(6): 1767–1771.

47. B. Ewing, L. Hillier, M.C. Wendl, and Phil Green. 1998. Base-Calling of Automated Sequencer Traces UsingPhred.I. AccuracyAssessment. *Genome Research* 8(3): 175–185.

48. Development-Core-Team, *R: A language and environment for statistical computing*, (Vienna: R.F.f.S. Computing, Editor, 2004).

49. M. Morgan, S. Anders, M. Lawrence, P. Aboyoun, H. Pagès, and R. Gentleman. 2009. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* 25(19): 2607-2608.

50. J.M. Catchen, A. Amores, P. Hohenlohe, W. Cresko, and J.H. Postlethwait. 2011. Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes, Genomes, Genetics* 1(3): 171–182.

51. R.C. Edgar. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19): 2460–2461.

52. Edgar, R. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

53. R.C. Edgar. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5): 1792–1797.

54. A. Stamatakis and N. Alachiotis. 2010. Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. *Bioinformatics* 26(12): i132-9.

55. J. Felsenstein, *Inferring Phylogenies* (Sunderland, Maryland: Sinauer Associates, Inc., 2004).

56. L.L. Knowles and L. Kubatko, eds. *Estimating Species Trees: Practical and Theoretical Aspects* (Hoboken: Wiley-Blackwell & Sons, Inc. 2010).

57. L.L. Knowles, Estimating Species Trees: Methods of Phylogenetic Analysis When There Is Incongruence across Genes, *Systematic Biology*, 2009. 58(5): 463–467.

58. C. Ané, B. Larget, D.A. Baum, S.D. Smith, and A. Rokas. 2007. Bayesian Estimation of Concordance among Gene Trees. *Molecular Biology and Evolution* 24(2): 412-426.

59. J.F. Wendel and J.J. Doyle, *Phylogenetic incongruence: window into genome history and molecular evolution, ch. 10*, in *Molecular systematics of plants II: DNA sequencing* D.E. Soltis, P.S. Soltis, and J.J. Doyle, Editors. (Norwell, Massachusetts: Kluwer Academic Publishers, 1998), 265–296.

60. K.C. Nixon. 2002. The Oak (Quercus) Biodiversity of California and Adjacent Regions. *USDA Forest Service Gen. Tech. Rep.*, PSW-GTR-184.

61. W. Trelease, *The American Oaks. Memoirs of the National Academy of Sciences*, 1924. 20: 1–255.