Molecular Ecology Resources

WILEY

MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE OPEN ACCESS

Variant Calling in the Goldilocks Zone: How Reference Genome Choice and Read Mapping Stringency Impact Heterozygosity Estimates and Phylogenetic Analyses

Rebekah A. Mohn[1,2] | Mira Garner[3] | Paul S. Manos[4] | Andrew L. Hipp[1,5]

[1]Science and Conservation, The Morton Arboretum, Lisle, Illinois, USA | [2]Department of Plant and Microbial Biology, University of Minnesota, St. Paul, Minnesota, USA | [3]Pritzker Laboratory for Molecular Systematics and Evolution, The Field Museum, Chicago, Illinois, USA | [4]Department of Biology, Duke University, Durham, North Carolina, USA | [5]Negaunee Integrative Research Center, The Field Museum, Chicago, Illinois, USA

Correspondence: Rebekah A. Mohn (rebekah.mohn@gmail.com)

Received: 6 January 2025 | **Revised:** 22 October 2025 | **Accepted:** 29 October 2025

Handling Editor: Alison Gonçalves Nazareno

Funding: This work was supported by the National Science Foundation under award (DEB) Division of Environmental Biology (2129237, 2129281).

Keywords: Bowtie 2 | BWA | *Quercus* | reference mapping | SNP-calling | WGS

ABSTRACT

The increasing numbers of published reference genomes and affordability of whole genome resequencing have enabled multispecies population genomic and phylogenomic studies on non-model organisms, but they raise a new question for comparative genomics: what reference genome and mapping method combination results in the most data with the least bias? We mapped short-read resequencing data from seven eastern North American white oak (*Quercus* sect. *Quercus*) and two related samples to four *Quercus* reference genomes (*Q. alba, Q. lobata, Q. mongolica,* and *Q. rubra*) which represent different degrees of phylogenetic relatedness to the samples. We used three different mapping methods: a global (Bowtie 2 --end-to-end) and two local (Bowtie 2 --local and BWA-MEM) alignment approaches. For the twelve resulting datasets, we analysed read mapping accuracy and efficiency, missing data, heterozygosity, and inferred phylogenies to evaluate the impact of reference genome and read-mapping method. We found that the genetic distance of the reference genome to the samples and mapping method together impacted heterozygosity and phylogenetic tree estimation. There were two notable effects. First, when using a global alignment method (Bowtie 2 --end-to-end), estimated heterozygosity negligibly decreased with increased genetic distance between the reference and sample. Second, the most distantly related reference genome had significantly reduced base pair recovery and resulted in under- or overestimating heterozygosity depending on the method, and a more unbalanced phylogeny. We conclude that using a closely related but not conspecific reference is ideal to minimise reference bias and using Bowtie 2 --end-to-end minimises mismapping, resulting in the most accurate variant calls.

1 | Introduction

As next-generation sequencing has become more affordable, increasing numbers of studies are using multi-species whole genome sequencing datasets to address questions in population genomics and phylogenomics. At the same time, increasing numbers of reference genomes and the numerous methods available for read-mapping require researchers to decide which to use. As mapping methods impact the number of mismatches or indels allowed between the sample reads and the reference genome (Canzar and Salzberg 2017; Wilton and Szalay 2022), and genetic distance between a sample and the reference genome to

which it is mapped directly influences the number of sequence mismatches and indels, the choices of mapping method and reference genome should be considered together. Comparative genomic studies, which sample multiple species of potentially varying genetic or phylogenetic distance to any single reference genome, make the choice of a reference and the appropriate mapping method even more complex than single-species studies.

Previous papers exploring the impact of reference genomes on subsequent analyses in genome resequencing data have found that reference genome choice impacts both phylogenetic inference and population genetic statistics. Rick et al. (2023) found that reference genome selection (ingroup vs. outgroup) was correlated with incorrect tree inference and increased tree imbalance in some scenarios, with imbalance and inaccuracy increasing under stringent minor allele count filtering. Studies looking at the impact of a reference on population genetic inferences have concluded that a reference that is more closely related to the study samples is preferable to a more genetically distant reference (Galla et al. 2019; Thorburn et al. 2023) but that a conspecific reference is not necessary for accurate estimations of population genetic statistics (Galla et al. 2019). Including multiple species in a single study may complicate reference genome selection, as variation in genetic distance between the reference and the study species may favour mapping in closely related species by increasing sequence recovery. Alternatives to selecting a single reference have been proposed, such as *de novo* assembly, pangenome references, and cross-species scaffolding assembly (Prasad et al. 2022). However, currently, these methods result in less accurate mapping (Manel et al. 2016), slower analyses (Eizenga et al. 2020), and more fragmented mapping scaffolds (Prasad et al. 2022), respectively.

