NRC Research Press

ARTICLE

Phylogenomic inferences from reference-mapped and de novo assembled short-read sequence data using RADseq sequencing of California white oaks (*Quercus* section *Quercus*)[1]

Sorel Fitz-Gibbon, Andrew L. Hipp, Kasey K. Pham, Paul S. Manos, and Victoria L. Sork

Abstract: The emergence of next generation sequencing has increased by several orders of magnitude the amount of data available for phylogenetics. Reduced representation approaches, such as restriction-sited associated DNA sequencing (RADseq), have proven useful for phylogenetic studies of non-model species at a wide range of phylogenetic depths. However, analysis of these datasets is not uniform and we know little about the potential benefits and drawbacks of de novo assembly versus assembly by mapping to a reference genome. Using RADseq data for 83 oak samples representing 16 taxa, we identified variants via three pipelines: mapping sequence reads to a recently published draft genome of *Quercus lobata*, and de novo assembly under two sets of locus filters. For each pipeline, we inferred the maximum likelihood phylogeny. All pipelines produced similar trees, with minor shifts in relationships within well-supported clades, despite the fact that they yielded different numbers of loci (68 000 – 111 000 loci) and different degrees of overlap with the reference genome. We conclude that both the reference-aligned and de novo assembly pipelines yield reliable results, and that advantages and disadvantages of these approaches pertain mainly to downstream uses of RADseq data, not to phylogenetic inference per se.

Key words: de novo clustering, GATK, phylogenetics, PyRAD, restriction-site associated DNA sequencing, variant discovery.

Résumé : L'avènement du séquençage de nouvelle génération a augmenté de plusieurs ordres de grandeur la quantité de données disponibles pour des études phylogénétiques. Des approches de réduction de la complexité, telles que le séquençage de l'ADN associé aux sites de restriction (RADseq), se sont avérées utiles pour les études phylogénétiques chez les espèces non-modèles à une vaste gamme de niveaux phylogénétiques. Cependant, l'analyse de ces jeux de données n'est pas uniforme et peu de choses sont connues quant aux bénéfices et désavantages potentiels de l'assemblage de novo par rapport à l'alignement sur un génome de référence. Au moyen de données RADseq pour 83 échantillons de chêne représentant 16 taxons, les auteurs ont identifié les variants à l'aide de trois pipelines : l'alignement des séquences sur une ébauche de génome publiée récemment pour le *Quercus lobata*, ainsi que l'assemblage de novo à l'aide de deux séries de filtres pour identifier les locus d'intérêt. Pour chaque pipeline, les auteurs ont déduit la phylogénie par une approche de vraisemblance maximale. Tous les pipelines ont produit des arbres semblables, avec des différences mineures des relations au sein de clades bien supportés, malgré le fait qu'ils ont produit un nombre assez différent de locus (68 000 –111 000 locus) et différents degrés de chevauchement avec le génome de référence. Les auteurs concluent que tant le pipeline reposant sur l'alignement sur génome de référence que ceux basés sur l'assemblage de novo produisent des résultats fiables. En définitive, les avantages et désavantages de ces approches concernent principalement les analyses envisagées avec les données RNAseq et non l'inférence phylogénétique en soi. [Traduit par la Rédaction]

Mots-clés : groupement de novo, GATK, phylogénétique, PyRAD, séquençage de l'ADN associé aux sites de restriction, identification de variants.

Introduction

Phylogenetic inference near the tips of the tree of life is often a problem of sampling sufficient numbers of loci (e.g., Corl and Ellegren 2013). This difficulty has become more evident in light of the past decade's efforts to account for lineage sorting and coalescent processes in the estimation of species trees from individual gene trees (Edwards 2009; Swenson and El-Mabrouk 2012). Next generation sequencing technologies make it possible to efficiently identify and genotype large numbers of variants for non-model organisms using reduced representation sequencing, such as restriction-site associated DNA sequencing (RADseq; Baird et al. 2008) or genotyping-by-sequencing (GBS; Elshire et al. 2011). These

Received 15 November 2016. Accepted 27 January 2017.

Corresponding Editor: Juan P. Jaramillo-Correa.

S. Fitz-Gibbon. Institute of Genomics and Proteomics, University of California, Los Angeles, CA 90095, USA.
A.L. Hipp.* The Morton Arboretum, 4100 Illinois Route 53, Lisle, IL 60532-1293, USA; The Field Museum, 1400 S Lake Shore Drive, Chicago, IL 60605, USA.
K.K. Pham. The Morton Arboretum, 4100 Illinois Route 53, Lisle, IL 60532-1293, USA; Department of Plant Biology, Michigan State University, East Lansing, MI 48824-1312, USA.
P.S. Manos. Department of Biology, Duke University, Durham, NC 27708, USA.
V.L. Sork. Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095-7239, USA; Institute of the Environment and Sustainability, University of California, Los Angeles, CA 90095-1496, USA.
Corresponding author: Victoria L. Sork (email: vlsork@ucla.edu).

*Andrew L. Hipp currently serves as a Guest Editor; peer review and editorial decisions regarding this manuscript were handled by Juan P. Jaramillo-Correa.

[1]This paper is part of a Special Issue entitled The Evolution of Tree Diversity.

datasets, which we treat inclusively under the term RADseq in this paper, are all based on relatively short sequence reads—typical datasets are based on 100-bp sequencing reads—that begin at a specific recognition site wherever it occurs in the genome (Baird et al. 2008; McCluskey and Postlethwait 2015). Thus the resulting sequences are anonymous, differing in this way from sequence-capture methods such as HybSeq (Weitemier et al. 2014) or anchored enrichment (Lemmon et al. 2012), which sequence loci targeted at the outset of the study. This anonymity is beneficial because little information about the study organism is needed at the outset of a project, but it also presents challenges in both alignment and phylogenetic analysis of datasets.

RADseq has emerged as a practical means of inferring phylogenetic trees among lineages of up to ca. 50 mya in depth (Eaton 2014; Hipp et al. 2013; Hou et al. 2015; Rubin et al. 2012), but analysis of RADseq datasets is not standardized and is perhaps unlikely ever to be fully standardized given the complexity and dimensions of these datasets (Ree and Hipp 2015). There are two general types of analytical approaches to aligning sequences and detecting variants for phylogenetic inference: (*i*) mapping sequence reads to a longer genome sequence and (*ii*) de novo clustering of sequence reads. In the first approach, which is appropriate for clades that are sufficiently closely related to a species with a sequenced genome (e.g., Hyma and Fay 2013; McCluskey and Postlethwait 2015; Nadeau et al. 2013), raw sequence reads are mapped to the genome using, for example, BOWTIE (Langmead et al. 2009), Stampy (Lunter and Goodson 2011), or BWA (Li 2013). Depending on the quality of annotation of the reference genome, this method has the potential to dramatically reduce the anonymity of the RADseq dataset, facilitating the identification of genes, estimation of linkage among markers, and distinguishing between orthologs and paralogs. GATK (McKenna et al. 2010), a commonly used toolkit for reference-aligned datasets, offers a robust way to detect variants, including insertions, and deletions relative to the reference genome, but it is unknown whether ascertainment bias associated with the use of a genome from a focal species creates a problem for construction of phylogenomic trees.

For most phylogenetic studies, the lack of a reference genome makes the second approach, de novo clustering of sequence reads, the only feasible approach to aligning raw sequence reads into a RADseq dataset. Several pipelines have been implemented for de novo clustering of RADseq data, including but not limited to Rainbow (Chong et al. 2012), RADtools (Baxter et al. 2011), Stacks (Catchen et al. 2013), PyRAD (Eaton 2014; Eaton and Ree 2013), AftrRAD (Sovic et al. 2015), and RADIS (Cruaud et al. 2016). Of these, Stacks appears to be by far the most widely used (based on citation number, 25 October 2016). Stacks, however, which was designed for linkage mapping and population genetic applications, is problematic for phylogenetics because it ignores indels, taking an "off-by-N" approach to clustering (see discussion in Eaton 2014; Eaton and Ree 2013). Of the remaining applications, PyRAD appears to be most widely used for creation of phylogenetic applications. PyRAD addresses a number of problems expected in analysis of divergent sequences: insertion–deletion events, the high potential for conflating orthology with paralogy, and efficient global clustering among phylogenetically structured populations (Eaton 2014). But it cannot make use of gene identity or position, as a mapping approach can.

Given that each of these two approaches has benefits and drawbacks, it would be useful to know whether they produce similar phylogenetic trees. A few researchers have utilized genetic mapping to screen loci identified by de novo clustering (Reitzel et al. 2013; Wagner et al. 2013), but we are not aware of any study comparing phylogenetic inferences from de novo clustering of RADseq reads to variant identification by mapping of RADseq reads to a genome. The goal of this paper is to compare phylogenomic trees of a set of California endemic white oak species (*Quercus* section *Quercus*) based on the reference-aligned versus de novo assembled short-read sequence data generated by the same restriction-site associated DNA sequencing methods. Oaks provide an excellent opportunity for this type of analysis because two research teams (one working on species-wide phylogeny of North American oaks and the other focusing on the genomics and introgression of California oaks) have generated separate datasets sequenced in the same manner, which when combined provide a powerful dataset for assessment. In this comparison, we utilize two pipelines for variant discovery (GATK for variant-calling from mapped reads, PyRAD for de novo clustering) followed by maximum likelihood on the concatenated data matrices in RAxML 8.2.0 (Stamatakis 2014), utilizing comparable filters and criteria for both pipelines. We illustrate the features of each approach and identify their limitations.

