View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by White Rose Research Online


This is a repository copy of *Phylogenomics using low-depth whole genome sequencing: a case study with the olive tribe*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/145071/

Version: Accepted Version

---

**Article:**
Olofsson, J.K., Cantera, I., Van de Paer, C. et al. (6 more authors) (2019) Phylogenomics using low-depth whole genome sequencing: a case study with the olive tribe. Molecular Ecology Resources. ISSN 1755-098X

https://doi.org/10.1111/1755-0998.13016

---

This is the peer reviewed version of the following article: Olofsson, J. K., Cantera, I. , Van de Paer, C. , Hong-Wa, C. , Zedane, L. , Dunning, L. T., Alberti, A. , Christin, P. and Besnard, G. (2019), Phylogenomics using low-depth whole genome sequencing: a case study with the olive tribe. Mol Ecol Resour, which has been published in final form at https://doi.org/10.1111/1755-0998.13016. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

**Reuse**
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

**Takedown**
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

# Phylogenomics using low-depth whole genome sequencing: a case study with the olive tribe

**Running title:** Phylogenomics of Oleeae using low-depth sequencing

Jill K. Olofsson[1*], Isabel Cantera [2], Céline Van de Paer[2], Cynthia Hong-Wa[3], Loubab Zedane[2#], Luke T. Dunning[1], Adriana Alberti[4], Pascal-Antoine Christin[1], Guillaume Besnard[2*]

[1] Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom

[2] Laboratoire Évolution & Diversité Biologique (EDB, UMR5174), Université de Toulouse, CNRS, UPS, IRD, 118 route de Narbonne, F-31062 Toulouse, France

[3] Delaware State University, Claude E. Phillips Herbarium, 1200 N Dupont Hwy, Dover, DE 19901, USA

[4] CEA - Institut de biologie François-Jacob, Genoscope, 2 Rue Gaston Cremieux, 91057 Evry Cedex, France

[#] Present address: Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, United Kingdom

[*]Authors for correspondence: j.k.olofsson@sheffield.ac.uk, guillaume.besnard@univ-tlse3.fr

## Abstract

Species trees have traditionally been inferred from a few selected markers, and genome-wide investigations remain largely restricted to model organisms or small groups of species for which sampling of fresh material is available, leaving out most of the existing and historic species diversity. The genomes of an increasing number of species, including specimens extracted from natural history collections, are being sequenced at low depth. While these datasets are widely used to analyse organelle genomes, the nuclear fraction is generally ignored. Here we evaluate different reference-based methods to infer phylogenies of large taxonomic groups from such datasets. Using the example of the Oleeae tribe, a worldwide-distributed group, we build phylogenies based on single-nucleotide polymorphisms (SNPs) obtained using two reference genomes (the olive and ash trees). The inferred phylogenies are overall congruent, yet present differences that might reflect the effect of the distance to the reference on the amount of missing data. To limit this issue, the genome complexity was reduced by using pairs of orthologous coding sequences as the reference, thus allowing combining SNPs obtained using two distinct references. Concatenated and coalescence trees based on these combined SNPs suggest events of incomplete lineage sorting and/or hybridization during the diversification of this large phylogenetic group. Our results show that genome-wide phylogenetic trees can be inferred from low-depth sequence datasets for eukaryote groups with complex genomes, and histories of reticulate evolution. This opens new avenues for large-scale phylogenomics and biogeographic analyses covering both the extant and historic diversity stored in museum collections.


**Keywords:** Coalescence, genome skimming, low-depth sequencing, natural history collections, phylogenomics, Oleaceae.

# 1 INTRODUCTION

One of the fundamental goals of evolutionary biology is to understand how organisms have changed over time and adapted to the environments in which they prosper, an endeavour which requires a robust phylogenetic framework (e.g., Revell, Johnson, Schulte, Kolbe, & Losos, 2007; Edwards et al., 2010; Jetz, Thomas, Joy, Hartmass, & Mooers, 2012). While the availability of sequence data has historically been limiting, high-throughput sequencing provides large datasets to infer phylogenetic trees from genetic markers spread across whole genomes (Delsuc, Brinkmann, & Philippe, 2005; Paterson, Freeling, Tang, & Wang, 2009), increasing the statistical confidence in solving phylogenetic problems (Dunn et al., 2008; Puttick et al., 2018). High-throughput sequencing also allows for simultaneous phylogenetic comparisons of different genomic regions, which can further help elucidate processes such as incomplete lineage sorting, hybridization, or horizontal gene transfer (Green et al., 2010; Christin et al., 2012a; Christin et al., 2012b; Scally et al., 2012; Marcussen et al., 2014). However, genome-wide phylogenetic analyses still remain mostly restricted to organisms for which high-quality sequencing datasets are available, currently excluding most of the extant, and almost all of the past, species diversity.

The limitations in producing genome-wide datasets for large numbers of species are twofold. Firstly, high-depth sequencing remains prohibitively expensive. Secondly, obtaining fresh and high-quality material for a large number of taxa is challenging, especially if these are rare and/or occur in remote areas. Most species are, however, available as conserved specimens in natural history museums, often as multiple samples from large geographic areas (e.g., Buerki & Baker, 2016; Bieker & Martin, 2018). The DNA stored in such samples is often degraded and/or of limited quantity but can still be accessed via low-depth whole genome shotgun sequencing of small DNA fragments (also known as genome skimming or genome scanning) or targeted capture approaches (Straub et al., 2012; McCormack, Tsai, & Faircloth, 2016). While the latter provides high coverage of markers selected *a priori*, low-depth whole genome sequencing (usually < 1-2× depth) is typically used to assemble genomic regions with high copy numbers, such as the organelle genomes

and the nuclear ribosomal DNA (nrDNA) cluster (e.g., Straub et al., 2012; Malé et al., 2014; Dodsworth, 2015; Van de Paer, Hong-Wa, Jeziorski, & Besnard, 2016). However, the reads corresponding to nuclear low-copy number loci included in these datasets are generally overlooked, although individual markers can be assembled using a reference-based approach (Besnard et al., 2014, 2018). In addition, genetic variation spread across the genome can be extracted from these datasets and used for phylogenetic inference or population genetics (e.g., Li, Sidore, Kang, Boehnke, & Abecasis, 2011; Buerkle & Gompert, 2013; Olofsson et al., 2016; Dunning et al., 2017). However, the usefulness of low-depth sequence datasets to infer phylogenies that cover a large number of taxa with deep divergence times remains unclear, especially for capturing the diversity stored in herbaria and museums. Indeed, previous studies have highlighted the effect of the distance to the reference on the assembled dataset, with different amounts of missing data potentially influencing the inferred phylogenetic trees (Bertels, Silander, Pachkov, Rainey, & van Nimwegen, 2015). In addition, previous methods typically focussed on sequencing coverage higher than that associated with genome-skimming sequence data (Allen et al., 2017; Zhang et al., 2019). Here, we evaluate the possibility of using low-depth sequencing (e.g. as low as $< 0.5\times$ depth) for phylogenetic inferences across large divergence times using the olive tribe (Oleeae) as an example.

In this study, we combine low-to-medium depth whole genome sequencing data from 72 fresh and 28 herbarium accessions of Oleeae. We use these data to build phylogenies based on two classical genetic markers used for investigating phyloeography, the plastome and nrDNA cluster. We then compare these phylogenies to those based on nuclear SNPs obtained using two different reference genomes from the Oleeae tribe (olive and ash trees). We use these comparisons to evaluate the effect of different (i) reference genomes including their complexity and (ii) filtering strategies to retain SNP positions on the inferred phylogenetic relationships. We further introduce a new approach that consists in reducing the complexity of the reference by only considering pairs of orthologous coding sequences (CDS) to allow combining sets of SNPs obtained with different references. We evaluate the usefulness of this method in (iii) removing the effect of the distance to

the reference genome, and (iv) allowing multigene coalescence-based phylogenetic analyses. The methods we present work for nuclear markers retrieved from low-depth whole genome data that can be obtained at relatively low-cost from fresh samples and specimens stored in natural history collections, making them highly attractive for inferring phylogenies over large divergence times and to trace functional trait evolution.

# 2 MATERIALS AND METHODS

## 2.1 The Oleeae tribe as a study system

The Oleeae tribe (Oleaceae, Lamiales) comprises 19 genera (one recently extinct) and encompasses approximately 300 species, predominately trees and shrubs, distributed worldwide (Wallander & Albert, 2000; Green, 2004). Oleeae has an allopolyploid origin ($n$ = 23 chromosomes; Taylor, 1945), although few subsequent events of polyploidization have been observed [usually in narrow endemics; for instance in olive (Besnard et al., 2008), *Fraxinus* (Taylor, 1945), and *Noronhia* (G. Besnard, unpub. data)]. The genome size of the group is typical of flowering plants (1C: mean 1.5 pg [range: 0.87-2.99 pg]; Bennett & Leitch, 2012). Four subtribes are currently recognized: Schreberinae, Ligustrinae, Fraxininae, and Oleinae (Wallander & Albert, 2000). The phylogenetic relationships within each subtribe are complex, with some of the recognized genera being polyphyletic (Besnard, Rubio de Casas, Christin, & Vargas, 2009; Yuan, Zhang, Han, Dong, & Shang, 2010; Guo et al., 2011; Hong-Wa & Besnard, 2013; Zedane et al., 2016) or paraphyletic (Li, Alexander, & Zhang, 2002). Furthermore, extensive incongruence between the plastid DNA (cpDNA) and nrDNA phylogeny has been reported, suggesting hybridization and/or incomplete lineage sorting within several genera (e.g., Besnard et al., 2009; Hinsinger et al., 2013; Hong-Wa & Besnard, 2013), although heterogeneous evolutionary rates might also account for biased inferences based on nrDNA (Zedane et al., 2016). Nuclear genomes were recently sequenced for two Oleeae species, *Olea europaea* subsp. *europaea* (Olive; Cruz et al., 2016) and *Fraxinus excelsior* (Ash; Sollars et al., 2017), offering the opportunity to develop phylogenomics in this plant group.

125

### 2.2 Sampling, DNA extraction, and sequencing

Low to medium depth whole genome sequences for a total of 100 accessions belonging to 86 species and representing the four subtribes of Oleeae and one outgroup species (*Forsythia mandschurica,* tribe Forsythieae) were obtained (Table S1). For 28 herbarium samples, a mature
130 leaf fragment was sampled from green-looking, presumably non-alcohol-treated, specimens collected between 1872 and 2013 (Table S1). This allowed us to sample lineages from remote areas in tropical Asia and Australasia (n = 13), and neotropical America (n = 9), some of which have not previously been included in phylogenetic studies. The other 72 samples were obtained from fresh leaf material dried in silica gel. Multiple accessions were included for nine species and served as
135 species controls, with the expectation that they would group together in all phylogenetic analyses. For 84 accessions, DNA was extracted from leaf fragments (ca. 5-10 mg) using the DNeasy Plant Mini kit (Qiagen, Valencia, CA, USA), quality checked, and sequenced at the Genopole platform of Toulouse or at the Genoscope platform of Evry as previously described (Besnard et al., 2014; Roquet et al., 2016). For the 72 samples sequenced at the Genopole platform of Toulouse, between
140 10 and 500 ng of double stranded DNA were used to construct sequencing libraries with the Illumina TruSeq Nano DNA LT Sample Prep kit (Illumina, San Diego, CA, USA), following the manufacturer's instructions. DNA was fragmented by sonication, except for extracts from herbarium specimens, which were already degraded. Each sample was paired-end sequenced (100, 125, or 150 bp) on 1/24$^{th}$ of an Illumina HiSeq2000, HiSeq2500 or HiSeq3000 lane (Table S1) and
145 multiplexed with samples from the same or different projects. For the 12 samples sequenced at the Genoscope platform of Evry, 250 ng of genomic DNA were sonicated using the E210 Covaris instrument (Covaris, Inc., USA). Fragments were end repaired and 3`-adenylated. NextFlex DNA barcodes (Bioo Scientific Corporation, Austin, TX, USA) were then added using the NEBNext DNA Modules Products (New England Biolabs, MA, USA) followed by clean up with 1x AMPure
150 XP. The ligated product was amplified with 12 cycles PCR using Kapa Hifi Hotstart NGS library Amplification kit (Kapa Biosystems, Wilmington, MA) followed by a 0.6x AMPure XP

purification. Each sample was paired-end sequenced (101 bp) on 1/48$^{th}$ of an Illumina HiSeq2000 lane (Illumina, USA; Table S1) and multiplexed with samples from the PhyloAlps project. Data for another 16 accessions generated with a similar protocol were retrieved from previous studies (Table S1; Van de Paer et al., 2016; Van de Paer, Bouchez, & Besnard, 2018; Zedane et al., 2016).