Materials and methods

Study species and samples

The ecology and evolutionary biology of California white oaks (*Quercus* section *Quercus*: Fagaceae) are described in Nixon (2002), Pavlik et al. (1995), Ortego et al. (2015), and Sork et al. (2016*a*). The specimens and genotypes used to construct the phylogenetic trees (Table 1) were collected for two separate studies based on RADseq data: a phylogenetic analysis of California white scrub oak species (Kim et al. in review), and a phylogeny of North American white oaks (McVay et al. 2017). Collectively, we used 83 samples from 11 California endemic white oak taxa, one California non-endemic white oak (*Q. engelmannii*), three outgroup white oak species from other geographic regions (*Q. alba*, *Q. robur*, and *Q. stellata*), and one distant outgroup species from a different section *Lobatae* (*Q. kelloggii*).

Genomic sequencing

RADseq DNA extraction, library preparation, and sequencing were conducted as presented previously by A. Hipp and colleagues (Cavender-Bares et al. 2015; Hipp et al. 2014). Using the same genomic methods, we sequenced 58 samples from the Sork Lab independently from the 25 Hipp/Manos samples. Briefly, DNA for all RADseq samples was extracted from fresh or frozen material using the DNeasy plant extraction protocol (DNeasy, Qiagen, Valencia, Calif.) in the home laboratories. DNA extractions were gel-quantified in agarose by visual comparison with the New England BioLabs 100 bp DNA Ladder (NEB, Ipswich, Mass.). Extraction concentrations ranged from 5 to 10 ng DNA/μL extraction. RAD sequencing library preparation for both sets of specimens was conducted at Floragenex, Inc. (Portland, Ore., USA) following the methods of Baird et al. (2008) with *Pst*I. RAD libraries were barcoded by individual and multiplexed on an Illumina Genome Analyzer IIx at Floragenex (samples from The Morton Arboretum) or an Illumina HiSeq 2000 at the Broad Stem Cell Research Center (BSCRC) at UCLA (samples from UCLA). Sequencing reads were 100 bp in length, single end initially but after removal of the barcode and recognition sequence, most of the analyzed sequences were 86 bp long, but some were 85 bp. The average number of sequences per sample was 1 740 000 (minimum 367 788). Quality was checked with FastQC (Andrews 2010).

Reference-mapped variant discovery

We used a collapsed version of the reference genome (v0.5, Sork et al. 2016*b*; available from https://valleyoak.ucla.edu/genomicresources/) for which haplotype redundancy had been aggressively removed. To facilitate variant calling, we also concatenated the 40 158 contigs, ordered by length and separated by runs of 10 Ns, into eight pseudocontigs of approximately 100 MB in length. The sequencing reads were aligned using BWA-MEM v.0.7.12-r1039 (Li 2013). Identification of targeted genome regions and variant discovery were done with GATK v3.6-0-g89b7209 (McKenna et al. 2010), as follows. We ran CallableLoci on each sample with default parameters, except that minimum read depth was set to 5 and maximum read depth was set to 100 (minDepth 5 and maxDepth

Table 1. List of oak sample identification code, taxonomic identification, and spatial locations.

Phylogeny IDs	Taxa	County*	Latitude	Longitude
OAK-MOR-204	*Quercus alba*	Sangamon, Ill.	39.9348	–89.8016
OAK-MOR-507	*Quercus berberidifolia*	Lake	39.1984	–123.0560
OAK-MOR-508	*Quercus berberidifolia*	Solano	38.4117	–122.0492
VLS-QB-10.3	*Quercus berberidifolia*	San Diego	33.1970	–116.5966
VLS-QB-12.1	*Quercus berberidifolia*	San Diego	33.2105	–116.5162
VLS-QB-14.3	*Quercus berberidifolia*	Riverside	33.7059	–116.7548
VLS-QB-18.6	*Quercus berberidifolia*	Riverside	33.6486	–117.4102
VLS-QB-20b.2	*Quercus berberidifolia*	Los Angeles	34.7520	–118.7151
VLS-QB-20b.5	*Quercus berberidifolia*	Los Angeles	34.7520	–118.7151
VLS-QB-20b.8	*Quercus berberidifolia*	Los Angeles	34.7520	–118.7151
VLS-QB-34.5	*Quercus berberidifolia*	Santa Barbara	34.7200	–119.9570
VLS-QB-59.1	*Quercus berberidifolia*	Butte	39.7824	–121.7308
VLS-QB-66b.8	*Quercus berberidifolia*	Mariposa	37.7430	–120.2444
VLS-QB-NEW.3	*Quercus berberidifolia*	Orange	33.5958	–117.8371
VLS-QB-NEW.4	*Quercus berberidifolia*	Orange	33.5958	–117.8371
VLS-QB-NEW.5	*Quercus berberidifolia*	Orange	33.5958	–117.8371
VLS-QB-PR.315	*Quercus berberidifolia*	San Diego	33.2253	–117.0271
VLS-QB-SD.1	*Quercus berberidifolia*	San Diego	33.1697	–117.2834
VLS-QB-SD.3	*Quercus berberidifolia*	San Diego	33.1697	–117.2834
VLS-QB-SD.4	*Quercus berberidifolia*	San Diego	33.1697	–117.2834
OAK-MOR-715	*Quercus cornelius-mulleri*	—	—	—
VLS-QCM-9.1	*Quercus cornelius-mulleri*	San Diego	33.0858	–116.5084
VLS-QCM-11.7	*Quercus cornelius-mulleri*	San Diego	33.2207	–116.4576
VLS-QCM-16.4	*Quercus cornelius-mulleri*	San Bernardino	33.9980	–116.0599
VLS-QCM-17b.11	*Quercus cornelius-mulleri*	San Bernardino	34.1413	–116.4760
OAK-MOR-725	*Quercus cornelius-mulleri*	San Bernardino	34.2761	–116.8254
OAK-MOR-531	*Quercus douglasii*	El Dorado	38.6962	–120.8873
OAK-MOR-696	*Quercus douglasii*	Solano	38.4117	–122.0492
OAK-MOR-90	*Quercus douglasii*	Solano	38.4541	–121.8410
VLS-QDO-SW.10	*Quercus douglasii*	Santa Barbara	34.6911	–120.0489
VLS-QDO-SW.4	*Quercus douglasii*	Santa Barbara	34.6886	–120.0454
VLS-QDO-SW.1	*Quercus douglasii*	Santa Barbara	34.6892	–120.0457
PM-141	*Quercus dumosa*	Los Angeles	34.0180	–118.3810
VLS-QDU-19.2	*Quercus dumosa*	Riverside	33.7925	–117.3475
VLS-QDU-SBBG.1	*Quercus dumosa*	Santa Barbara	34.4590	–119.7094
OAK-MOR-336	*Quercus durata* var. *durata*	Napa	38.5127	–122.4830
VLS-QDD-37b.2	*Quercus durata* var. *durata*	San Luis Obispo	35.3463	–120.6448
VLS-QDD-45.5	*Quercus durata* var. *durata*	Santa Clara	37.1757	–121.8644
VLS-QDD-48.2	*Quercus durata* var. *durata*	Stanislaus	37.4089	–121.4166
VLS-QDD-48.7	*Quercus durata* var. *durata*	Stanislaus	37.4089	–121.4166
VLS-QDD-50b.1	*Quercus durata* var. *durata*	Sonoma	38.5639	–122.6854
VLS-QDD-62.4	*Quercus durata* var. *durata*	Placer	38.9015	–121.0611
VLS-QDD-62.6	*Quercus durata* var. *durata*	Placer	38.9015	–121.0611
PM-1	*Quercus durata* var. *gabrielensis*	San Bernardino	34.2325	–117.7573
VLS-QDG-45b.10	*Quercus durata* var. *gabrielensis*	Los Angeles	34.1775	–118.095
VLS-QDG-45b.12	*Quercus durata* var. *gabrielensis*	Los Angeles	34.1548	–117.8439
VLS-QDG-45b.9	*Quercus durata* var. *gabrielensis*	Los Angeles	34.1949	–118.1141
VLS-QDG-MtB.1	*Quercus durata* var. *gabrielensis*	Los Angeles	34.2719	–118.1586
VLS-QDG-MtB.5	*Quercus durata* var. *gabrielensis*	Los Angeles	34.2815	118.1808
VLS-QDG-MtB.8	*Quercus durata* var. *gabrielensis*	Los Angeles	34.2424	–118.1885
VLS-QE-15.2	*Quercus engelmannii*	San Diego	32.7305	–116.8764
VLS-QE-16.1	*Quercus engelmannii*	San Diego	32.7430	–116.8107
VLS-QE-17.1	*Quercus engelmannii*	San Diego	32.8120	–116.7782
VLS-QE-Ju.161	*Quercus engelmannii*	San Diego	33.0490	–116.3442
OAK-MOR-422	*Quercus garryana* var. *breweri*	El Dorado	38.6785	–120.8125
OAK-MOR-723	*Quercus garryana* var. *garryana*	Humboldt	41.1891	–123.6826
PM-218	*Quercus john-tuckeri*	Los Angeles	34.4755	–118.1166
PM-219	*Quercus john-tuckeri*	Los Angeles	34.4755	–118.1166
OAK-MOR-709	*Quercus john-tuckeri*	Kern	34.6284	–119.1929
VLS-QJT-21b.4	*Quercus john-tuckeri*	Kern	34.8716	–119.2719
VLS-QJT-21b.6	*Quercus john-tuckeri*	Kern	34.8716	–119.2719
VLS-QJT-22.3	*Quercus john-tuckeri*	Santa Barbara	34.6781	–119.3680
PM-266	*Quercus kelloggii*	San Diego	33.1583	–116.6744
PM-316	*Quercus kelloggii*	El Dorado	38.7941	–120.2351
OAK-MOR-123	*Quercus lobata*	San Benito	36.6167	–120.8500
OAK-MOR-502	*Quercus lobata*	El Dorado	38.6962	–120.8873

Table 1 (*concluded*).

Phylogeny IDs	Taxa	County*	Latitude	Longitude
OAK-MOR-697	*Quercus lobata*	—	—	—
VLS-QL-COL.1	*Quercus lobata*	Tehama	40.0628	–122.4929
VLS-QL-HV.6	*Quercus lobata*	Ventura	34.1530	–118.9073
VLS-QL-OAK-E	*Quercus lobata*	Madera	37.2636	–119.6904
VLS-QL-PAY.6	*Quercus lobata*	Tehama	40.3130	–121.9984
VLS-QL-SW.786	*Quercus lobata*	Santa Barbara	34.6882	–120.0362
VLS-QL-UCD.8	*Quercus lobata*	Yolo	38.5341	–121.7519
VLS-QP-Cat.4	*Quercus pacifica*	Los Angeles	33.3548	–118.3529
VLS-QP-Catb.19	*Quercus pacifica*	Los Angeles	33.3548	–118.3529
PM-3b	*Quercus pacifica*	Los Angeles	33.3755	–118.4180
VLS-QP-SR33.3	*Quercus pacifica*	Santa Barbara	33.9793	–120.0472
OAK-MOR-470	*Quercus pacifica*	Santa Barbara	33.9825	–120.0736
VLS-QP-SRb.2	*Quercus pacifica*	Santa Barbara	34.0071	–120.0511
VLS-QP-SR.7	*Quercus pacifica*	Santa Barbara	34.0071	–120.0511
VLS-QP-SC3.20	*Quercus pacifica*	Santa Barbara	34.0270	–119.6949
OAK-MOR-392	*Quercus robur*	Moscow, Russia	55.8386	37.6005
OAK-MOR-340	*Quercus stellata*	Crawford Co., Mo.	38.0349	–91.5203

*All counties are inside California, USA, unless state and country are listed.

100). An overall set of callable regions, hereafter referred to as reference-mapped loci, included all regions of exactly 86 bp determined to be callable in at least four samples. HaplotypeCaller variant discovery was run using emitRefConfidence GVCF mode separately for each sample. The whole cohort was concurrently genotyped over the reference-mapped loci regions using GenotypeGVCFs with a minimum confidence of 30. Based on inspection of variant calls and read alignments (Robinson et al. 2011), we chose to only filter the variants by QD as well as requiring them to be biallelic. VariantFiltration was used to apply a QD < 2.0 filter tag. VCFtools v0.1.15 (Danecek et al. 2011) was used to filter non-biallelic variants and to calculate transition to transversion ratios. This filtering decreased the number of variants from 522 199 (Ts/Tv = 1.47) to 477 819 (Ts/Tv = 1.52). For each sample, we created the sequences for phylogenetic analysis by concatenating the reference-mapped loci with the following adjustments: (*i*) Sites with filtered variants, including non-biallelic variants were converted to Ns for all samples (these sites are filtered out later by the phylogenic analysis software). (*ii*) For each sample, GATK loci not considered callable by the above CallableLoci step were converted to Ns. (*iii*) Sequences were printed with passing variants applied according to homozygous genotype or if heterozygous IUPAC ambiguity codes were used. (*iv*) Homozygous indel variants were included using dashes to maintain the overall alignment. However, there is no IUPAC ambiguity code for heterozygous indels and so they were instead encoded as Ns. A Phylip-formatted file was created as input for phylogenetic analysis.

De novo variant discovery

Data were analyzed using the PyRAD pipeline (Eaton 2014; www.dereneaton.com/software). Sequences were clustered first by individual using VSEARCH (Rognes et al. 2016), which allows sequences within clusters to vary in indels, nucleotide polymorphisms, and sequences strand (direction). After clustering, heterozygosity and sequencing error were jointly estimated from the base counts observed across all sequences, sites, and clusters using the likelihood equation of Lynch (2008). Consensus sequences of each cluster were generated for each individual. Heterozygous positions in the consensus sequences were called using a binomial probability based on the global (across samples) error rate and heterozygosity estimates. Bases that could not be assigned with ≥95% probability were treated as unknown (N). Alleles that differ only by an indel in the PyRAD dataset were coded as the base that is present; thus allele variation based on indels only within an individual is invisible without additional analysis. Any locus possessing more than two haplotypes within individuals after correcting for sequencing errors was discarded, under the assumption that it included one or more paralogous sequences. Consensus sequences were clustered among individuals to generate a data matrix for each locus, using the same clustering parameters as used within individuals. We set the minimum depth of reads per within-sample cluster to 6, maximum number of sites in a read which can have a quality score of less than 20 to 4, clustering threshold (percent similarity) to 0.85, and the minimum number of samples in each across-sample cluster to 4.

For comparison with the reference-mapped pipeline, we used two sets of PyRAD parameters, which we refer to as the PyRAD-default and the PyRAD-relaxed parameter sets. In the PyRAD-default parameter set, we set the maximum number of individuals with a shared heterozygous site (MaxSH) in an across-sample cluster to 3 and the maximum number of heterozygous sites in consensus sequences (maxH) to 5. We relaxed these latter two parameters in the PyRAD-relaxed parameter set: MaxSH was set to the number of individuals in the cluster, and maxH was set to 85. These parameters, which are intended to weed out hypervariable sites or sites that are incorrectly clustered, were relaxed in the PyRAD-relaxed parameter set because the filters are not implemented in our reference-mapped pipeline. All other settings used default values. The resulting loci from both clustering runs were concatenated and exported for analysis as PHYLIP files. Missing loci were treated as Ns.

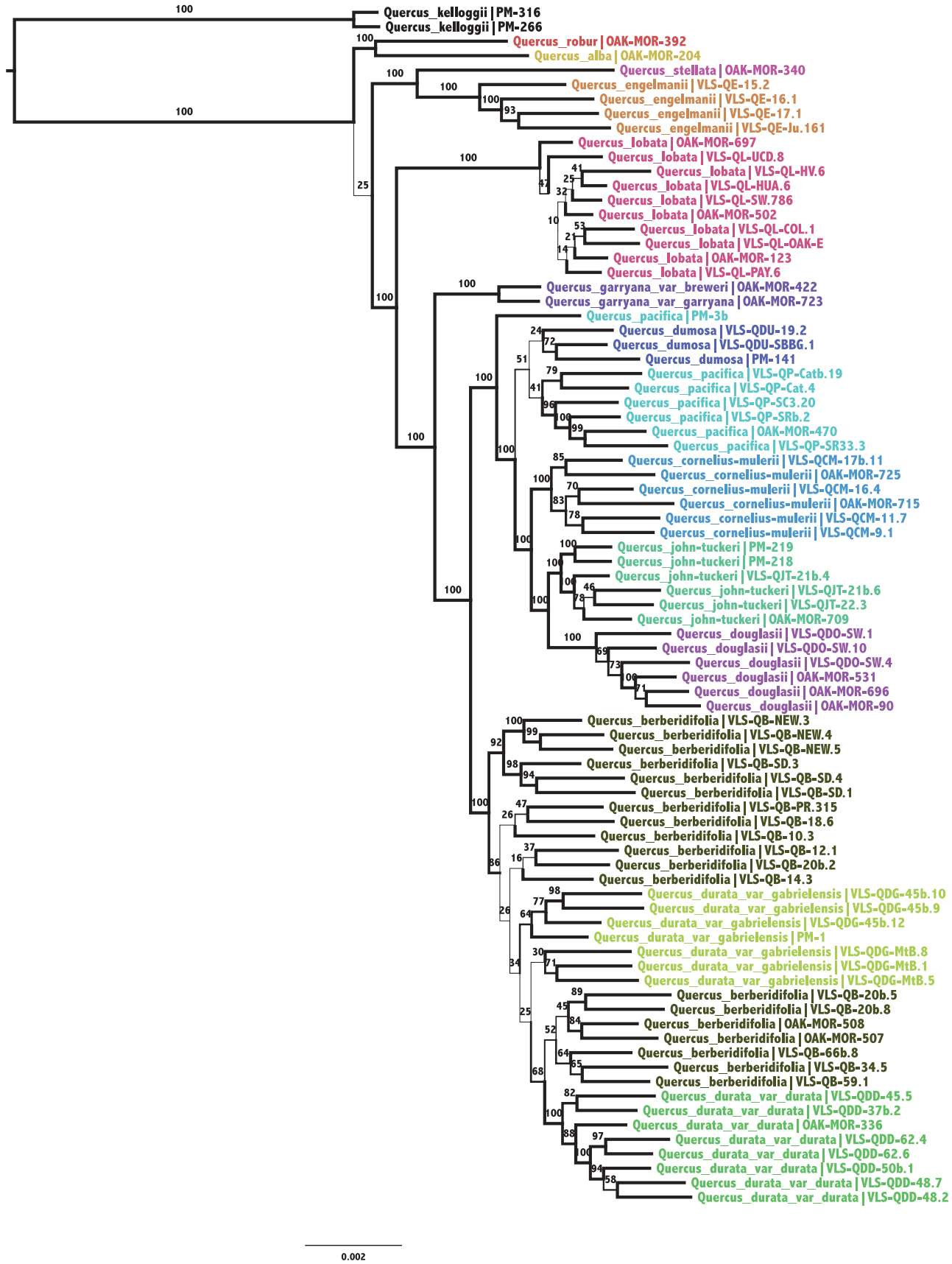
Phylogenetic inference

We used the same tree inference methods for each of the aligned sequence sets from the three pipelines. Maximum likelihood analysis was conducted with RAxML 8.2.0 (Stamatakis 2014), using the -f a option with -N 100. This runs 100 rapid bootstrap replicate trees and adds the resulting support values to the best-scoring of 10 maximum likelihood trees built from10 distinct maximum parsimony starting trees. We used the GTRCAT model of nucleotide substitution and specified the two *Q. kelloggii* samples as outgroups. Trees were plotted with FigTree v1.4.2. (http://tree.bio.ed.ac.uk/software/figtree/), as well as the ape (Paradis et al. 2004), and phytools (Revell 2012) packages in R version 3.3.2 ("Sincere Pumpkin Patch", R Core Team 2016).