### 2.3 Assembly and phylogenetic analyses of whole plastomes and nrDNA clusters

For each accession, paired-end reads were used to assemble the whole plastome and nrDNA cluster independently, following the method of Zedane et al. (2016). Alignments including both monomorphic and polymorphic sites were produced independently for the plastome and the nrDNA cluster using MUSCLE as implemented in MEGA v. 7 (Kumar, Stecher, & Tamura, 2016) followed by manual refinement (particularly in regions containing inversions). Phylogenetic trees were inferred independently for the plastome (after removing one inverted repeat) and nrDNA using the maximum-likelihood method implemented in RAxML v. 8 (Stamatakis, 2014) with the best substitution model (GTR + G + I), as determined with Smart Model Selection (SMS) v. 1.8.1 (Lefort, Longueville, & Gascuel, 2017) in PhyML v. 20120412 (Guindon et al., 2010). Node support was evaluated with 100 rapid bootstrap iterations. A time-calibrated phylogeny was obtained from the plastome dataset using Bayesian inference as implemented in BEAST v. 2.4.3 (Bouckaert et al., 2014; For details see Supplemental Material 1).

### 2.4 Nuclear SNP calling

#### *2.4.1 Overview of method*

Due to the low-coverage of our data, we adopted a reference-based approach to call SNPs. Existing genotype-calling algorithms that control for the quality of the mapping to the reference and of the genotype tend to favour the reference allele for certain genomic regions even when SNPs are present (Bertels et al., 2015). Preliminary tests conducted here indicated that the prevalence of this problem increased with low-depth sequencing, but it is possible to bioinformatically reconstruct

genotypes directly from mapped reads. This approach however, often ignores quality scores, potentially increasing the amount of sequencing errors incorporated into final SNP alignments. Here, we combine the two methods by first defining a set of high quality SNP positions using a genotyping algorithm and then bioinformatically reconstructing genotypes from uniquely mapped reads using low to medium depth sequencing data (Olofsson et al., 2016; Dunning et al., 2017; Figure 1). In short, quality filtered reads were mapped onto a reference genome and high-quality SNP positions were extracted from uniquely mapped reads taking differences in sequencing depth between samples into account (Figure 1). Genotypes were then reconstructed for the high-quality positions from the mapped reads using an in-house developed bioinformatic pipeline (Figure 1; For more details see Supplemental Material 2). The results of our genotyping method were compared to those of the likelihood method implemented in the program ANGSD (Korneliussen, Albrechtsen, & Nielsen, 2014).

### *2.4.2 Reference genomes*

Two different whole genome reference assemblies were used, the olive tree (Oe6, cultivar ‘Farga’; Cruz et al., 2016) and the ash tree (BATG-0.5; Sollars et al., 2017; Figure 1), both of which belong to the Oleeae tribe. Neither of these genomes is assembled to the chromosome level, and they therefore consist of a collection of contigs and scaffolds. The olive genome includes recently diverged copies of many genes, which has been interpreted as the remnants of a recent polyploidization event followed by a rapid rediploidization (Julca, Marcet-Houben, Vargas, & Gabaldón, 2018). However, these duplicated regions could also be an assembly artefact where divergent alleles have not been merged due to the high heterozygosity of the olive cultivar 'Farga' (Besnard et al., 2013; Cruz et al., 2016; Diez et al., 2015).

Differences in genome complexity can alter the mapping of short reads and the subsequent SNP calling between different reference genomes. We therefore constructed pairs of orthologous protein-coding sequences from the two genomes. A total of 13,907 orthologous CDS

pairs were identified from their reciprocal best hits using BLAST v. 2.2.28 (e-value < 1e-10; Altschul, Gish, Miller, Myers, & Lipman, 1990) (Figure 1). Each pair was aligned in MAFFT v. 7.123b (Katoh & Standley, 2013), and the alignments were trimmed in GBLOCKS v. 0.91b (default parameters; Castresana, 2000) removing overhang bases and insertion/deletion. The trimmed orthologous CDS of each of the two species were then used separately as references to call SNPs (Figure 1). The called SNPs were finally merged, producing a single dataset (see below; Figure 1).

### *2.4.3 Details of nuclear genotyping*

Raw reads were first cleaned and trimmed individually using the NGS QC toolkit v. 2.3.3 (Patel & Jain, 2012). Reads with ambiguous base calls and where more than 20% of the bases had a quality score below 20 were removed. Low quality bases (Q < 20) were further trimmed from the 3' end of each read. The cleaned and trimmed pair-end reads were then mapped onto each of the four references (see above) using the default settings for pair-end reads in BOWTIE2 v. 1.1.1 (Figure 1; Langmead & Salzberg, 2012) and uniquely mapped reads in proper pairs were identified using SAMtools v. 1.3.1 (Li et al., 2009) and Picard tools v. 1.92 (http://picard.sourceforge.net/; Figure 1). The relationship between mapping success and divergence time to the two reference genomes was evaluated using dates retrieved from the time-calibrated plastome phylogeny (see below). Furthermore, an estimated nuclear sequencing depth was calculated given the number of cleaned reads, the length of the reads, and the size of the reference genome (Table S1).

The genomic position of each high quality nuclear SNP was determined using the mpileup function in SAMtools and the consensus variant caller algorithm in BCFtools v. 1.3.1 (Li, 2011b). Because ploidy levels were not known, all samples were treated as if they were diploids, only including SNPs with a maximum number of two alleles within a sample. This could cause some locus drop-outs in allopolyploids but should not significantly affect the SNP calling efficiency of autopolyploids. Mapping polyploids to a diploid genome might also increase the frequency of allele

drop-outs in polyploids due to unequal mapping success between divergent alleles. However, allelic drop-outs due to the low sequencing depth (Olofsson et al., 2016) are likely to be more frequent than locus-drop outs due to polyplodization. The effect on SNP calling when treating all samples as diploids will therefore likely be minimal in our low depth sequencing data set. The median coverage of all SNPs called in at least 50% of the individuals was computed for each sample using a custom Perl script (Supplemental Material 2). This means that coverage is estimated from positions present in the majority of species which ensures that the effect of sequence errors is kept to a minimum. For each sample, the raw genotyped SNPs were then filtered so that only sites with coverage between 0.5 and two times the median coverage and a minimum quality score of 20 were retained (Supplemental Material 2). This approach takes into account variation in sequencing depth among samples and the upper threshold excludes reads originating from the organelle genomes or repetitive regions of the nuclear genome, both of which reach high coverage. The individual genotypes were merged in BCFtools and filtered in VCFtools v. 0.1.14 (Danecek et al., 2011), keeping positions with a minor allele count of at least three, thereby removing private SNPs and sequencing errors which are indistinguishable from each other in low-coverage data. This filter retains variants shared by at least two individuals and ensure that phylogenetically informative sites for lineages with few sampled taxa are retained. To assess the impact of missing data on our downstream analysis we also applied five different missing data cut-offs for the proportion of missing data (10, 20, 50, 80, and 90%). Genotypes of the identified high-quality SNP positions were then directly reconstructed from the uniquely mapped reads using a previously published shell pipeline (Olofsson et al., 2016) with small modifications (Dunning et al., 2017; Supplemental Material 2). The two sets of SNPs called using the CDS as reference were combined into a single SNP alignment where orthologous positions were only represented once. The two SNP sets were imported into Geneious v. 8.1.7 and the consensus genotype at each position was determined without consideration for the individual base frequencies. If more than two bases were detected at a position, the SNP was considered ambiguous.

For comparison, nuclear SNPs were also called from uniquely mapped reads using ANGSD v. 0.921 (Korneliussen et al., 2014) using the recommend settings, treating all samples as diploids, and calling genotypes with a probability above 0.34 (i.e. not allowing sites with equal genotype probabilities). SNPs were called from uniquely mapped reads in proper pairs with per-Base Alignment Quality (BAQ; Li, 2011a) activated. Genotype likelihoods were estimated using the SAMtolls model (-GL 1) and the reference allele was used as the major allele (-doMajorMinor 4). Only SNPs with a p-value (-SNP_pval) below 1e-6 were kept. The genotypes were then filtered to make them comparable to the best filter-set for our SNP calling method (i.e. < 80% missing data and minor allele count > 3; see Results).

*2.4.4 Phylogenetic inference based on nuclear markers*

Phylogenetic trees were inferred using RAxML v. 8 under the GTR + G substitution model, as described above, for each of the three concatenated datasets of nuclear SNPs (mapped to the ash genome, the olive genome, and the combined CDS). The significance of topological differences between the three nuclear SNP datasets and those based on plastomes and nrDNA were evaluated using Shimodaira-Hasegawa tests corrected for multiple testing (SH tests; Shimodaira & Hasegawa, 1999) as implemented in baseml in the PAML v. 4.7 package (Yang, 2007).

A multigene coalescence phylogenetic tree was further inferred from the CDS SNPs treating each CDS as an individual gene. Only CDS with an alignment of at least 50 nucleotides was included, and a bootstrapped (100) maximum likelihood phylogeny was inferred in PhyML v. 20120412 using a GTR + G + I substitution model. Nodes with bootstrap support below ten were collapsed using NEWICK utilities v. 1.6 (Junier & Zdobnov, 2010). The gene trees were then used to reconstruct a coalescence species tree using ASTRAL v. 5.6.2 (Mirarab et al., 2014), reporting the posterior support value for the main topology as well as the percentage of gene trees supporting each of the three alternative quartets for each node.

# 3 RESULTS

## 3.1 Sequencing and analyses of plastomes and nrDNA clusters

The quality and quantity of input DNA differed among samples resulting in a large variation in the number of paired-end raw reads (range: 2.67-58.4 M). Specimens from herbarium collections generally had shorter insert sizes than those from fresh material (mean 226 [range: 109-364] bp vs. 310 [range: 146-441] bp; Table S1), although their DNA was not sonicated prior to library preparation. Complete plastomes (154 to 165 kb in length) and nearly complete nuclear ribosomal clusters (5 to 10 kb in length), two markers classically used to contrast plastid and nuclear evolutionary history (e.g., Alvarez & Wendel, 2003; Nieto Feliner & Rosselló, 2007; Christin et al., 2012b; Lundgren et al., 2015), were assembled for all newly sequenced accessions. These markers are present in high copy numbers (Straub et al., 2012) and therefore the coverage of these regions is relatively high (64× to 6,514× for cpDNA, and 58× to 9,652× for nrDNA; Table S1).

The phylogeny inferred from the plastomes is well resolved, congruent with previous works (Wallander & Albert, 2000; Hong-Wa & Besnard, 2013; Zedane et al., 2016), and all accessions belonging to the same species cluster together (Figures 2 and S1). In contrast, the phylogeny inferred from nrDNA is less well resolved and shows statistically supported differences to the plastome tree (Figures 2 and S2; Table S2). In particular, the subtribe Schreberinae is nested within the subtribe Oleinae, while it is sister to all other Oleeae in the plastome phylogeny (Figures 2 and S2). The nrDNA of Schreberinae and the clade in which it is nested are characterized by high GC content in their intergenic spacers (> 65%; Figure S2), which is correlated with branch lengths (pgls: $R^2 = 0.3105$ and $p < 0.001$; Figures S2 and S3).

## 3.2 Genotyping of nuclear SNPs

The percentage of read-pairs retained after cleaning varied among samples (43-98%; Table S1), reflecting variation in DNA and sequencing quality. Similarly, the mapping success also varied among samples and between the two reference genomes (Table S1). The latter is especially

noticeable for species closely related to one of the two reference genomes, which have very high mapping success to one of the two genomes (Table S1). Overall, the mapping success decreases with evolutionary distance to the reference genome, as estimated from plastome divergence times, and there were no differences in the slope between samples extracted from fresh silica dried material and herbarium specimens (Figure 3; Table S1). However, despite this rapid decrease in mapping success, some conserved genomic regions were still retrieved from the most divergent samples in the dataset (Figure 3), indicating that some mapping is still possible for species that diverged up to 45 million years ago. The median coverage of the raw SNPs is generally low (median 2×; 1-20×), and the number of SNPs varies among samples; ~27,600 to 16 million using the olive genome as reference and ~97,300 to 8.4 million using the ash genome as reference (Table S1). Variation in coverage is mainly explained by the estimated sequencing depth, whereas the large variation in genotyped positions mainly reflects differences in divergence times to the reference genomes (linear regressions, $p < 0.001$, $R^2 = 0.23$ and $p < 0.001$, $R^2 = 0.62$, respectively; Figure S4). Estimated sequencing depth is only mildly indicative of the number of typed SNPs (Figure S4B). The trade-off between number of sequenced samples and the sequencing depth should therefore be considered on a per study basis as the divergence time to the reference genome is causing the most variation in typed SNPs (Figure S4).

The amount of missing data per SNP can affect the robustness of phylogenetic inferences, but retaining only positions genotyped in all samples drastically reduces the size of the dataset. No SNPs were retained when a maximum of 10 or 20% of missing data was allowed (Table S3). Higher levels of missing data (50-90%) produced similar topologies, and 80% missing data was chosen (Figure 4) as it gave consistent relative branch lengths across the phylogenies (Figures S5-S12). Concatenated SNP alignments obtained using our pipeline (Figures 4, S6, and S9) and ANGSD produced highly similar topologies although more SNPs were obtained with ANGSD (Figures S13-S15). While genotype likelihoods can be incorporated in population genomic analyses, it is not possible to do so for phylogenetic inferences that require alignment matrices. We

therefore focus on the phylogenies produced using our pipeline.

The final alignments consisted of 319,869 and 222,461 nuclear SNPs, for the mapping to the olive and ash genomes, respectively. The percentage of missing data varied among samples, from 0.2 to 94.4% (median 32.8%) and from 1.2 to 94.3% (median 29%) for the olive and ash references, respectively (Table S1). Level of missing data was influenced by multiple factors including coverage and mapping success, the latter being determined by the divergence time to the reference genome (Figures 3 and S4), and the amount of missing data was consequently phylogenetically clustered around the reference genome (Figures S6 and S9). Most of the positions in the final nuclear SNP alignments are localized in annotated exons (80 and 87%, respectively for the olive and ash datasets) scattered across the genomes.

Mapping success and the number of called SNPs similarly varied among the sets of orthologous CDSs extracted from the ash and genome trees (Tables S1 and S4), and the number of SNPs was similarly inversely correlated to the distance from the reference genome (Figure 5). After merging and filtering the two CDS SNP sets, a total of 233,829 nuclear SNPs were retained and the amount of missing data varied between 0.3 and 93.1% (median 24%; Table S4) among samples, a range similar to that observed when mapping onto the two whole genomes (see above). Importantly, the relationship between the number of SNPs and distance to the references disappeared in this combined dataset (linear regressions, $p = 0.23$ and 0.36 for the divergence to the olive and ash, respectively; Figure 5C). Instead, the amount of missing data in the combined CDS matrix was inversely correlated to the estimated sequencing depth (linear regression, $p < 0.00001$, $R^2 = 0.19$; Figure 5D). All samples with an estimated sequencing depth based on the size of the olive reference genome above 3× had less than 10% missing data, and these were spread across the phylogeny (Figure 4). These relationships are based on the assumption of a conserved genome size in the group, which is unlikely to be true. The real correlation between sequencing depth and missing data is likely stronger, but evaluating it requires obtaining good quality material (i.e. fresh leaves) to measure the genome size of all species.

360

### 3.3 Phylogenetic inference based on nuclear SNPs

The phylogenies inferred from the three different nuclear SNP alignments were overall similar to the one based on plastomes, although a total of 18 branches from the plastome tree differed from the three nuclear topologies (Figures 2, 4, S6, and S9) and SH tests confirmed that the plastome topology fitted the plastome alignment significantly better than any other topology ($p < 0.05$; Table S2). While the differences concerned are mainly branches close to the tips (e.g., relationships among *Olea* sect. *Olea* and among *Fraxinus* sect. *Ornus*), deeper nodes also differed, such as the grouping of *Noronhia*, *Olea*, and *Chionanthus* (Figures 2, 4, S6, and S9).

The topologies inferred from the SNP alignments based on the olive and ash whole genome references were highly similar, with only eight differences, three of which were within the *Fraxinus* genus and two others concerned terminal triplets (Figures S6 and S9). The mapping success of the *Fraxinus* taxa is very different between the two whole genomes, which can partially explain the few observed discrepancies in the phylogenetic placements (Table S1). Furthermore, the identity of the most basal node within *Olea* sect. *Olea*, the position of *Chionanthus virginicus*, and the relationships among the *Olea* sensu stricto, core *Chionanthus*, and *Noronhia* clades also differed (Figures S6 and S9). Both of the whole genome topologies fitted significantly ($p < 0.05$) better the data from which they were inferred than any of the alternative topologies (Table S2).

The topology inferred for the orthologous CDS alignment did not fit this dataset better than those based on the olive or ash whole genome reference SNP alignments ($p > 0.1$), but was significantly better than the plastome and nrDNA topologies (Table S2). For the eight branches differing between the two whole genome topologies, the CDS topology agreed with the ash topology in four cases and with the olive topology in the four others (Figures 4, S6, and S9). All four differences between the ash and CDS topologies are supported by less than 90% of bootstrap replicates in at least one of the two trees, while all four differences between the olive and CDS topologies are supported by more than 90% of bootstrap replicates in both topologies (Figures 4,

S6, and S9). In addition, two branches, the grouping of *Fraxinus insularis* and *Fr. bungeana* and the position of *Chionanthus ligustrinus* as sister to *Forestiera* were identical in the two whole genome topologies, but differed to the CDS topology (Figures 4, S6, and S9). Both these branches are, however, associated with support values below 80% in the CDS tree (Figures 4, S6, and S9).

The multigene coalescence species tree based on CDS SNPs is well resolved, but not all nodes are supported by a majority of the gene trees (Figure 6). Some lineages (e.g., *Fraxinus*, *Syringa* + *Ligustrum*, and the core *Chionanthus*) are supported by most gene trees (Figure 6). However, the relationships that differ among the three concatenated datasets also vary among gene trees (Figure 6). The discrepancies between the multiple references can therefore, at least partially, reflect incomplete sorting and/or hybridization (including allopolyploidization). Many of the nodes that differ between the nuclear and plastid datasets are also supported by different gene trees (Figures 2, 4, 6, S6, and S9). The most notable exception is the relationship among *Picconia* and *Phillyrea*. Both genera are monophyletic in all nuclear SNP phylogenies, a relationship that is supported by the majority of gene trees. However, in the plastome dataset *Phillyrea* is nested within *Picconia* (Figures 2, 4, 6, S6, and S9).

## 4 DISCUSSION

In this study, we combined new and existing whole-genome sequencing datasets of low to medium depth (0.2-15×; Table S1) to evaluate their power to infer phylogenetic relationships among distantly related taxa. The plastome phylogeny, which is one of the most widely used genetic marker for phylogeography reconstruction, significantly differs from all phylogenies inferred from nuclear SNP alignments (Figures 2, 4, S6, and S9). By contrast, phylogenies of SNP alignments obtained using two different reference genomes present relatively few differences (Figures S6 and S9). Reducing the genome complexity down to orthologous CDS sequences decreases the phylogenetic clustering of missing data caused by differences in divergence time to the reference genomes and allows multigene coalescence analyses (Figures 4 and 6). We therefore conclude that

consistent, robust phylogenetic relationships of the nuclear genome can be inferred from low-depth sequencing data for groups of eukaryotes with complex genome histories, including events of reticulate evolution, that span at least 45 Mya.

## 4.1 Low-depth sequencing can infer nuclear genome phylogenies

### *4.1.1 Impact of reference and filtering for nuclear SNP alignments*

Retrieving phylogenetic information from genome-wide low coverage scans is challenging, but identifying variants (SNPs) by mapping reads to a reference genome can provide markers spread across the nuclear genome (Olofsson et al., 2016; Dunning et al., 2017). Here, we obtain concatenated SNP alignments from reads mapped to genome assemblies for each of the ash and olive genomes. By controlling for coverage, we remove highly repeated markers, such as transposable elements, but we allow for SNPs to be called from both coding and non-coding DNA. However, as expected, the vast majority of the SNPs retrieved in our analyses are located in annotated protein-coding genes (>80%), which are sufficiently conserved to be compared across the evolutionary scale considered (Olofsson et al., 2016).

Phylogenies inferred from genome-wide SNPs can be affected by different methodological and analytical problems. For example, it has been shown that phylogenies inferred from alignments excluding invariant positions can be biased under some circumstances (Lewis, 2001; Bertels et al., 2015; Leaché, Banbury, Felsenstein, de Oca, & Stamatakis, 2015). Obtaining SNP alignments using reference-based methods can also introduce phylogenetic biases due to the divergence between the samples and the reference. Because the mapping success depends on the divergence to the reference genome and hence affects the number of filtered SNPs per accession, the amount of missing data across our phylogeny varies among samples and is phylogenetically clustered (Table S1; Figures 3 and S4A). This could cause biases similar to long-branch attraction resulting in slight incongruences between phylogenies obtained using different reference genomes, potentially contributing to the observed differences (Figures 2, 4, S6, and S9; Xi, Liu, & Davis, 2016; Nute, Chou, Molloy, &

Warnow, 2018). In addition, heterogeneous sampling densities across the phylogeny can affect the distribution of the retained SNPs, as variants present in clades with many species are more likely to pass our filters. The variants existing in species-poor lineages are indeed more difficult to distinguish from sequencing errors in datasets of low sequencing depth. To compensate for this our filters retain only those positions that are shared by at least two individuals. While this filter will exclude most sequencing errors, it leads to a likely underestimation of terminal branch lengths, especially for lineages with few samples (e.g., subtribes Schreberinae and Ligustrinae). The SNP alignments we generate are therefore not suitable for analyses that rely on evolutionary rate information, such as molecular dating. While these potential problems result from the low-depth sequencing combined with the large evolutionary scale considered, our analytical pipeline deals with these difficulties by focusing on phylogenetically-informative markers that are conserved across large evolutionary scales, variable, and parsimony informative. Topology inferences are therefore expected to be robust, and the nuclear SNP phylogenies we infer are indeed consistent using different reference genomes and SNP filters.

The SNPs obtained with different reference genomes cannot be directly compared, but the two datasets obtained from the olive and ash whole references differ in the total number of SNPs and missing data among samples (Table S1). The topologies inferred using the two whole genome datasets are still mostly congruent and present only eight differences all associated with short branches (Figures S6 and S9). Four of these differences occur within the *Olea* sect. *Olea* or *Fraxinus* group, the two lineages that include one of the reference genomes (Figures S6 and S9). Furthermore, branch lengths within these two groups vary among the two whole genome datasets, while those in the rest of the tree are largely unaffected by the use of different reference genomes (Figures S6 and S9).

The topological incongruence between the phylogenies obtained with the two whole reference genomes, both of which are fragmented, might stem from the high number of closely related duplicates in the olive tree genome, either due to a recent polyploidization (Julca et al., 2018) or

failure to collapse divergent alleles in the genome assembly (e.g., Hahn, Zhang, & Moyle, 2014; Pryszcz & Gabaldón, 2016). The latter hypothesis can be argued for because the basic chromosome number ($n$ = 23; Taylor, 1945) is stable in Oleeae and the sequenced olive cultivar ('Farga'; Cruz et al., 2016) is highly heterozygous due to recent admixture between distinct gene pools (Besnard et al., 2013; Diez et al., 2015). Similar persistence of duplicates might exist in the *Fraxinus* genome. Retention of duplicates in the reference genome assemblies would result in reads from accessions closely related to the reference having reads that uniquely map to the most similar duplicate. However, reads from more distantly related species would map to neither or both of the duplicates equally well, rendering SNP calling from these regions problematic as they would have low quality scores. Obtaining SNP alignments from reference genomes with different complexities could therefore cause phylogenetic incongruences related to, for example, clustering of missing data (Bertels et al., 2015; Xi et al., 2016; Nute et al., 2018). However, the overall congruence between the topologies obtained with the two references shows that relying on a single reference is a viable option for groups where genomic resources are sparse.

#### *4.1.2 Using orthologous CDS reduces the genome complexity*

Previous methods, such as REALPHY (Bertels et al., 2015), have been designed to incorporate multiple reference genomes in large scale prokaryote phylogenies. The REALPHY pipeline relies on mapping of pseudo-reads obtained from multiple references to a set of genome assemblies and reduces the final SNP alignment down to orthologous positions based on reciprocal mapping (Bertels et al., 2015). Here we expand on this concept to fit complex genomes of eukaryotes, by establishing orthology via reciprocal best-hit BLAST searches between annotated CDS extracted from two available genomes. We therefore consider only sets of sequences descended from a single gene in the last common ancestor of the compared genomes (i.e., co-orthologs). Lineage-specific duplicates or un-collapsed alleles would either be discarded, or represented as a single co-ortholog, which would remove some of the variation in mapping between highly divergent samples. This

reduction in the complexity of the reference comes at the expense of the number of sites considered, but because reads from distant relatives almost exclusively map to coding sequences, focusing on CDS also decreases the disparity among taxa. The effect of the distance to the reference genome and the resulting phylogenetic clustering indeed disappeared from the combined SNPs alignment (Figures 4, 5, S6, and S9). Instead, the amount of missing data in this dataset was a function of the sequencing depth, and samples with depth above 3× all had less than 10% missing data, which can be partially accounted for by gene losses. Our method can therefore be used to generate SNP alignments with very low amounts of missing data by slight increases of the sequencing depth, which is doable even for herbarium samples that are more than 130 years old (see for example *Schrebera swietenioides*; Table S1).