Loci analyses

Consensus sequences for the PyRAD loci were mapped to the reference genome using BWA-MEM (Li 2013). Overlapping and adjacent loci were merged with BEDTools v2.25.0 (Quinlan 2014). Intersections between PyRAD loci and reference-mapped loci were determined using BEDTools and plotted using BioVenn (Hulsen et al. 2008). GATK's DepthOfCoverage was used to identify regions with 1, 2, or more consensus PyRAD loci mapped. Measures of spacing

Fig. 1. Phylogenetic tree using a maximum likelihood analysis of white oak sequences (*Quercus* section *Quercus*) and one red oak sequence (*Q. kelloggii*), using GATK for variant discovery relative to a reference genome, *Q. lobata* v0.5 (Sork et al. 2016*b*).



between mapped loci were done with BEDTools, using the original reference genome contig coordinates rather than the concatenated pseudo-contig coordinates, and distances to the ends of contigs were ignored.

Results and discussion

General phylogenetic trends

The phylogenetic trees inferred from reference-mapped (Fig. 1) and de novo assembled short-read sequence data (Figs. 2, 3) show essentially the same structure, with only minor differences among trees from fundamentally different analysis methods and no evidence within trees of a laboratory group effect in the position of samples within a clade. It has been argued that RADseq data "are essentially one-off datasets" (Harvey et al. 2016) that cannot be readily repurposed for downstream studies. Extractions and sequencing in this study were conducted in different facilities, while libraries were all prepared at Floragenex using the same enzyme (*Pst*I) and methods (described above). Thus this does not constitute a strong test of data combinability, but it demonstrates as we have shown previously (Hipp et al. 2014) that laboratories can collaborate using RADseq. It remains to be seen how difficult it is to combine RADseq data from different research teams when library preparation is also conducted in different laboratories. Our suspicion is that with true RADseq (sensu Baird et al. 2008), where size selection does not select for different regions of the genome, data combinability will not be a great problem. This advantage of sequence data, which is not true for many genetic markers such as microsatellite or AFLPs, allows systematists to collaborate on large-scale phylogenetic studies and repurpose data for years to come.

The phylogenetic trees presented here show similar patterns to those of Kim et al. (in review). Our study, like theirs, reveals that the endemic California scrub oaks, which are all related to each other, as suggested by Nixon (2002) and observed by Ortego et al. (2015), cluster into two main clades: the "berberidifolia" clade, a strongly supported clade comprised of *Q. berberiidifolia* and both varieties of *Q. durata* with *Q. berberidifolia* and *Q. durata* var. *gabrielensis* not reciprocally monophyletic; and the "dumosa" clade, a strongly supported clade comprised of *Q. cornelius-mulleri*, *Q. douglasii*, *Q. dumosa*, *Q. john-tuckeri*, and *Q. pacifica*, with *Q. dumosa* and *Q. pacifica* not consistently differentiated from each other. As discussed in Kim et al. (in review), the most unexpected finding is that *Q. douglasii*, a California endemic tree oak, is not a sister taxa to the other common tree oak *Q. lobata* as suggested by Nixon, but is instead embedded within the scrub oak species. All of these studies also reveal that the scrub oaks share a common ancestor more recently than they do with *Q. garryana* and *Q. lobata*. The relationships among the California endemic oaks are discussed in greater detail in other studies by our group (Kim et al. in review; Sork et al. 2016*a*). Here, we present these data primarily to compare the phylogenetic implications of the two primary alternatives to identifying loci from reduced representation sequencing methods such as RADseq.

Comparison among phylogenetic trees

Trees constructed with variants discovered using reference-aligned assembly versus de novo assembly (Fig. 4) and different filters applied during de novo assembly (Fig. 5) are largely congruent. All points of incongruence are within strongly supported clades, with one exception: a single specimen of *Q. pacifica* (PM-3b, Santa Catalina Island, Santa Barbara Co., Calif.) is found with 100% bootstrap support outside of one of the two major scrub oak clades both in the reference-mapped tree and in the PyRAD-relaxed tree, while in the PyRAD-default tree, this sample is found well within the major clade and neighboring the clade with the rest of the *Q. pacifica* samples. Why this sample does not cluster with the other samples on Santa Catalina Island for the two larger datasets is not clear. The PyRAD-default filters, designed to exclude regions with high heterozygosity, may be removing signs of introgression, recent divergence, or founder effect. Given that *Q. dumosa* and *Q. pacifica* have very low bootstrap support, the inconsistent clustering suggests very recent divergence. The remaining points of incongruence are mostly within species, with a few exceptions. In all trees, for example, *Q. durata* var. *gabrielensis* is polyphyletic, while *Q. durata* var. *durata* forms a strongly supported clade derived from within a *Q. berberidifolia* grade. However, the reference-mapped tree recovers just two *Q. durata* var. *gabrielensis* clades, which the two de novo assembly trees resolve the variety into four small clades. The low support for this variety and its alternative resolutions under different analysis approaches raise questions about the validity of the variety (see Kim et al. in review). Overall, the incongruent cases among trees seem to reflect the sensitivity of all pipelines to the evolutionary history of the lineages, rather than systematic differences in the pipelines used to generate the trees.

We further explored the impact of the PyRAD paralog filters by applying them to the final loci resulting from the reference-mapped pipeline. We removed loci with more than five heterozygous sites (similar to maxH = 5) and removed loci for which more than three individuals shared a single heterozygous site (similar to maxSharedH = 3) and rebuilt the tree (Fig. S1[2]). Implementing these filters had little effect on topology or congruence although the relative positions of the *Q. lobata* and *Q. engelmannii* clades were altered, but with zero bootstrap support. In both the PyRAD and the reference-mapped pipelines, implementing the more stringent filters reduced the total amount of data available for phylogenetic analysis (Table 2), resulting in generally lower branch support (Figs. 3 and 1 versus Fig. 2 and Fig. S1[2]). These effects mirror the observation that "cleaning up" RADseq datasets to reduce missing data often has the effect of simply reducing phylogenetic support (Ree and Hipp 2015). The differences between the two sets of parameters we test in this study were not great. However, for datasets smaller than the one used here, the choice of parameters might have a bigger affect on the tree.

The effects of generating reference-mapped pipeline datasets with two alternate references was also negligible (Figs. S2, S3)[2]. The highly redundant *Q. robur* draft genome and the more redundant (non-collapsed) version of the *Q. lobata* genome, version 1.0, both produce essentially the same topology as our initial analyses (Figs. S2, S3)[2]. Mapping to *Q. robur* resulted in fewer final loci than mapping to *Q. lobata* as would be expected using a more distant reference (Table 2). Interestingly, mapping to the more redundant and also more complete version of the reference genome, *Q. lobata* v1.0, also resulted in fewer final loci. Because *Q. lobata* v1.0 is thought to be 10% more complete than *Q. lobata* v0.5 (Sork et al. 2016*b*), additional loci are expected. At the same time, 30% of the genome in the v1.0 assembly is thought to be represented twice in the v1.0 assembly (Sork et al. 2016*b*), which is likely to account for the lower coverage of loci (Table 2). Most importantly, our analysis suggests that the use of reference-mapped or de novo assemblies is not a key issue for producing a robust phylogenetic tree, at least at fine phylogenetic scales. Instead, all of these methods and the filters used in variant discovery affects the number of loci and selection of loci and that these subsamples of loci will yield slightly different trees. Thus, it seems if the bootstrap support distinguishing clusters is not strong, phylogenetic trees based on different pipelines will also vary depending on the assumptions and pipelines used. The variance across trees does not seem to be due primarily to the use of a reference genome versus de novo

[2]Supplementary data are available with the article through the journal Web site at http://nrcresearchpress.com/doi/suppl/10.1139/gen-2016-0202.

Fig. 2. Phylogenetic tree using the same method and samples as Fig. 1, but using PyRAD approach with default settings as described in text.

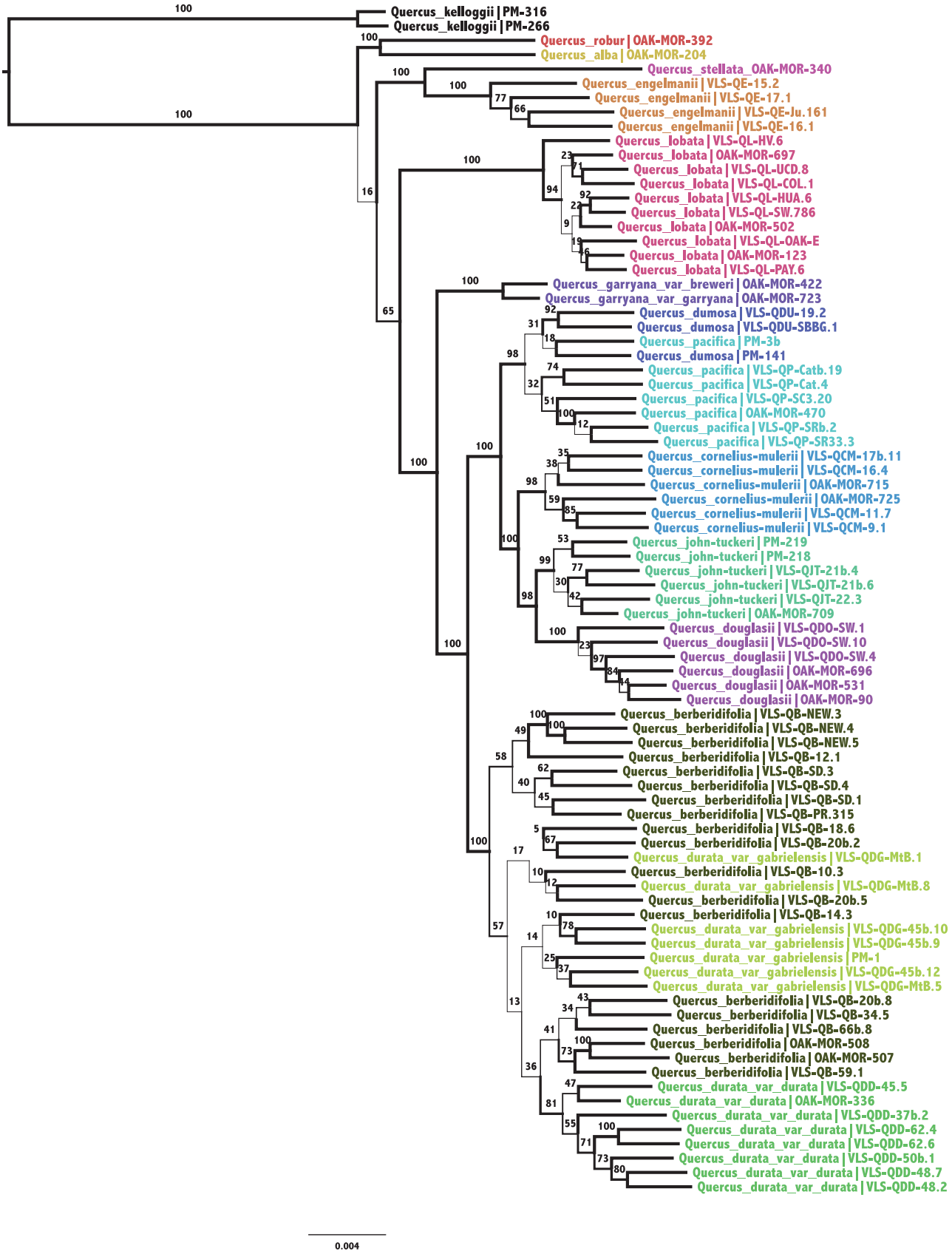


Fig. 3. Phylogenetic tree using the same method and samples as Fig. 1, but using PyRAD approach with "relaxed" settings as described in text.

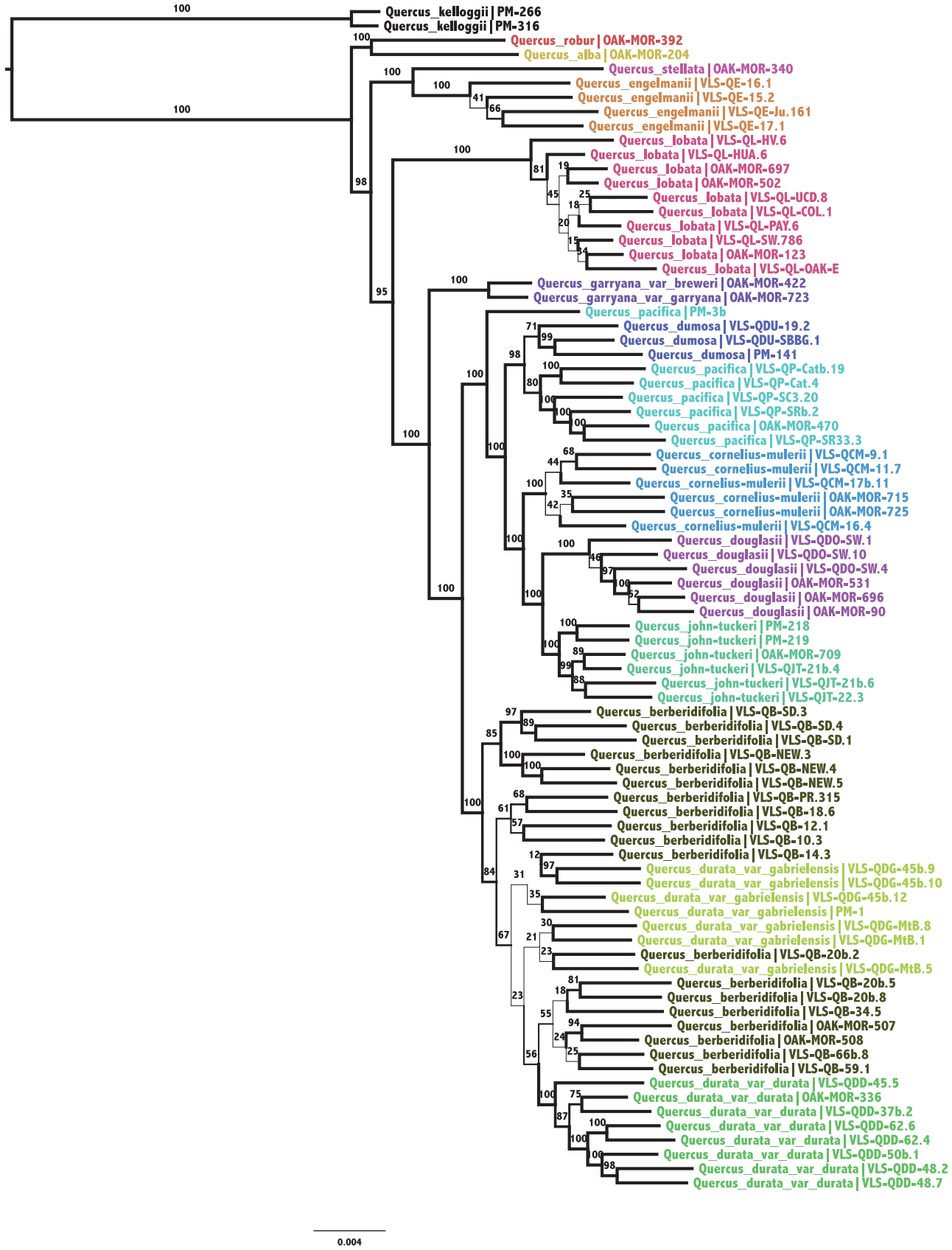
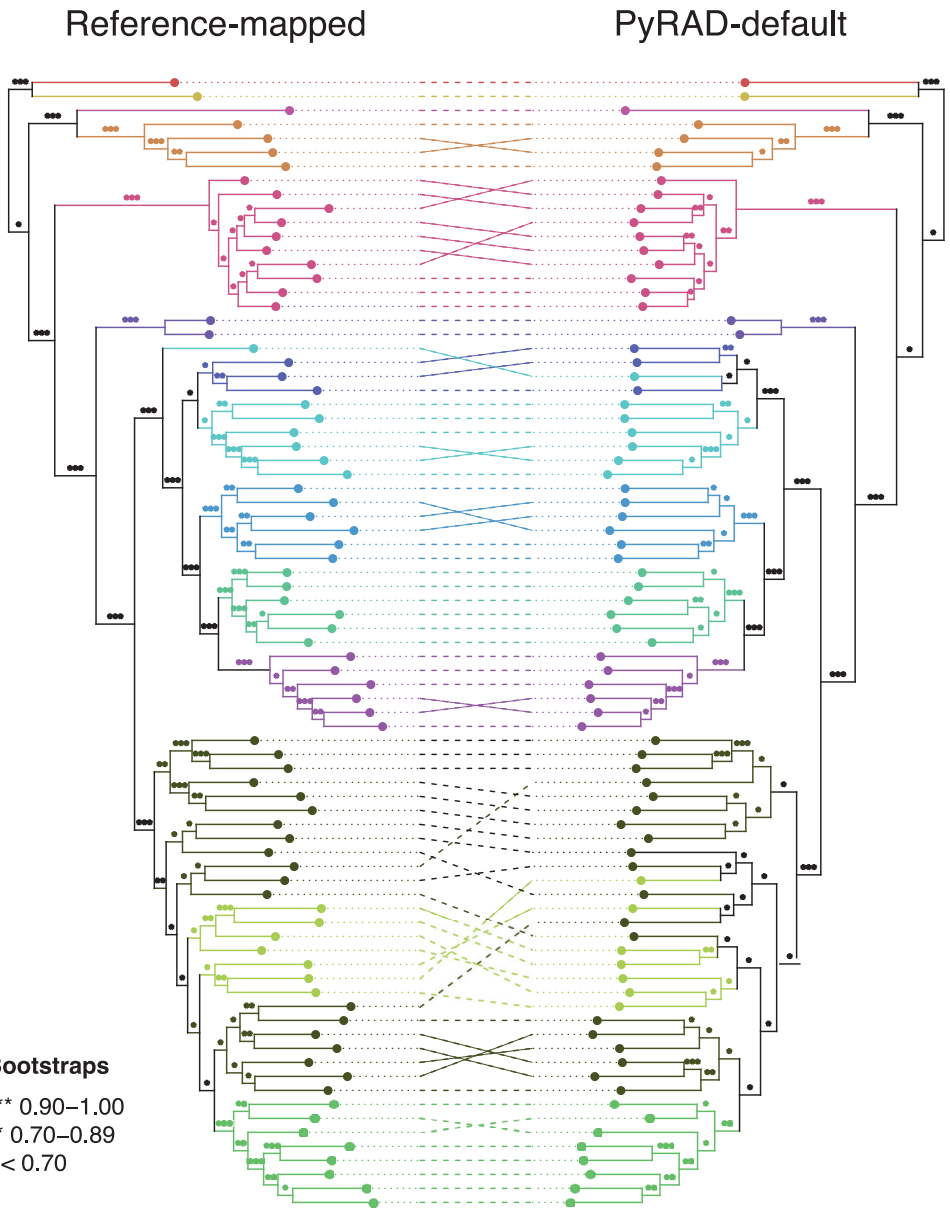


Fig. 4. Co-phylogeography plots comparing Fig. 1 tree versus Fig. 2 tree. Species colors are the same as in Figs. 1–3.



assembly because we also see variation when we change the filters used for the PyRAD-based trees.