The SNPs obtained from mapping to orthologous CDS are only partially overlapping with those obtained from the mapping to the whole genomes. This is linked to the reduced complexity of the CDS genomes, and, hence fewer bases are considered, to the filters used to remove low quality SNPs from the alignments. In addition, even when the positions are overlapping the identity of the genotypes can differ as the consensus of the two genomes is used in the CDS alignment (e.g. a SNP called as 'C' with the olive genome but 'T' with the ash genome would be 'Y' in the consensus CDS alignment). Despite this, the topology obtained from the orthologous CDS SNP alignment is highly congruent with those based on either the whole ash or olive genomes (Figures 4, S6, and S9). It presents the same number of differences to each of them (four branches), which are not statistically supported (non-significant SH tests; Table S2; Figure 4). We therefore conclude that reducing the genome complexity prior to mapping and merging SNPs called from multiple orthologous CDS references is preferable, and should be considered when analysing low-depth sequence data spanning large evolutionary groups. In the absence of multiple reference genomes, orthologous CDS extracted from transcriptomes can be used to increase the number of references.

Although we show that it is viable to obtain concatenated SNP alignments from a large

number of taxa mapped to a single reference genome, these SNPs cannot easily be used to infer single gene trees. By contrast, such gene trees can directly be inferred from SNPs obtained from a set of co-orthologs. While these trees can be used to detect functional genetic changes (e.g., Besnard et al., 2014, 2018; Christin et al., 2012a; Dunning et al., 2013; Yokoyama, Tada, Zhang, & Britt, 2008), the multitude of gene trees can also be used for coalescence analyses. Our coalescence species tree is compatible with all three nuclear SNP topologies (Figures 4, 6, S6, and S9). However, many nodes, including all of those differing among the three concatenated SNP alignments, vary among gene trees (Figures 4, 6, S6, and S9). The only nodes supported by a large majority of gene trees are those associated with long branches, which combine long evolutionary times allowing coalescence with the accumulation of informative mutations (Figure 6). The high number of nodes where gene trees differ likely stems from a combination of incomplete lineage sorting and/or hybridization in the group, and in some cases a lack of sequence information due to missing data.

The approach presented here treats all individuals as diploid, and generates a single set of SNPs for each individual gene marker. In cases where the phylogenetic origin of the two alleles at each locus is the same, the history of the species would be correctly inferred by the concatenated and/or coalescence species trees. This encompasses the possibilities of ancient hybridization followed by homogenization of the alleles through recombination and/or losses of one of the alleles, which would be evidenced by gene tree discordance in the coalescence analysis. Similarly, autopolyploids would be correctly placed in the phylogeny as the multiple alleles at each of their loci come from the same parental species. Considering a single sequence per locus will by definition not identify cases where the two alleles belong to distinct phylogenetic groups, as might be the case of recent hybrids or allopolyploids involving parents from distinct clades, which will both possess sets of non-recombining alleles. Resolving such cases requires allele phasing, which can be done only with consequent sequencing efforts. Such detailed investigations should be conducted in the future for groups with proven neopolyploidy and where topological incongruence

might be linked to allopolyploidization.

### 4.2 Hard discrepancies suggest reticulate evolution

The phylogenetic incongruence between plastomes and nrDNA mirror previous reports (Figures 2 and S2; Hong-Wa & Besnard, 2013; Zedane et al., 2016), and probably mostly result from heterogeneous evolutionary rates in the nrDNA cluster, with an acceleration in GC-rich groups resulting in a form of long-branch attraction of these lineages (Stiller & Hall, 1999; Bergsten, 2005). We therefore conclude that the nrDNA cluster is not a reliable marker for the group. On the other hand, the plastome and nuclear SNP phylogenies are similar (Figures 2, 4, S6, and S9). . However, the nuclear SNP and plastome topologies present a number of differences concerning relationships among terminal branches, as well as in some deeper nodes (Figures 2, 4, S6, and S9). For intraspecific or intrageneric relationships, these discrepancies probably mirror the different dispersal abilities of pollen+seed- and seed-transported markers that can be extenuated by increased genetic drift of the organelle genomes. Notably, the nuclear datasets strongly improve the resolution within the olive tree lineage (*Olea europaea*; Figures 2, 4, S6, and S9). The vast majority of the incongruences can be explained by incomplete lineage sorting and/or hybridization within the tribe as detected in the coalescence analysis. Some differences are, however, supported by all nuclear SNP datasets (Figures 2, 4, S6, and S9), indicating that the evolutionary history of the plastome in some cases differs from that of the majority of the nuclear genome. In particular, each of the genera *Picconia* and *Phillyrea* is supported as monophyletic by all nuclear SNP alignments and the vast majority gene trees (Figures 4, 6, S6, and S9), yet the Mediterranean *Phillyrea* is nested within the Macaronesian *Picconia* in the plastome phylogeny (Figure 2). This hard incongruence suggests that cytoplasmic capture might also have played a role in shaping the differences in evolutionary histories between the two genomic regions.

## 5. CONCLUSIONS

Obtaining well supported nuclear phylogenies that accurately capture the history of the group is essential for evolutionary studies. Using the complex plant tribe Oleeae (family Oleaceae) as a model system, we show that reliable phylogenetic trees can be obtained from low-depth sequencing data. Reticulated evolution and incomplete lineage sorting coupled with a phylogenetic clustering of missing data might, however, cause slight discrepancies in phylogenetic topologies when different reference genomes are used to obtain SNP alignments. We show that using orthologous CDS from multiple genomes can overcome such problems, by removing the effect of the distance to the reference and allowing inferences of multigene coalescence-based species tree. Importantly, our analyses suggest that very low levels of missing data can be achieved with sequencing depths around 3×, which can be achieved with herbarium samples. We further predict that some of the observed issues caused by SNP filtering necessary for low-depth sequencing datasets will likely improve when more taxa are included in the phylogenies, allowing better estimates of terminal branch lengths. Therefore, analyses of low-depth shotgun sequencing can infer nuclear phylogenies, potentially shedding new light on the evolutionary history of functional traits. As this sequencing approach is also suitable for samples obtained from natural history collections, it will allow for increased species sampling especially of rare or recently extinct lineages, as well as taxa occurring in remote areas. Low-depth sequence data are continuously generated from herbarium samples with the purpose to assemble plastomes, and widespread application of our approach will therefore allow the inference of large nuclear phylogenetic trees, fuelling diverse evolutionary and ecological investigations.

## ACKNOWLEDGEMENTS

JKO and LTD are funded by ERC grant ERC-2014-STG-638333 and NERC grant NE/M00208X/1. GB and JKO received funding from the TULIP LABEX visiting scientist program (AO "Visiting Scientist Printemps 2017") to enhance collaboration on the Oleaceae phylogeny. LZ was funded by

the University of Al Furat, Syria. PAC is supported by a Royal Society Research Fellowship (grant number URF120119). CVDP and GB are members of the Laboratoire Evolution & Diversité Biologique (EDB), part of the LABEX TULIP managed by Agence Nationale de la Recherche (ANR; no. ANR-10-LABX-0041). They were funded by the Regional Council Midi-Pyrénées (AAP 13053637, 2014-EDB-UT3-DOCT) and the ERA-NET BiodivERsa framework with INFRAGECO (Inference, Fragmentation, Genomics, and Conservation, ANR-16-EBI3-0014). We also acknowledge an Investissement d'Avenir grant of the ANR (CEBA: ANR-10-LABX-25-01). This work was performed within the framework of the PhyloAlps project, whose sequencing was funded by France Genomique (ANR-10-INBS-09-08). L. Csiba and E. Kepos (Jodrell Laboratory) provided DNA extracts from accessions of the living collection of the Royal Botanic Gardens, Kew. We are also grateful to D. Stadie (Eisleben), M. Dosmann and K. Richardson (Arnold Arboretum, Harvard), E. Bellefroid (National Botanical Garden of Belgium), O. Maurin, M.S. Vorontsova, and D. Goyder (Royal Botanical Gardens, Kew), T. Haevermans and M. Gaudeul (Muséum National d'Histoire Naturelle de Paris), H. Esser (Munich Botanical Gardens), J. Razanatsoa and F. Rakotonasolo (Parc de Tsimbazaza, Antananarivo), P. Saumitou-Laprade and P. Vernet (Evo-Eco-Paléo Lille), T. Josseberger and A. Krämer (Botanische Gärten der Universität Bonn), A. Rinfret-Pilo (Botanical Garden of Montreal), S. Blackwell (Botanical Garden of Phoenix), J. Munzinger (IRD Montpellier), R. Lima and G. Frey (Universidade de São Paulo), the herbarium of the Royal Botanical Garden of Edinburgh, The United States Department of Agriculture, and the Charles R. Keith Arboretum for providing samples. We also thank E. Coissac, J. Hackel, R. Kiew, S. Lavergne and J. Murienne for constructive discussions on this project, and A. Iribar, H. Holota and O. Bouchez for lab assistance.

## AUTHOR'S CONTRIBUTIONS

GB conceived and planned the study with input from PAC. GB, LZ and CHW sampled and prepared DNA for HiSeq sequencing. JKO analyzed the nuclear data with input from PAC and

LTD. IC, CVDP, and LZ analyzed and interpreted the plastome data with input from GB. LZ and JKO analyzed the nuclear ribosomal data with the help of GB. JKO and GB interpreted the analyses and wrote the paper with the input from all authors.

## DATA ACCESSIBILITY

All raw reads are available in the short sequence archive under accession no. PRJNA506987 and PRJEB30497. In the NCBI nucleotide database, all newly assembled chloroplast genomes and ribosomal DNA clusters are available under accession numbers specified in Table S1. All scripts are available on Github https://github.com/jill-olofsson/low-depth-sequencing_analyses.

## REFERENCES

Allen, J. M., Boyd, B., Nguyen, N. P., Vachaspati, P., Warnow, T., Huang, D. I., Grady, P. G., Bell, K. C., Cronk, Q. C., Mugisha, L., & Pittendrigh, B. R. (2017). Phylogenomics from whole genome sequences using aTRAM. *Systematic Biology, 66,* 786–798.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment tool. *Journal of Molecular Biology*, *215,* 403–410.

Álvarez, I., & Wendel, J. F. (2003). Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution*, *29*, 417–434.

Bakker, F. T., Lei, D., Yu, J., Mohammadin, S., Wei, Z., van de Kerke, S., . . . Holmer, R. (2016). Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biological Journal of the Linnean Society*, *117*, 33–43.

Bennett, M.D., & Leitch, I. J. (2012). Plant DNA *C*-value database (release 6.0, Dec. 2012). http://data.kew.org/cvalues/CvalServlet?querytype=1

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, *21*, 163–193.

Bertels, F., Silander, O. K., Pachkov, M., Rainey, P.B., & van Nimwegen, E. (2015). Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Molecular Biology and Evolution*, *31*, 1077–1088.

Besnard, G., Rubio de Casas, R., Christin, P-A., & Vargas, P. (2009). Phylogenetics of *Olea* (Oleaceae) based on plastid and nuclear ribosomal DNA sequences: Tertiary climatic shifts and lineage differentiation times. *Annals of Botany*, *104*, 143–160.

Besnard, G., El Bakkali, H., Haouane, H., Baali-Cherif, D., Moukhli, A., & Khadari, B. (2013). Population genetics of Mediterranean and Saharan olives: geographic patterns or differentiation and

evidence for early generations of admixture. *Annals of Botany*, *112*, 1293–1302.

Besnard, G., Christin, P-A., Malé, P-J. G., Lhuillier, E., Lauzeral, C., Coissac, E., & Vorontsova, M.S. (2014). From museums to genomics: old herbarium specimens shed light on a $C_3$ to $C_4$ transition. *Journal of Experimental Botany*, *65*, 6711–6721.

Besnard, G., Bianconi, M. E., Hackel, J., Manzi, S., Vorontsova, M. S., & Christin, P-A. (2018). Herbarium genomics retraces the origins of $C_4$-specific carbonic anhydrase in Andropogoneae (Poaceae). *Botany Letters*, *165*, 419–433.

Bieker, V., & Martin, M. (2018). Implications and future prospects for evolutionary analyses of DNA in historic herbarium collections. *Botany Letters*, *165*, 409–418.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, CH., Suchars, A. R, . . . Drummond, A. J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, *10*, e1003537.

Buerki, S., & Baker, W. J. (2016). Collections-based research in the genomic era. *Biological Journal of the Linnean Society*, *117*, 5–10.

Buerkle, C. A., & Gompert, Z. A. (2013). Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, *22*, 3028–3035.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*, 540–553.

Christin, P-A., Edwards, E. J., Besnard, G., Boxall, S. F., Gregory, R., Kellogg, E. A., Hartwell, J., . . . Osborne, C. P. (2012a). Adaptive evolution of $C_4$ photosynthesis through recurrent lateral gene transfer. *Current Biology*, *22*, 445–449.

Christin, P-A., Wallace, M. J., Clayton, H., Edwards, E. J., Furbank, R. T., Hattersley, P. W., Sage, R. F., . . . Ludwig, M. (2012b). Multiple photosynthetic transitions, polyploidy, and lateral gene transfer in the grass subtribe Neurachninae. *Journal of Experimental Botany*, *63*, 6297–6308.

Cruz, F., Julca, I., Gómez-Garrido, J., Loska, D., Marcet-Houben, M., Cano, E., . . . Gabaldón, T. (2016). Genome sequence of the olive tree, *Olea europaea*. *GigaScience*, *5*, 29.

Danecek, P., Auton, A., Abecasis, G., Banks, E., DePristo, M. A., . . . Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158.

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, *6*, 361–375.

Diez, C., Trujillo, I., Martinez-Urdiroz, N., Barranco, D., Rallo, L., Marfil, P., & Gaut, B. S. (2015). Olive domestication and diversification in the Mediterranean Basin. *New Phytologist*, *206*, 436–447.

Dodsworth, S. (2015). Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science*, *20*, 525–527.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. B., Smith, S. A., . . . Giribet, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, *452*,

745–749.

Dunning, L. T., Dennis, A. B., Thomson, G., Sinclair, B. J., Newcomb, R. D., & Buckley, T. R. (2013). Positive selection in glycolysis among Astralasian stick insects. *BMC Evolutionary Biology, 13*, 215.

Dunning, L. T., Liabot, A. L., Olofsson, J. K., Smith, E. K., Vorontsova, M. S., Besnard, G., . . . Lehmann, C. E. R. (2017). The recent and rapid spread of *Themeda triandra*. *Botany Letters*, *164*, 327–337.

Edwards, E. J., Osborne, C. P., Strömberg, C. A., Smith, S. A., & Grasses Consortium (2010). The origin of $_4$ $C_4$ grasslands: Integrating evolutionary and ecosystem science. *Science*, *328*, 587–591.

Green, P. S. (2004). Oleaceae. In: Kubitzki, K. & Kadereit, J. W. (eds), *The Families and Genera of Vascular Plants*. Vol. VII: *Flowering Plants, Dicotyledons*. New York: Springer, pp. 296–306.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircger, M., . . . Pääbo, S. (2010). A draft sequence of the Neanderthal genome. *Science*, *328*, 710–722.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0, *Systematic Biology*, *59*, 307–321.

Guo, S. Q., Xiong, M., Ji, C. F., Zhang, Z. R., Li, D. Z., & Zhang, Z. Y. (2011). Molecular phylogenetic reconstruction of *Osmanthus* Lour. (Oleaceae) and related genera based on three chloroplast intergenic spacers. *Plant Systematics and Evolution*, *294*, 57–64.

Hahn, M. W., Zhang, S. V., & Moyle, L. C. (2014). Sequencing, assembling, and correcting draft genomes using recombinant populations. *Genes Genomes Genetics*, *4*, 669–679.

Hinsinger, D. D., Bask, J., Gaudeul, M., Cruaud, C., Bertolino, P., Frascaria-Lacoste, N., . . . Bousquet, J. (2013). The phylogeny and biogeographic history of ashes (*Fraxinus*, Oleaceae) highlight the roles of migration and vicariance in the diversification of temperate trees. *PLoS ONE*, *8*, e80431.

Hong-Wa, C., & Besnard, G. (2013). Intricate patterns of phylogenetic relationships in the olive family as inferred from multi-locus plastid and nuclear DNA sequence analyses: A close-up on *Chionanthus* and *Noronhia* (Oleaceae). *Molecular Phylogenetics and Evolution*, *67*, 367–378.

Jeffroy, O., Brinkmann, H., Delsuc, F., & Philippe, H. (2006). Phylogenomics: the beginning of incongruence? *Trends in Genetics, 22*, 225–231.

Jetz, W., Thomas, G. H., Joy, J. B., Hartmass, K., & Mooers, A. O. (2012). The global diversity of birds in space and time. *Nature*, *491*, 444–448.

Julca, I., Marcet-Houben, M., Vargas, P., & Gabaldón, T. (2018). Phylogenomics of the olive tree (*Olea europaea*) revels the relative contribution of ancient allo- and autopolyploidization events. *BMC Biology*, *16*, 15.

Junier, T., & Zdobnov, E. M. (2010). The Newick Utilities: High-throughput phylogenetic tree processing in the UNIX shell, *Bioinformatics, 26*, 1669-1670.

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

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics, 15*, 356.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, *33*, 1870–1874.

Langmead, B., & Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–359.

Leaché, A. D., Banbury, B. L., Felsenstein, J., de Oca, A. N. M., & Stamatakis, A. (2015). Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. *Systematic Biology*, *64*, 1032–1047.

Lefort, V., Longueville, J.M., & Gascuel, O. (2017). SMS: Smart model selection in PhyML. *Molecular Biology and Evolution*, *34*, 2422–2424.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, *50*, 913–925.

Li, H. (2011a). Improving SNP discovery by base alignment quality. *Bioinformatics, 27*, 1157–-1158.

Li, H. (2011b). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27***,** 2987–2993.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079.

Li, J., Alexander, J. H., & Zhang, D. (2002). Paraphyletic *Syringa* (Oleaceae): evidence from sequences of nuclear ribosomal DNA ITS and ETS regions. *Systematic Botany*, *27*, 592–597.

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., & Abecasis, G. R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome Research*, *21*, 940–951.

Lundgren, M. R., Besnard, G., Ripley, B. S, Lehmann, C. E. R., Chatelet, D. S., Kynast, R. G., . . . Christin, P-A. (2015). Photosynthetic innovation broadens the niche within a single species. *Ecology Letters*, *18*, 1021–1029.

Malé, P-J. G., Bardon, L., Besnard, G., Coissac, E., Delsuc, F., Engel, J., . . . Chave, J. (2014). Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. *Molecular Ecology Resources*, *14*, 966–975.

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., International Wheat Genome Sequencing Consortium, . . . Olsen, O. A. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, *345*, 1250092.

McCormack, E., Tsai, W. L. E., & Faircloth, B. C. (2016). Sequence capture of ultraconserved

elements from bird museum specimens. *Molecular Ecology Resources*, *16*, 1189–1203.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M., S., & Warnow, T. (2014). ASTRAL: Genome-Scale Coalescent-Based Species Tree. *Bioinformatics, 30,* i541–i548.

Nieto Feliner, G., & Rosselló, J. A. (2007). Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. *Molecular Phylogenetics and Evolution*, *44*, 911–919.

Nute, M., Cho, J., Molloy, E. K., & Warnow, T. (2018). The performance of coalescent-based species tree estimation methods under models of missing data. *BMC Genomics, 19*, 286.

Olofsson, J. K., Bianconi, M., Besnard, G., Dunning, L. T., Lundgren, M. R., Holota, H., . . . Christin, P-A. (2016). Genome biogeography reveals the intraspecific spread of adaptive mutations for a complex trait. *Molecular Ecology*, *25*, 6107–6123.

Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE*, *7*, e30619.

Paterson, A. H., Freeling, M., Tang, H. B., & Wang, X. Y. (2009). Insights from the comparison of plant genome sequences. *Annual Review of Plant Biology*, *61*, 349–372.

Pryszcz, L. P., & Gabaldón, T. (2016). Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, *8*, e113.

Puttick, M. N., Morris, J. L., Williams, T. A., Cox, C. J., Edwards, D., Kenrick, P., . . . Donoghue, P. C. J. (2018). The interrelationships of land plants and the nature of the ancestral embrophyte. *Current Biology*, *28*, e2.

Revell, L. J., Johnson, M. A., Schulte, J. A., Kolbe, J. J., & Losos, J. B. (2007). A phylogenetic test for adaptive convergence in rock-dwelling lizards, *Evolution*, *61*, 2898–2912.

Roquet, C., Coissac, É., Cruaud, C., Boleda, M., Boyer, F., Alberti, A., . . . Lavergne, S. (2016). Understanding the evolution of holoparasitic plants: the complete plastid genome of the holoparasite *Cytinus hypocistis* (Cytinaceae), *Annals of Botany, 118,* 885–896.

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E, Goodhead, I., Herrero, J., . . . Durbin, R. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, *483*, 169–175.

Shimodaira, H., & Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, *16*, 1114–1116.

Sollars, E. S. A., Harper, A. L., Kelly, L. J., Sambles, C. M., Ramirez-Gonzalez, R. H., Swarbreck, D., . . . Buggs, R. J. A. (2017). Genome sequence and genetic diversity of European ash trees. *Nature*, *541*, 212–216.

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

Stiller, J. W., & Hall, B. D. (1999). Long-branch attraction and the rDNA model of early eukaryotic evolution. *Molecular Biology and Evolution, 16*, 1270–1279.

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

Taylor, H. (1945). Cyto-taxonomy and phylogeny of the Oleaceae. *Brittonia*, *5*, 337–367.

Van de Paer, C., Hong-Wa, C., Jeziorski, C., & Besnard, G. (2016). Mitogenomics of *Hesperelaea*, an extinct genus of Oleaceae. *Gene, 594*, 197–202.

Van de Paer, C., Bouchez, O., & Besnard, G. (2018). Prospects on the evolutionary mitogenomics of plants: a case study on the olive family (Oleaceae). *Molecular Ecology Resources*, *18*, 409–423.

Wallander, E., & Albert, V. A. (2000). Phylogeny and classification of Oleaceae based on *rps16* and *trnL-F* sequence data. *American Journal of Botany*, *87*, 1827–1841.

Welch, A. J., Collins, K., Ratan, A., Drautz-Moses, D. I., Schuster, S. C., & Lindqvist, C. (2016). The quest to resolve recent radiations: plastid phylogenomics of extinct and endangered Hawaiian endemic mints (Lamiaceae). *Molecular Phylogenetics and Evolution*, *99*, 16–33.

Xi, Z., Liu, L., & Davis, C. C. (2016). The impact of missing data on species tree estimation. *Molecular Biology and Evolution, 33*, 838–860.

Yang, Z. (2007). PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*, 1586–1591.

Yokoyama, S., Tada, T., Zhang, H., & Britt, L. (2008). Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. *Proceedings of the National Academy of Sciences of the United States of America, 105*, 13480–13485.

Yuan, W. J., Zhang, W. R., Han, Y. J., Dong, M. F., & Shang, F. D. (2010). Molecular phylogeny of *Osmanthus* (Oleaceae) based on non-coding chloroplast and nuclear ribosomal internal transcribed spacer regions. *Journal of Systematics and Evolution*, *48*, 482–489.

Zedane, L., Hong-Wa, C., Murienne, J., Jeziorski, C., Baldwin, B. G., & Besnard, G. (2016). Museomics illuminate the history of an extinct, paleoendemic plant lineage (*Hesperelaea*, Oleaceae) known from an 1875 collection from Guadalupe Island, Mexico. *Biological Journal of the Linnean Society*, *117*, 44–57.

Zhang, F., Ding, Y., Zhu, C.F., Zhou, X., Orr, M.C., Scheu, S., & Luan, Y.X. (2019). Phylogenomics from low-coverage whole-genome sequencing. *Methods in Ecology and Evolution,* doi.org/10.1111/2041-210X.13145.

# FIGURE CAPTIONS

**FIGURE 1** Flow-chart showing the different steps in our approach for low depth whole genome sequencing data. SNP = single nucleotide polymorphism; CDS = coding sequences.