It is important to note that while our resolution of the Californian oaks is very much in keeping with other work we have been conducting, our overall topology of oaks is not in keeping with what we know about the topology of oaks as a whole. This discrepancy is presumably due to the skewed nature of our taxon sampling. In analyses of alternative clustering parameters for both de novo and reference-mapped datasets, we encountered variance among trees in the relative placement of *Q. alba*, *Q. robur*, and the *Q. engelmannii*/*Q. stellata* clade (Figs. S4–S6)[2]. Our work with broader samples of oak species demonstrates that all California white oaks other than *Q. engelmannii* form a clade sister to the remainder of the white oaks, with the possible exception of the Eurasian *Q. robur* and allies (Hipp et al. 2014; Pearse and Hipp 2009). Placement of these Eurasian white oaks is unstable, tacking between a position sister to the white oaks as a whole (Pearse and Hipp 2009) or sister to an eastern North American white oak group, possibly centered on *Q. alba* (Hipp et al. 2014). This question is currently being investigated (McVay et al. 2017). Thus, while the topologies for the California white oak species in this study seem reliable, one should be cautious about interpreting the findings for the remainder of the tree when sampling is strongly skewed to the in-group.

Comparison of loci generated by each pipeline

The PyRAD-default pipeline yielded the smallest number of loci contributing to the final sequence alignment, ~68 000. The reference-mapped pipeline had 10% more loci, ~75 000, while the PyRAD-relaxed pipeline had by far the most, 64% more than the PyRAD-default pipeline, ~111 000 (Table 2). To compare the set of loci used in each method, the PyRAD loci were mapped to the reference genome and their mapped positions were compared with the set of loci produced by the reference-mapped pipeline. We found that 15% of the PyRAD-default loci and 13% of PyRAD-relaxed loci did not map anywhere on the reference genome, which is similar to the 14% of raw reads that did not map to the reference genome (Table S1[2]). This percentage of unmapped reads is expected as the reference genome, *Q. lobata* assembly version 0.5, was aggressively "collapsed" to remove haplotype redundancy, resulting in concurrent loss of some

Fig. 5. Co-phylogeography plots comparing PyRAD-default (Fig. 2) versus PyRAD-relaxed settings (Fig. 3). Species colors are the same as in Figs. 1–3.

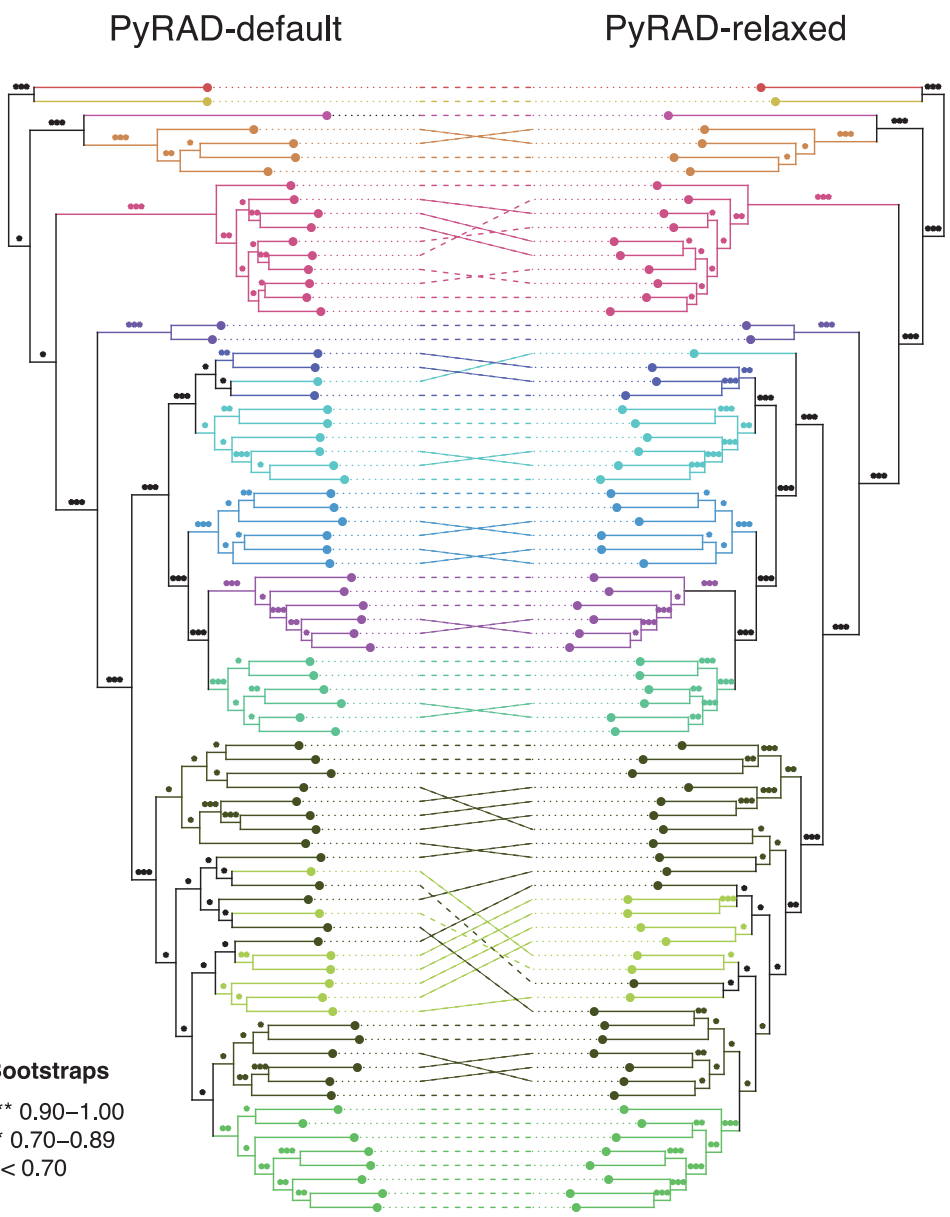


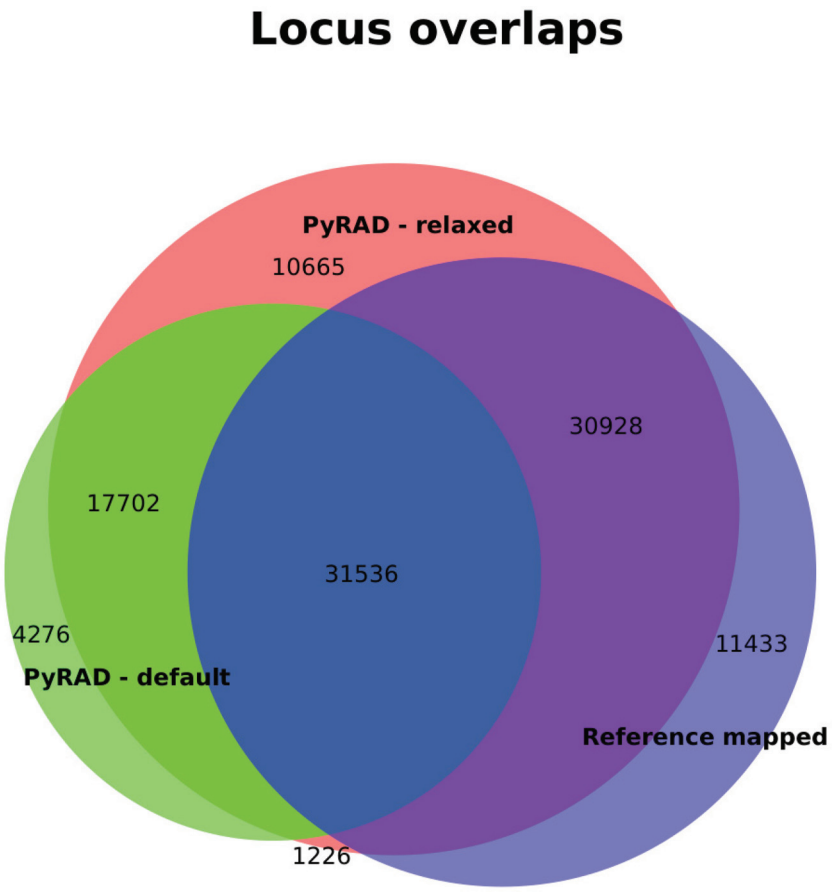
Table 2. Final sets of loci and their variation produced by the different pipelines, including mapping to three different reference genomes.

	Total loci	Total length (bps)	RAxML distinct alignment patterns	RAxML gaps and completely undetermined characters
PyRAD-default	68 071	5 941 401	952 330	74.57%
PyRAD-relaxed	111 505	9 751 753	2 019 299	58.64%
Reference-mapped *Q. lobata* v0.5	75 123	6 477 357	1 229 090	45.91%
Reference-mapped *Q. robur*	62 230	5 364 181	987 368	47.36%
Reference-mapped *Q lobata* v1.0	73 826	6 361 890	1 107 808	51.05%

Note: The number of distinct alignment patterns as well as gaps and undetermined characters are as reported by the RAxML software (Stamatakis 2014).

genome regions. Match rates of mapped and unmapped reads to *Arabidopsis* and other genomes suggest at least two thirds of the unmapped reads are likely to be *Quercus* reads, while others are possibly human and bacterial contamination (data not shown). For the loci that map to the reference genome, 29% are shared between the three pipelines (Fig. 6). Relaxing the PyRAD paralog filters adds another 29% of loci shared with the reference-mapped loci and adds only 10% of unshared loci. Therefore the PyRAD filters, maxSharedH and maxH, mostly remove loci that are retained in the reference-mapped pipeline.

Fig. 6. Overlap of final loci sets resulting from the three variant discovery pipelines. PyRAD-retained loci were mapped to the reference genome for comparison with the reference-mapped pipeline. Not represented are 13% of PyRAD-relaxed loci and 15% of PyRAD-default loci that failed to map to the reference genome. Also 5% of PyRAD loci that overlapped on the reference genome (usually globally) were counted as single loci. Figure created using BioVenn (Hulsen et al. 2008). See Table 2 for sample sizes.