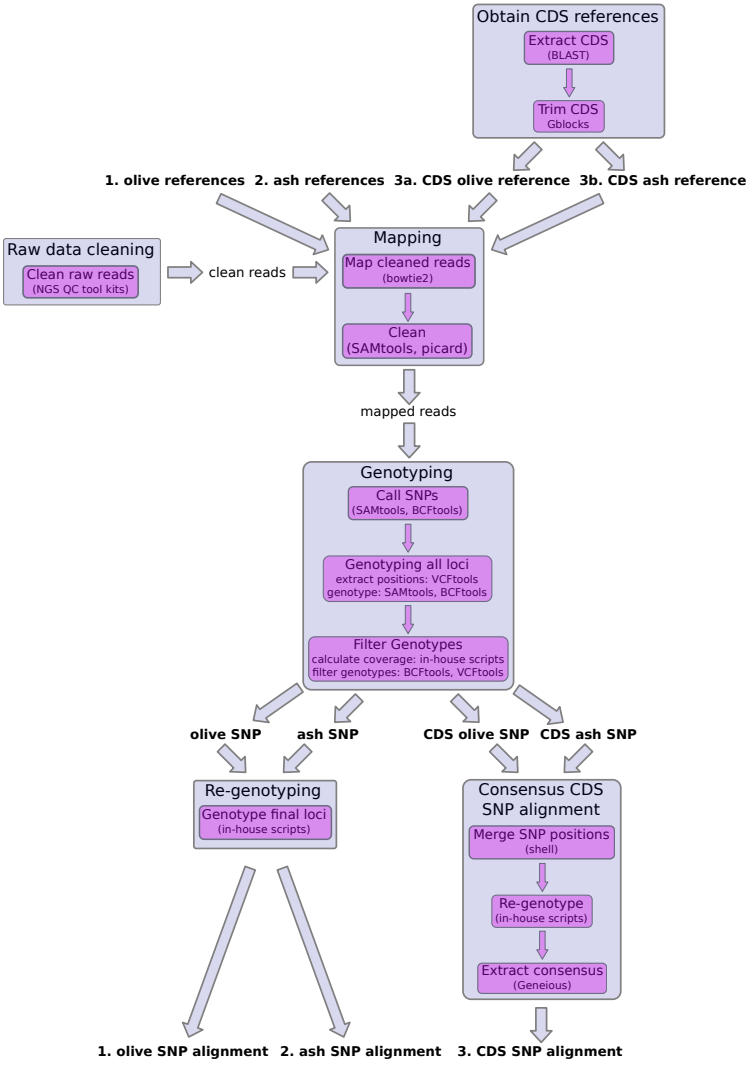
**FIGURE 2** Maximum-likelihood phylogenetic tree inferred in RAxML v. 8 (Stamatakis, 2014) under a GTR + G + I substitution model from the alignment of whole plastomes. Monophyletic clades and genera are denoted and geographic origin of each taxon is shown by coloured circles. Nodes support was evaluated with 100 bootstrap replicates and is indicated near branches (* = 100%). Nodes denoted in red represent difference to at least one of the nuclear SNP topologies (Figures 4, S6, and S9). Branch length is given as expected number of substitution per site.

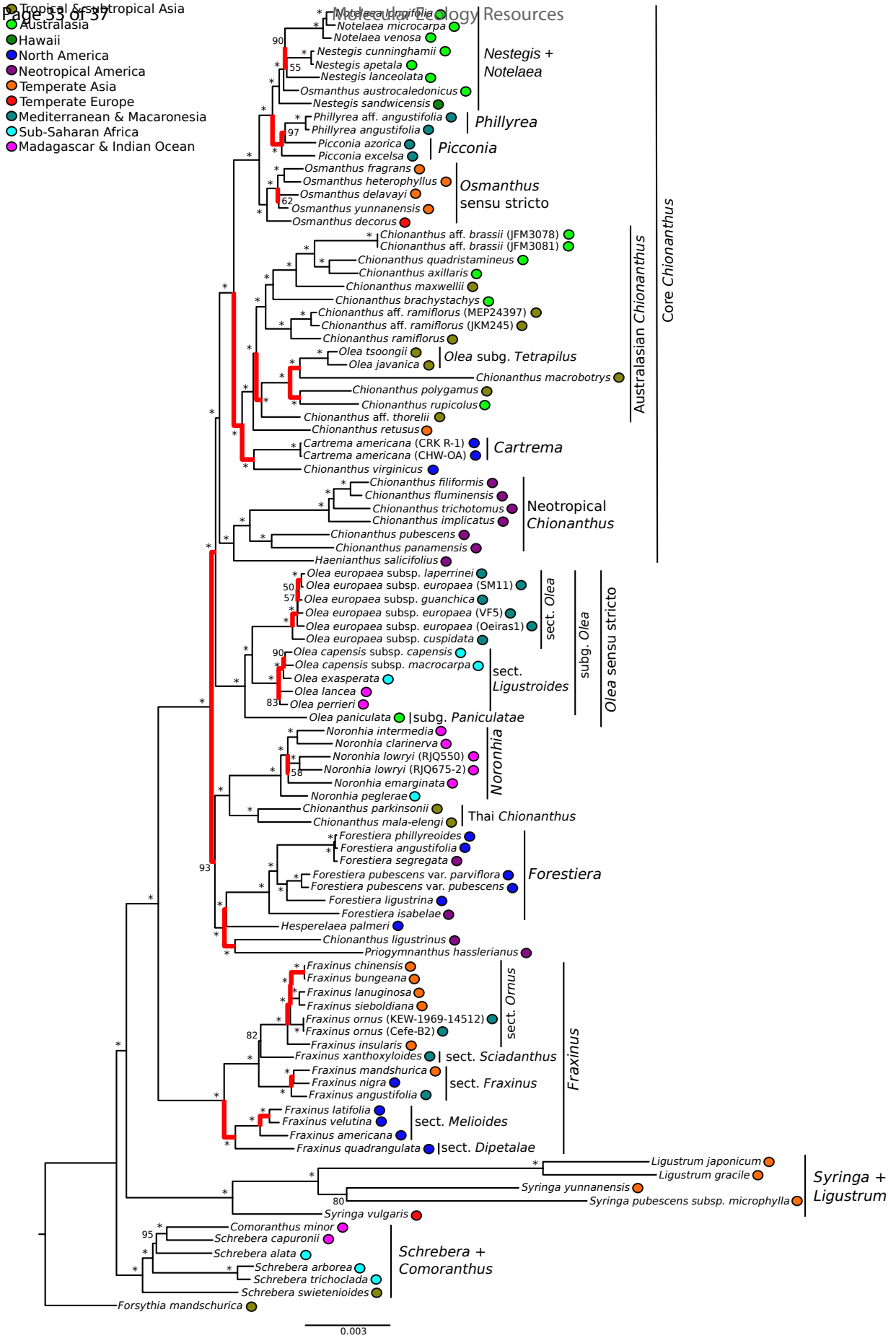
**FIGURE 3** Decrease in mapping success with increasing divergence from the reference genome. Divergence times are taken from the dated phylogenetic tree obtained from whole plastome alignments (Figure S1).

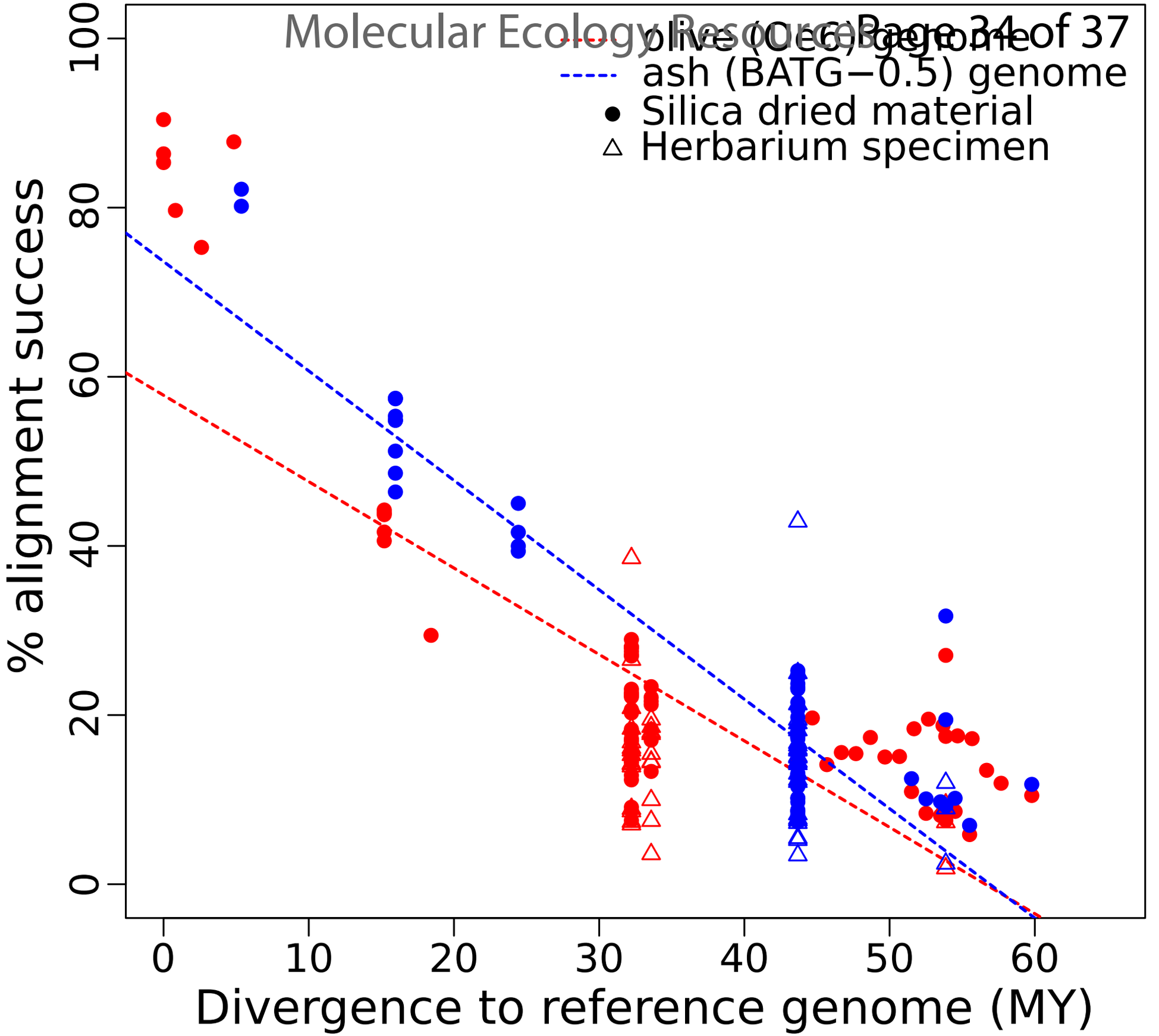
**FIGURE 4** Maximum-likelihood phylogenetic tree inferred in RAxML v. 8 (Stamatakis, 2014) under GTR + G substitution model from the consensus alignment of nuclear single nucleotide polymorphisms (SNPs) obtained after mapping reads to the best reciprocal blast hits between annotated coding sequences (CDS) in the olive (Oe6; Cruz et al., 2016) and ash genome (BATG-0.5; Sollars et al., 2017). Monophyletic clades and genera are denoted and the geographic origin of each taxon is denoted by coloured circles. Nodes support was evaluated with 100 bootstrap replicates and is indicated near branches (* = 100%). Nodes denoted in dark blue represent differences to the nuclear SNP topology obtained after mapping to the whole olive genome (Oe6; Cruz et al., 2016; Figure S6). Nodes denoted in light blue represent differences to the nuclear SNP topology obtained after mapping to the whole ash genome (BATG-0.5; Sollars et al., 2017; Figure S9). Branch length is given as expected number of substitution per site.

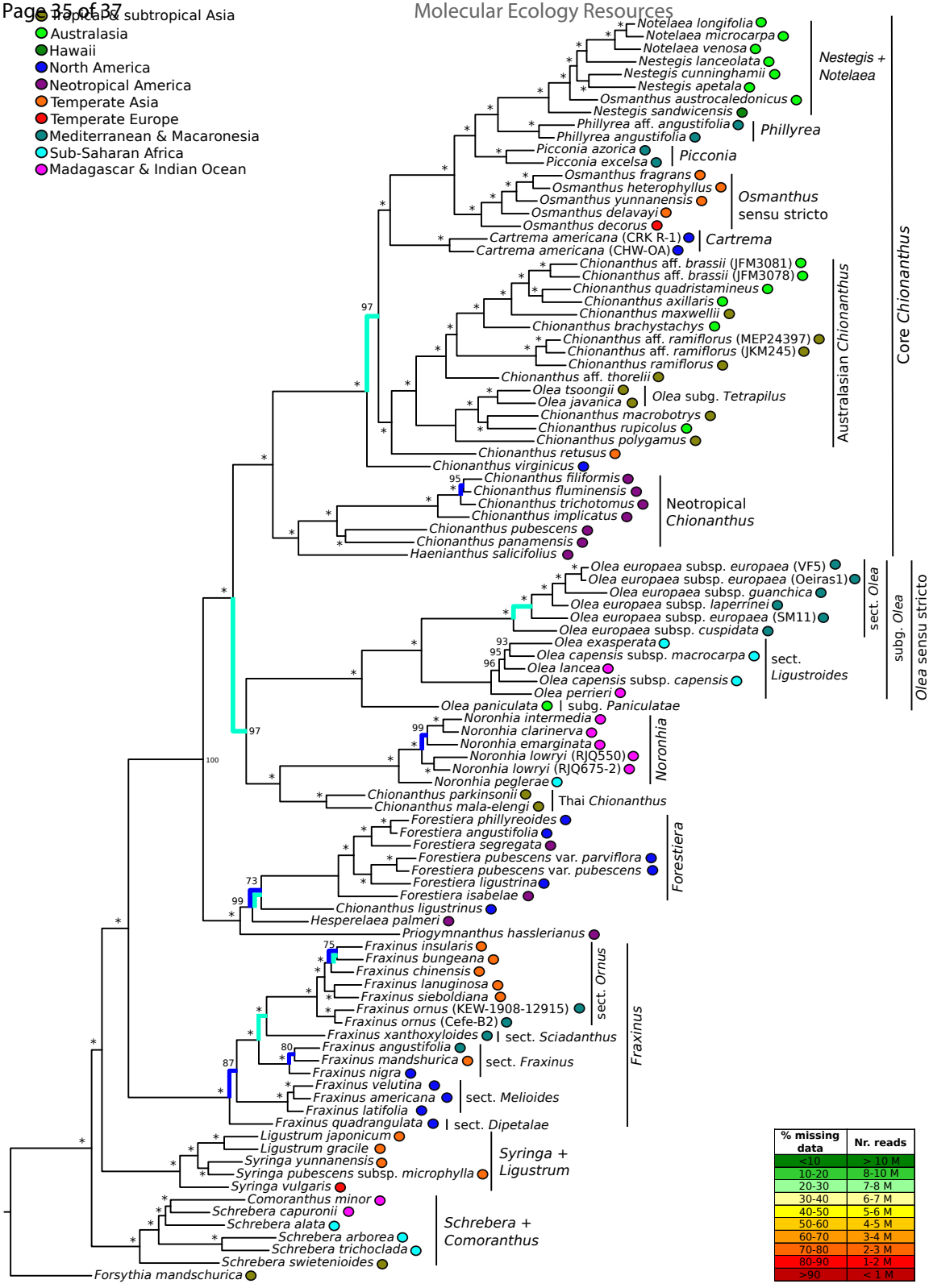
**FIGURE 5** Impact of the reference genome, the divergence time and the sequencing depth on the level of missing data. Correlation between A) number of genotyped single nucleotide polymorphisms (SNP) from coding sequences (CDS) and divergence to the olive reference genome (Oe6; Cruz et al., 2016); B) number of genotyped SNPs from CDS and divergence to the ash genome (BATG-0.5; Sollars et al., 2017); C) percentage of missing data in the final CDS alignment and divergence from the reference genome (red - olive; blue - ash); and D) percentage of missing data in the final CDS alignment and the estimated sequencing depth for the olive genome.

**FIGURE 6** Coalescence species tree obtained from single nuclear polymorphism (SNP) alignments (> 50 bp) from 1,400 gene trees of orthologous coding sequences (CDS) in the olive (Oe6; Cruz et al., 2016) and ash genome (BATG-0.5; Sollars et al., 2017) using ASTRAL v. 5.6.2 (Mirarab et al., 2014). Posterior support values are indicated near branches. Proportion of gene trees supporting the three alternative quartet topologies are indicated at nodes with the blue proportion indicating the quartet supporting the species topology. Branch length is given in coalescent units.


Obtain CDS references
Extract CDS
(BLAST)
Trim CDS
Gblocks
1. olive references
2. ash references
3a. CDS olive reference
3b. CDS ash reference
Raw data cleaning
Clean raw reads
(NGS QC tool kits)
clean reads
Mapping
Map cleaned reads
(bowtie2)
Clean
(SAMtools, picard)
mapped reads
Genotyping
Call SNPs
(SAMtools, BCFtools)
Genotyping all loci
extract positions: VCFtools
genotype: SAMtools, BCFtools
Filter Genotypes
calculate coverage: in-house scripts
filter genotypes: BCFtools, VCFtools
olive SNP
ash SNP
CDS olive SNP
CDS ash SNP
Re-genotyping
Genotype final loci
(in-house scripts)
Consensus CDS
SNP alignment
Merge SNP positions
(shell)
Re-genotype
(in-house scripts)
Extract consensus
(Geneious)
1. olive SNP alignment
2. ash SNP alignment
3. CDS SNP alignment

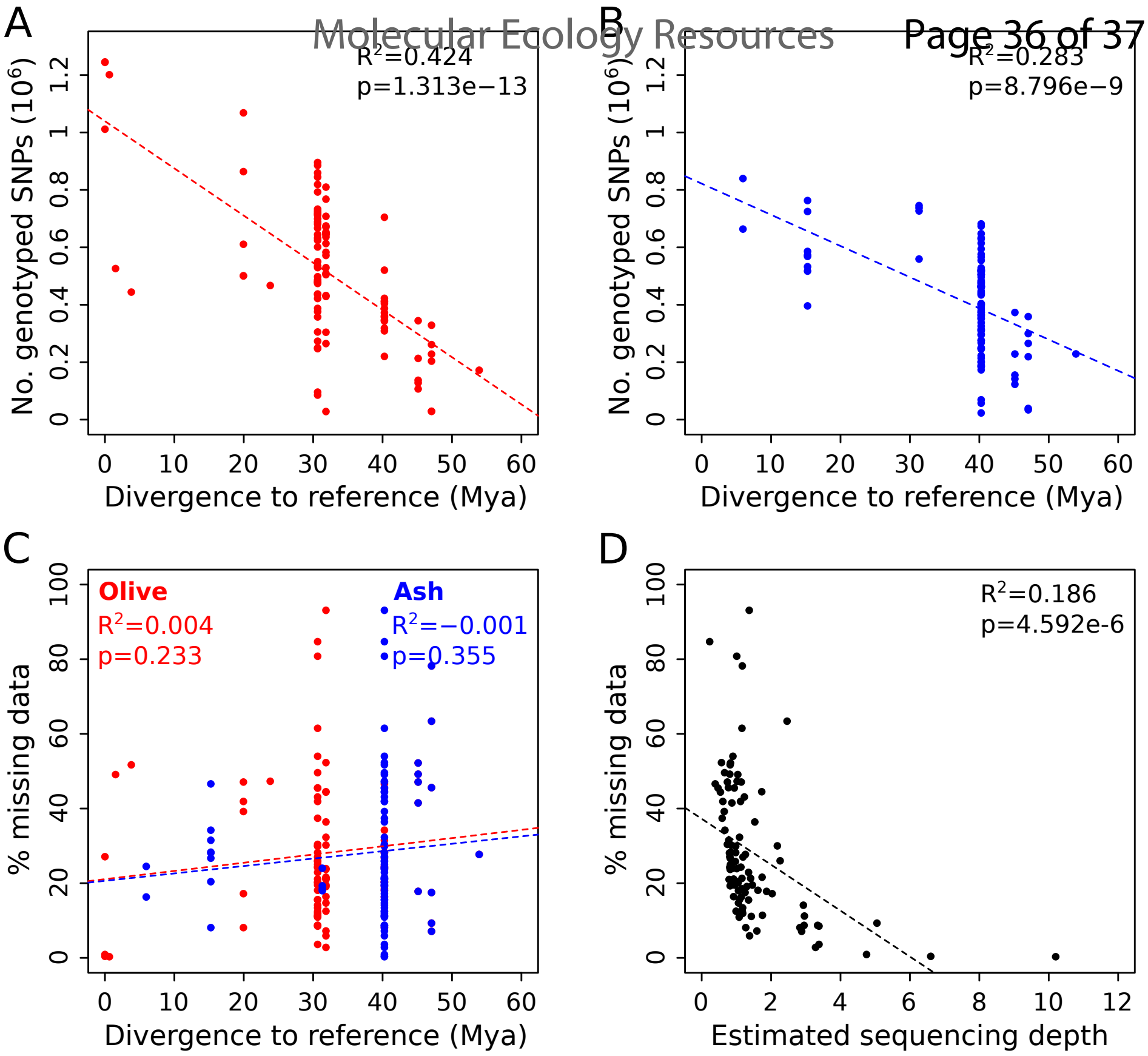

Tropical & subtropical Asia
Australasia
Hawaii
North America
Neotropical America
Temperate Asia
Temperate Europe
Mediterranean & Macaronesia
Sub-Saharan Africa
Madagascar & Indian Ocean
Notelaea ligustrifolia
Notelaea microcarpa
Notelaea venosa
Nestegis cunninghamii
Nestegis apetala
Nestegis lanceolata
Osmanthus austrocaledonicus
Nestegis sandwicensis
Nestegis + Notelaea
Phillyrea aff. angustifolia
Phillyrea angustifolia
Phillyrea
Picconia azorica
Picconia excelsa
Picconia
Osmanthus fragrans
Osmanthus heterophyllus
Osmanthus delavayi
Osmanthus yunnanensis
Osmanthus decorus
Osmanthus sensu stricto
Chionanthus aff. brassii (JFM3078)
Chionanthus aff. brassii (JFM3081)
Chionanthus quadristamineus
Chionanthus axillaris
Chionanthus maxwellii
Chionanthus brachystachys
Chionanthus aff. ramiflorus (MEP24397)
Chionanthus aff. ramiflorus (JKM245)
Chionanthus ramiflorus
Olea tsoongii
Olea javanica
Olea subg. Tetrapilus
Chionanthus macrobotrys
Chionanthus polygamus
Chionanthus rupicolus
Chionanthus aff. thorelii
Chionanthus retusus
Cartrema americana (CRK R-1)
Cartrema americana (CHW-OA)
Cartrema
Chionanthus virginicus
Australasian Chionanthus
Chionanthus filiformis
Chionanthus fluminensis
Chionanthus trichotomus
Chionanthus implicatus
Neotropical Chionanthus
Chionanthus pubescens
Chionanthus panamensis
Haenianthus salicifolius
Core Chionanthus
Olea europaea subsp. laperrinei
Olea europaea subsp. europaea (SM11)
Olea europaea subsp. guanchica
Olea europaea subsp. europaea (VF5)
Olea europaea subsp. europaea (Oeiras1)
Olea europaea subsp. cuspidata
sect. Olea
Olea capensis subsp. capensis
Olea capensis subsp. macrocarpa
Olea exasperata
Olea lancea
Olea perrieri
sect. Ligustroides
subg. Olea
Olea sensu stricto
Olea paniculata
subg. Paniculatae
Noronhia intermedia
Noronhia clarinerva
Noronhia lowryi (RJQ550)
Noronhia lowryi (RJQ675-2)
Noronhia emarginata
Noronhia peglerae
Noronhia
Chionanthus parkinsonii
Chionanthus mala-elengi
Thai Chionanthus
Forestiera phillyreoides
Forestiera angustifolia
Forestiera segregata
Forestiera pubescens var. parviflora
Forestiera pubescens var. pubescens
Forestiera ligustrina
Forestiera isabelae
Forestiera
Hesperelaea palmeri
Chionanthus ligustrinus
Priogymnanthus hasslerianus
Fraxinus chinensis
Fraxinus bungeana
Fraxinus lanuginosa
Fraxinus sieboldiana
Fraxinus ornus (KEW-1969-14512)
Fraxinus ornus (Cefe-B2)
sect. Ornus
Fraxinus insularis
Fraxinus xanthoxyloides
sect. Sciadanthus
Fraxinus mandshurica
Fraxinus nigra
Fraxinus angustifolia
sect. Fraxinus
Fraxinus latifolia
Fraxinus velutina
Fraxinus americana
sect. Melioides
Fraxinus quadrangulata
sect. Dipetalae
Fraxinus
Ligustrum japonicum
Ligustrum gracile
Syringa yunnanensis
Syringa pubescens subsp. microphylla
Syringa vulgaris
Syringa + Ligustrum
Comoranthus minor
Schrebera capuronii
Schrebera alata
Schrebera arborea
Schrebera trichoclada
Schrebera swietenioides
Schrebera + Comoranthus
Forsythia mandschurica
0.003