Mapping to a reference genome provides sequence context information for the individual loci and allows a greater possibility to control for non-independent variation. Indeed, RADseq loci are often obtained from both sides of a restriction cut site leading to strongly linked pairs of loci. We quantified this issue by counting the number of loci with a neighbor exactly six base pairs away and found the rate of these co-located final loci varies from 38% of the PyRAD-default set to 74% of the reference-mapped set (Table 3; Fig. 7). The differences are not due to systematic pipeline effects causing altered spacing as similar numbers are seen if we count all loci within 10 base pairs as co-located. Both sets of PyRAD loci are less frequently co-located than would be expected based on the frequency seen in the reference-mapped set of loci. Thus PyRAD must be filtering co-located loci at a higher rate than singly placed loci. This tendency suggests systematic differences in coverage or quality correlated to whether one or both sides of a cut site are represented, perhaps related to average distance between cut sites and the fragment size selection process. Knowing the frequency of co-located loci may improve any analysis that otherwise assumes random independent sampling of loci, but may not be a concern for phylogenetic analysis.

By examining a visual representation of the mapped positions of final loci along the genome we get an overview of the distribution, density, and concordance for each of the pipelines. Figure 7 shows multiple scales of genome regions with mapped loci shown in red (co-located) or blue (single). The fragmentation of the reference genome is indicated in the bottom tracks (reference scaffolds and scaffold breaks), and the distribution of annotated genes is indicated in the gene tracks. We see that the single and paired loci

Table 3. Quantifying co-located loci. PyRAD loci were mapped to the reference genome (*Q. lobata* v.0.5).

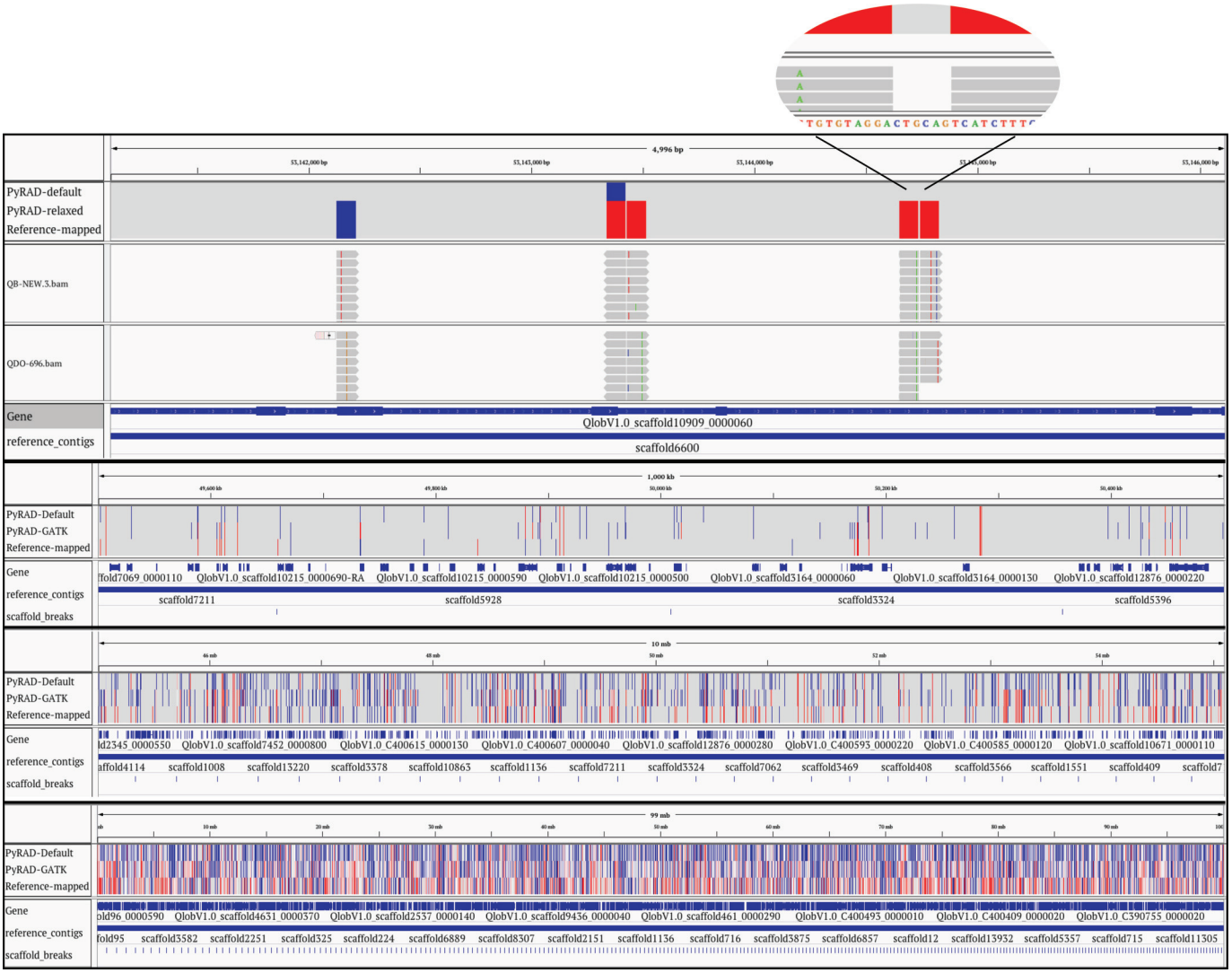
	Total loci after merging	Either inter-loci distance = 6 bp	Percent co-located
PyRAD-default	54 650	20 632	38%
PyRAD-relaxed	90 654	54 842	65%
Reference-mapped	75 123	55 632	74%

Note: Loci within each PyRAD set that mapped to overlapping genome positions were merged into a locus. A locus was counted as co-located if either of it's neighbors was six base pairs away. Numbers were similar if neighbors were allowed to be any distance less than 10 base pairs away.

are fairly evenly distributed along the genome, and that loci are found in and around most of the genes along the genome. Inspection of the sequences between co-located loci confirms the presence of the palindromic cut site sequence (CTGCAG).

One concern with de novo methods is a higher probability of the two haplotypes for single genome regions being treated as separate loci. With such "over-splitting", variation is partitioned into two regions and fails to inform relationships among the individuals partitioned into separate region. While this outcome may manifest as reduced among-clade support, if over-splitting tends to segregate individuals with a probability that is governed by phylogenetic relationships, it may also simply become noise if splitting is not phylogenetically structured. When we map the PyRAD loci to the reference genome, we would see this over-splitting as multiple PyRAD loci mapping to the same genome

Fig. 7. Alignment of loci to the reference genome, *Quercus lobata* v0.5. Each of the four main panels shows a different scale and each has three or more tracks. For the bottom three panels, the lowest track shows the position of the reference genome scaffolds, the middle track shows the gene models, and the top shows the position of each RAD-seq locus retained in each of the three pipelines. The gene models were transferred from annotation of the *Q. lobata* v1.0 genome and have scaffold numbers referring to that assembly. Co-located loci (around a cut site) are shown in red, other loci are shown in blue. The top panel, which includes a blowout, additionally shows read alignments for two of the samples. Bases that do not match the reference sequence are represented as colored bars. The blowout shows the bases of the reference sequence with the co-located reads aligned on either side of the palindromic cut site (CTGCAG). Figure was created using Integrative Genomics Viewer (Thorvaldsdóttir et al. 2013).



position. This phenomenon is especially likely to occur in highly heterozygous taxa, such as oaks, but we found only 5% of PyRAD loci from either set were mapped to non-unique positions on the genome. Further, this 5% represents an upper-bound of the problem as many of these multiply mapped regions are likely caused by over-collapsing of repeats in the reference genome rather than haplotype separation in PyRAD. Thus, over-splitting of loci appears in our dataset to be, at worst, a minor problem of PyRAD de novo clustering, even in highly outcrossed tree species with heterozygosity estimated at 1.25% (Sork et al. 2016*b*). It is worth noting that our data were all clustered at a similarity threshold of 85%, and then filtered to reduce paralogy. Clustering at a higher similarity threshold would undoubtedly increase over-splitting.

Conclusions

Building phylogenetic trees using mapping of RADseq data to a reference genome has several advantages, including avoidance of treating the same sequences separately and identification of genes. However, mapping to a reference appears neither to be needed nor particularly advantageous over de novo clustering for the basic problem of phylogenetic inference. The risk of ascertainment bias in calling variants based on a single genome appears not to have manifested itself in our analysis. Thus the main disadvantage of reference-mapped variant discovery would be the unavailability of a reference genome.

De novo assemblies are confirmed to provide a valuable tool for developing phylogenetic trees. They are advantageous to systematists who wish to develop a robust phylogeny but do not have access to an appropriate reference genome. Biologists using loci generated from RADseq datasets should be aware that 64% or more of the loci may be linked together and not represent independent samples. However, we do not see this linkage of contigs to be problematic for phylogenetic analyses.

Acknowledgements

We thank all the individuals that contributed to the datasets used in this paper, especially K. Beckley, P. Gugger, J. Ortego, and

X. Wei. The research was partially supported by National Science Foundation grants DEB-1146488 (A.L.H.), DEB-1146102 (P.S.M.), and IOS-1444611 (V.L.S.).

References

Andrews, S. 2010. FastQC: a quality control tool for high throughput sequence data. Available from http://www.bioinformatics.babraham.ac.uk/projects/fastqc [accessed 8 November 2016].

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE, **3**(10): e3376. doi:10.1371/journal.pone.0003376. PMID:18852878.

Baxter, S.W., Davey, J.W., Johnston, J.S., Shelton, A.M., Heckel, D.G., Jiggins, C.D., and Blaxter, M.L. 2011. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. PloS ONE, **6**(4): e19315. doi:10.1371/journal.pone.0019315. PMID:21541297.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., and Cresko, W.A. 2013. Stacks: an analysis tool set for population genomics. Mol. Ecol. **22**(11): 3124–3140. doi:10.1111/mec.12354. PMID:23701397.