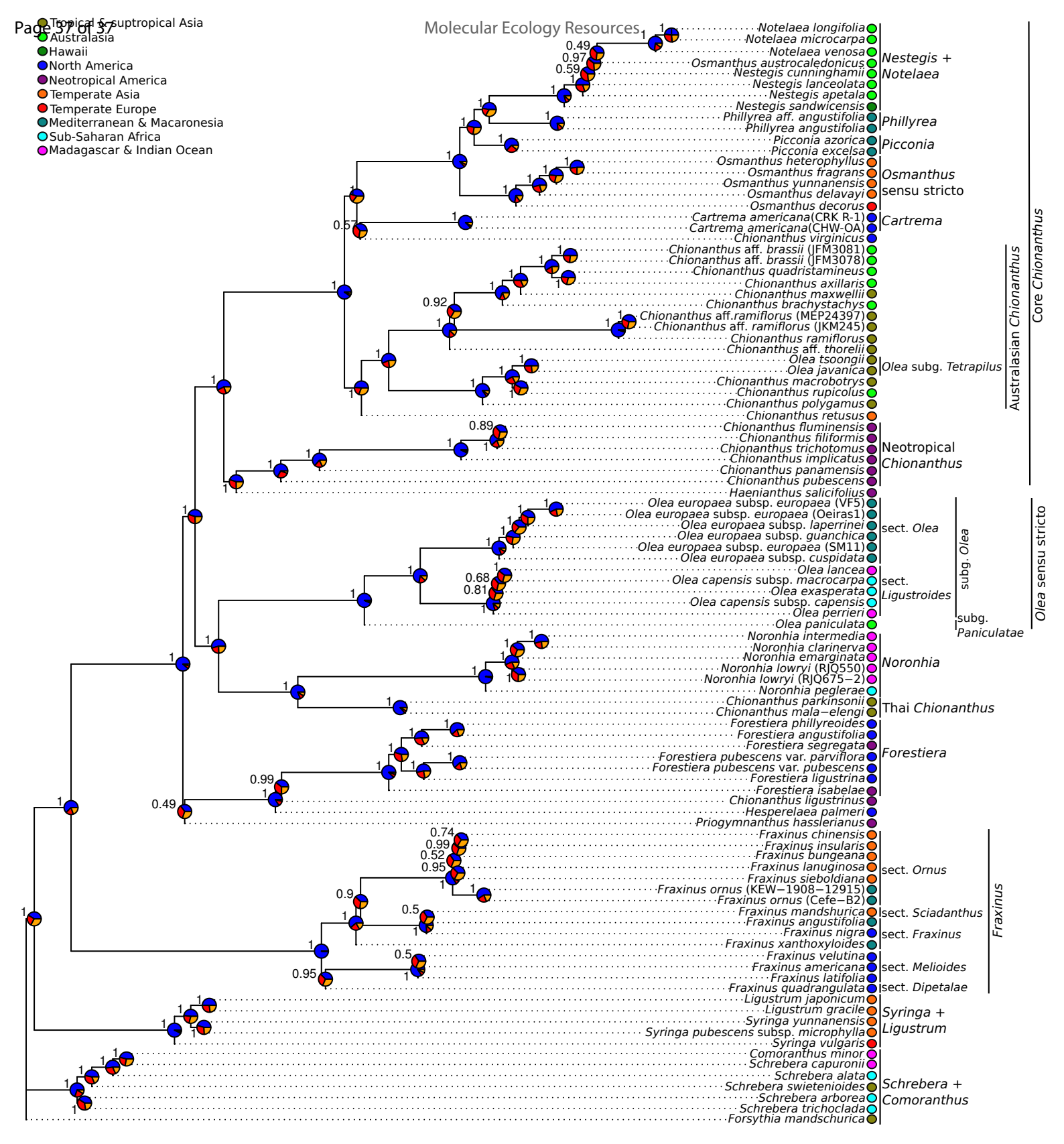

olive (Ole6) genome
ash (BATG−0.5) genome
Silica dried material
Herbarium specimen
% alignment success
0
20
40
60
80
100
0
10
20
30
40
50
60
Divergence to reference genome (MY)




| % missing data | Nr. reads |
|---|---|
| <10 | > 10 M |
| 10-20 | 8-10 M |
| 20-30 | 7-8 M |
| 30-40 | 6-7 M |
| 40-50 | 5-6 M |
| 50-60 | 4-5 M |
| 60-70 | 3-4 M |
| 70-80 | 2-3 M |
| 80-90 | 1-2 M |
| >90 | < 1 M |

| % missing data | No. reads |
|---|---|
| 7.2 | 10,526,465 |
| 5.9 | 9,092,062 |
| 21.6 | 7,664,720 |
| 44.4 | 3,561,894 |
| 30.4 | 4,868,071 |
| 45.5 | 3,069,103 |
| 15.6 | 7,196,115 |
| 16.4 | 6,098,934 |
| 41.9 | 4,036,067 |
| 11.1 | 9,461,143 |
| 20.3 | 6,039,200 |
| 13.4 | 7,770,885 |
| 24.2 | 5,340,876 |
| 11.9 | 7,775,106 |
| 10.9 | 7,136,072 |
| 12.4 | 7,703,343 |
| 20.1 | 6,920,815 |
| 28.2 | 4,302,243 |
| 49.6 | 6,439,190 |
| 45.5 | 6,224,472 |
| 54.0 | 5,944,158 |
| 29.9 | 6,459,054 |
| 43.1 | 6,480,964 |
| 61.5 | 6,139,529 |
| 11.4 | 11,551,595 |
| 8.7 | 17,654,720 |
| 18.1 | 10,698,816 |
| 22.9 | 8,903,815 |
| 8.5 | 14,865,409 |
| 24.3 | 7,451,990 |
| 21.1 | 6,061,957 |
| 26.0 | 11,946,047 |
| 8.7 | 15,563,731 |
| 30.0 | 11,536,626 |
| 37.4 | 3,882,458 |
| 25.8 | 6,376,918 |
| 25.0 | 5,451,409 |
| 3.6 | 14,878,651 |
| 11.2 | 13,017,388 |
| 19.5 | 6,418,857 |
| 80.8 | 4,439,982 |
| 27.0 | 7,880,699 |
| 84.7 | 1,524,530 |
| 0.9 | 20,885,954 |
| 0.4 | 29,025,914 |
| 49.1 | 4,586,111 |
| 0.3 | 44,896,338 |
| 27.1 | 5,366,264 |
| 51.7 | 5,369,778 |
| 47.1 | 4,802,971 |
| 17.2 | 13,282,830 |
| 8.1 | 18,460,844 |
| 39.2 | 4,260,300 |
| 41.9 | 7,320,843 |
| 47.3 | 6,751,237 |
| 23.7 | 5,409,714 |
| 14.7 | 6,990,805 |
| 21.0 | 5,201,803 |
| 52.3 | 3,700,201 |
| 19.5 | 6,210,891 |
| 30.2 | 5,615,616 |
| 14.1 | 12,912,166 |
| 15.5 | 8,902,262 |
| 21.3 | 9,313,535 |
| 2.8 | 14,428,252 |
| 32.3 | 7,217,842 |
| 23.9 | 6,639,747 |
| 12.5 | 6,507,570 |
| 21.3 | 7,632,289 |
| 44.5 | 7,629,853 |
| 19.1 | 8,551,700 |
| 36.4 | 10,116,007 |
| 93.1 | 9,032,605 |
| 26.7 | 5,393,415 |
| 28.1 | 5,267,476 |
| 20.4 | 6,760,458 |
| 34.2 | 4,410,328 |
| 31.5 | 5,152,897 |
| 28.3 | 5,476,623 |
| 8.1 | 8,403,940 |
| 46.6 | 2,562,927 |
| 20.3 | 5,909,491 |
| 24.5 | 5,424,724 |
| 16.3 | 7,510,336 |
| 24.0 | 5,944,491 |
| 18.0 | 7,018,544 |
| 18.5 | 7,623,040 |
| 19.3 | 5,393,568 |
| 52.2 | 5,462,260 |
| 47.1 | 7,492,984 |
| 49.2 | 5,362,357 |
| 17.8 | 9,887,751 |
| 41.5 | 5,769,957 |
| 17.5 | 8,217,218 |
| 7.1 | 12,659,254 |
| 45.6 | 5,057,664 |
| 78.2 | 5,169,909 |
| 63.4 | 10,832,082 |
| 9.3 | 22,211,793 |
| 27.7 | 8,283,607 |


A
No. genotyped SNPs ($10^6$)
$R^2$=0.424
p=1.313e−13
Divergence to reference (Mya)
B
No. genotyped SNPs ($10^6$)
$R^2$=0.283
p=8.796e−9
Divergence to reference (Mya)
C
% missing data
Olive
$R^2$=0.004
p=0.233
Ash
$R^2$=−0.001
p=0.355
Divergence to reference (Mya)
D
% missing data
$R^2$=0.186
p=4.592e-6
Estimated sequencing depth


Tropical & suptropical Asia
Australasia
Hawaii
North America
Neotropical America
Temperate Asia
Temperate Europe
Mediterranean & Macaronesia
Sub-Saharan Africa
Madagascar & Indian Ocean
Notelaea longifolia
Notelaea microcarpa
Notelaea venosa
Osmanthus austrocaledonicus
Nestegis cunninghamii
Nestegis lanceolata
Nestegis apetala
Nestegis sandwicensis
Nestegis + Notelaea
Phillyrea aff. angustifolia
Phillyrea angustifolia
Phillyrea
Picconia azorica
Picconia excelsa
Picconia
Osmanthus heterophyllus
Osmanthus fragrans
Osmanthus yunnanensis
Osmanthus delavayi
Osmanthus decorus
Osmanthus sensu stricto
Cartrema americana(CRK R-1)
Cartrema americana(CHW-OA)
Cartrema
Chionanthus virginicus
Chionanthus aff. brassii (JFM3081)
Chionanthus aff. brassii (JFM3078)
Chionanthus quadristamineus
Chionanthus axillaris
Chionanthus maxwellii
Chionanthus brachystachys
Chionanthus aff.ramiflorus (MEP24397)
Chionanthus aff. ramiflorus (JKM245)
Chionanthus ramiflorus
Chionanthus aff. thorelii
Olea tsoongii
Olea javanica
Olea subg. Tetrapilus
Chionanthus macrobotrys
Chionanthus rupicolus
Chionanthus polygamus
Chionanthus retusus
Australasian Chionanthus
Chionanthus fluminensis
Chionanthus filiformis
Chionanthus trichotomus
Chionanthus implicatus
Chionanthus panamensis
Chionanthus pubescens
Neotropical Chionanthus
Core Chionanthus
Haenianthus salicifolius
Olea europaea subsp. europaea (VF5)
Olea europaea subsp. europaea (Oeiras1)
Olea europaea subsp. laperrinei
Olea europaea subsp. guanchica
Olea europaea subsp. europaea (SM11)
Olea europaea subsp. cuspidata
sect. Olea
Olea lancea
Olea capensis subsp. macrocarpa
Olea exasperata
Olea capensis subsp. capensis
sect. Ligustroides
subg. Olea
Olea perrieri
Olea paniculata
subg. Paniculatae
Olea sensu stricto
Noronhia intermedia
Noronhia clarinerva
Noronhia emarginata
Noronhia lowryi (RJQ550)
Noronhia lowryi (RJQ675−2)
Noronhia peglerae
Noronhia
Chionanthus parkinsonii
Chionanthus mala−elengi
Thai Chionanthus
Forestiera phillyreoides
Forestiera angustifolia
Forestiera segregata
Forestiera pubescens var. parviflora
Forestiera pubescens var. pubescens
Forestiera ligustrina
Forestiera isabelae
Forestiera
Chionanthus ligustrinus
Hesperelaea palmeri
Priogymnanthus hasslerianus
Fraxinus chinensis
Fraxinus insularis
Fraxinus bungeana
Fraxinus lanuginosa
Fraxinus sieboldiana
Fraxinus ornus (KEW−1908−12915)
Fraxinus ornus (Cefe−B2)
sect. Ornus
Fraxinus mandshurica
sect. Sciadanthus
Fraxinus angustifolia
Fraxinus nigra
sect. Fraxinus
Fraxinus xanthoxyloides
Fraxinus velutina
Fraxinus americana
sect. Melioides
Fraxinus latifolia
Fraxinus quadrangulata
sect. Dipetalae
Fraxinus
Ligustrum japonicum
Ligustrum gracile
Syringa yunnanensis
Syringa pubescens subsp. microphylla
Syringa vulgaris
Syringa + Ligustrum
Comoranthus minor
Schrebera capuronii
Schrebera alata
Schrebera swietenioides
Schrebera arborea
Schrebera trichoclada
Schrebera + Comoranthus
Forsythia mandschurica