Cavender-Bares, J., Gonzalez-Rodriguez, A., Eaton, D.A.R., Hipp, A.A.L., Beulke, A., and Manos, P.S. 2015. Phylogeny and biogeography of the American live oaks (*Quercus* subsection *Virentes*): a genomic and population genetics approach. Mol. Ecol. **24**(14): 3668–3687. doi:10.1111/mec.13269. PMID:26095958.

Chong, Z., Ruan, J., and Wu, C.-I. 2012. Rainbow: an integrated tool for efficient clustering and assembling RAD-seq reads. Bioinformatics, **28**(21): 2732–2737. doi:10.1093/bioinformatics/bts482. PMID:22942077.

Corl, A., and Ellegren, H. 2013. Sampling strategies for species trees: the effects on phylogenetic inference of the number of genes, number of individuals, and whether loci are mitochondrial, sex-linked, or autosomal. Mol. Phylogenet. Evol. **67**(2): 358–366. doi:10.1016/j.ympev.2013.02.002. PMID:23410742.

Cruaud, A., Gautier, M., Rossi, J.-P., Rasplus, J.-Y., and Gouzy, J. 2016. RADIS: analysis of RAD-seq data for interspecific phylogeny. Bioinformatics, **32**(19): 3027–3028. doi:10.1093/bioinformatics/btw352. PMID:27312412.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., et al. 2011. The Variant Call Format and VCFtools. Bioinformatics, **27**(15): 2156–2158. doi:10.1093/bioinformatics/btr330. PMID:21653522.

Eaton, D.A.R. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics, **30**: 1844–1849. doi:10.1093/bioinformatics/btu121. PMID:24603985.

Eaton, D.A.R., and Ree, R.H. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae). Syst. Biol. **62**(5): 689–706. doi:10.1093/sysbio/syt032. PMID:23652346.

Edwards, S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution, **63**(1): 1–19. doi:10.1111/j.1558-5646.2008.00549.x. PMID:19146594.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE, **6**(5): e19379. doi:10.1371/journal.pone.0019379. PMID:21573248.

Harvey, M.G., Smith, B.T., Glenn, T.C., Faircloth, B.C., and Brumfield, R.T. 2016. Sequence capture versus restriction site associated DNA sequencing for shallow systematics. Syst. Biol. **65**: 910–924. doi:10.1093/sysbio/syw036. PMID:27288477.

Hipp, A.L., Eaton, D.A., Cavendar-Bares, J., Nipper, R., and Manos, P.S. 2013. Using phylogenomics to infer evolutionary history of oaks. Int. Oak J. **24**: 61–71.

Hipp, A.L., Eaton, D.A.R., Cavender-Bares, J., Fitzek, E., Nipper, R., and Manos, P.S. 2014. A framework phylogeny of the American oak clade based on sequenced RAD data. PLoS ONE, **9**(4): e93975. doi:10.1371/journal.pone.0093975. PMID:24705617.

Hou, Y., Nowak, M.D., Mirré, V., Bjorå, C.S., Brochmann, C., and Popp, M. 2015. Thousands of RAD-seq loci fully resolve the phylogeny of the highly disjunct arctic–alpine genus *Diapensia* (Diapensiaceae). PloS ONE, **10**(10): e0140175. doi:10.1371/journal.pone.0140175. PMID:26448557.

Hulsen, T., de Vlieg, J., and Alkema, W. 2008. BioVenn — a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. BMC Genomics, **9**(1): 488. doi:10.1186/1471-2164-9-488. PMID:18925949.

Hyma, K.E., and Fay, J.C. 2013. Mixing of vineyard and oak-tree ecotypes of *Saccharomyces cerevisiae* in North American vineyards. Mol. Ecol. **22**(11): 2917–2930. doi:10.1111/mec.12155. PMID:23286354.

Kim, B., Wei, X., Fitz-Gibbon, S., Lohmueller, K., Ortego, J., Gugger, P.F., and Sork, V.L. Phylogeny and ancient introgression of the Californian scrub white oak species complex (*Quercus* sect. *Quercus*: Fagaceae) inferred from RADseq data. [In review.]

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10**(3): R25. doi:10.1186/gb-2009-10-3-r25. PMID:19261174.

Lemmon, A.R., Emme, S.A., and Lemmon, E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. **61**: 727–744. doi:10.1093/sysbio/sys049. PMID:22605266.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

Lunter, G., and Goodson, M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. **21**(6): 936–939. doi:10.1101/gr.111120.110. PMID:20980556.

Lynch, M. 2008. Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. Mol. Biol. Evol. **25**(11): 2409–2419. doi:10.1093/molbev/msn185. PMID:18725384.

McCluskey, B.M., and Postlethwait, J.H. 2015. Phylogeny of zebrafish, a "model species," within *Danio*, a "model genus". Mol. Biol. Evol. **32**(3): 635–652. doi:10.1093/molbev/msu325. PMID:25415969.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. **20**: 1297–1303. doi:10.1101/gr.107524.110. PMID:20644199.

McVay, J.D., Hipp, A.L., and Manos, P.S. 2017. A genetic legacy of introgression confounds phylogeny and biogeography in oaks. Proc. R. Soc. B, **284**(1854): 20170300. doi:10.1098/rspb.2017.0300. PMID:28515204.

Nadeau, N.J., Martin, S.H., Kozak, K.M., Salazar, C., Dasmahapatra, K.K., Davey, J.W., et al. 2013. Genome-wide patterns of divergence and gene flow across a butterfly radiation. Mol. Ecol. **22**(3): 814–826. doi:10.1111/j.1365-294X.2012.05730.x. PMID:22924870.

Nixon, K. 2002. The oak (*Quercus*) biodiversity of California and adjacent regions. *In* Proceedings of the Fifth Symposium on Oak Woodlands: Oaks in California's Changing Landscape, 22–25 October 2001, San Diego, Calif. *Edited by* R.B. Standiford, D. McCreary, and K.L. Purcell. USDA Forest Service Gen. Tech. Rep, San Diego, Calif. pp. 3–20.

Ortego, J., Noguerales, V., Gugger, P., and Sork, V. 2015. Evolutionary and demographic history of the Californian scrub white oak species complex: an integrative approach. Mol. Ecol. **24**: 6188–6208. doi:10.1111/mec.13457. PMID:26547661.

Paradis, E., Claude, J., and Strimmer, K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics, **20**(2): 289–290. doi:10.1093/bioinformatics/btg412. PMID:14734327.

Pavlik, B.M., Muick, P.C., Johnson, S.G., and Popp, M. 1995. Oaks of California. Cachuma Press, Oakland, Calif.

Pearse, I.S., and Hipp, A.L. 2009. Phylogenetic and trait similarity to a native species predict herbivory on non-native oaks. Proc. Natl. Acad. Sci. **106**(43): 18097–18102. doi:10.1073/pnas.0904867106. PMID:19841257.

Quinlan, A.R. 2014. BEDTools: the Swiss-army tool for genome feature analysis. Current Protocols in Bioinformatics, **47**: 11.12.11–11.12.34. doi:10.1002/0471250953.bi1112s47. PMID:25199790.

R Core Team. 2016. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. Available from https://www.R-project.org.

Ree, R.H., and Hipp, A.L. 2015. Inferring phylogenetic history from restriction site associated DNA (RADseq). *In* Next-generation sequencing in plant systematics. *Edited by* E. Hörandl and M.S. Appelhans. pp. 181–204.

Reitzel, A., Herrera, S., Layden, M., Martindale, M., and Shank, T. 2013. Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. Mol. Ecol. **22**(11): 2953–2970. doi:10.1111/mec.12228. PMID:23473066.

Revell, L.J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. **3**(2): 217–223. doi:10.1111/j.2041-210X.2011.00169.x.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. 2011. Integrative genomics viewer. Nat. Biotechnol. **29**(1): 24–26. doi:10.1038/nbt.1754. PMID:21221095.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ, **4**: e2584. doi:10.7717/peerj.2584. PMID:27781170.

Rubin, B.E., Ree, R.H., and Moreau, C.S. 2012. Inferring phylogenies from RAD sequence data. PLoS ONE, **7**(4): e33394. doi:10.1371/journal.pone.0033394. PMID:22493668.

Sork, V.L., Riordan, E., Gugger, P.F., Fitz-Gibbon, S., Wei, X., and Ortego, J. 2016*a*. Phylogeny and introgression of California scrub white oaks (*Quercus* sect. *Quercus*). Int. Oak J. **27**: 61–74.

Sork, V.L., Fitz-Gibbon, S.T., Puiu, D., Crepeau, M., Gugger, P.F., Sherman, R., et al. 2016*b* First draft assembly and annotation of the genome of a California endemic oak *Quercus lobata* Née (Fagaceae). G3: Genes Genomes Genet. **6**(11): 3485–3495. doi:10.1534/g3.116.030411.

Sovic, M.G., Fries, A.C., and Gibbs, H.L. 2015. AftrRAD: a pipeline for accurate and efficient de novo assembly of RADseq data. Mol. Ecol. Resour. **15**(5): 1163–1171. doi:10.1111/1755-0998.12378. PMID:25641221.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, **30**(9): 1312–1313. doi:10.1093/bioinformatics/btu033. PMID:24451623.

Swenson, K.M., and El-Mabrouk, N. 2012. Gene trees and species trees: irreconcilable differences. BMC Bioinform. **13**(S19): S15. doi:10.1186/1471-2105-13-S19-S15.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings Bioinform. **14**(2): 178–192. doi:10.1093/bib/bbs017.

Wagner, C.E., Keller, I., Wittwer, S., Selz, O.M., Mwaiko, S., Greuter, L., et al. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. **22**(3): 787–798. doi:10.1111/mec.12023. PMID:23057853.

Weitemier, K., Straub, S.C., Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A., and Liston, A. 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. Appl. Plant Sci. **2**(9): 1400042. doi:10.3732/apps.1400042.