Studies published to date have found that using minor allele frequency and missing data to filter mapped reads and SNPs can impact phylogenetic inference and population genetic statistics (Huang and Knowles 2016; Linck and Battey 2019; Rick et al. 2023; Shafer et al. 2017). A RAD-seq study investigating the effects of different reference mapping and SNP-filtering methods on population statistics was unable to identify any one combination that best matched prior expectations (Shafer et al. 2017). We are unaware, however, of studies evaluating how reference genome and read mapping methods jointly impact our ability to address multispecies population and phylogenetic questions with whole genome sequencing. There is reason to expect that they might. Mapping methods differ, for example, in relative weights for match, mismatch, gap, and gap length, which together determine whether a mismatch or gap is favoured in read mapping (Canzar and Salzberg 2017; Wilton and Szalay 2022). This then impacts which genomic regions are recovered, potentially favouring variants or insertions/deletions. Additionally, methods differ in whether read mapping is based on the whole read or, alternatively, only the sections of each read that map successfully, ignoring ends of each read that do not map. The former approach, global alignment, requires that each sequence match the reference end-to-end (Langmead and Salzberg 2012), making it more stringent than a local alignment. The latter approach, local alignment, which requires only part of the read to match the reference, might have the advantage of allowing more of the sample reads to map to a reference especially in the presence of structural variance than in the global alignment. However, local alignment may come at the cost of accuracy. At the same time, the distance between the sample and the reference will affect the relative frequency of gaps or mismatches inferred in a sample.

To maximise the number of reads mapped to the reference and the accuracy of the mapping, a species identical to one of the species studied might be ideal. However, this may bias the read mapping and base pair calling toward the reference species, whose reads would map best. Alternatively, a reference species equally distantly related to all species in the sample would ensure the mapping is not biased, but may decrease mapping accuracy, outweighing some of the benefits of reduced bias. The third option is a species not included in the studied group but more closely related to some studied species than others, which might balance the risk of bias against the benefits of read-mapping accuracy. If the reference genome produces a substantial bias, it may influence the interpretation of analyses.

The concerns about genome reference and mapping stringency are especially relevant when studying groups like the eastern North American (ENA) white oaks, in which population genetics, species delimitation, hybridisation, and phylogenetic history cannot easily be considered separately from each other. With a most recent common ancestor ca. 36 Ma ago, the 16 species of the ENA white oaks are a largely sympatric but polyphyletic group within the global white oak clade (*Quercus* sect. *Quercus*). Nested among them are clades from Europe, Asia, and Mexico (Hipp et al. 2020). In spite of the phylogenetic distance separating the ENA white oaks, they interact ecologically, and many of them readily hybridise. Even the California valley oak, *Quercus lobata* (also *Q.* sect. *Quercus*), exhibits evidence of a history of introgression with *Q. macrocarpa*, one of the ENA species (McVay et al. 2017). However, the directionality of the introgression is highly uncertain (McVay et al. 2017). The biological and genetic complications presented by the co-occurring, interfertile ENA white oak species require addressing the analytical challenges of producing comparative genomic datasets that are accurate and unbiased despite the 36 Ma divergence.

In this study, we investigate the effects of mapping sequence reads from ENA white oak samples to four different references: (1) northern red oak, *Q. rubra*, which as a member of *Q.* sect. *Lobatae* diverged from all white oaks (*Q.* sect. *Quercus*) ca. 48–54 Ma and does not hybridise with them; (2) California valley oak, *Q. lobata* (*Q.* subsect. *Dumosae, Q.* sect. *Quercus*), which diverged an estimated 45 Ma from the ENA white oak lineages (Hipp et al. 2020) but exhibits ancient introgression with the ENA white oak *Q. macrocarpa*; (3) the east Asian Mongolian oak, *Q. mongolica* (a Eurasian 'roburoid' oak *sensu* Denk and Grimm 2010; *Q.* sect. *Quercus*), which diverged from the ENA white oaks of *Q.* subsect. *Albae* (*sensu* Manos and Hipp, 2021) approximately 19 Ma (Hipp et al. 2020); and (4) eastern white oak, *Q. alba* (*Q.* subsect. *Albae*, Q. sect. *Quercus*), one of the ENA species. These reference genomes provide a range of genetic distances and introgression histories to test the impact of distance between the reference genome and the mapped samples on the diversity statistics and phylogenomics.

We also investigate the effects of mapping methods. While a few previous studies have explored questions concerning reference choice (Galla et al. 2019; Rick et al. 2023; Thorburn et al. 2023), each used a single mapping method, raising the question of what impact the method has in relation to genome choice. Bowtie 2 and

TABLE 1 | Differences in default scoring weights between Bowtie 2 --end-to-end, Bowtie 2 --local, and BWA-MEM.

Method	Bowtie 2 --end-to-end	Bowtie 2 --local	BWA-MEM
Global vs. local	global	local	local
Match	+0	+2	+1
Mismatch	−6	−6	−4
Gap open	−5	−5	−6
Gap extension	−3	−3	−1
Clipping penalty	NA	0	−5 total
Source	bowtie-bio.sourceforge.net/bowtie2/manual.shtml	bowtie-bio.sourceforge.net/bowtie2/manual.shtml	bio-bwa.sourceforge.net/bwa.shtml

Note: Global alignments align the whole read while local alignments clip the read ends if it improves the alignment score.

TABLE 2 | Hypotheses of the impact of reference and methods on assembly statistics, heterozygosity, and phylogenetics.

	Assembly statistics	Heterozygosity	Phylogenetics
Reference	**Assembly Stats by Reference** More reads will map and more sites will be called with less error as sample/reference genetic distance decreases.	**Heterozygosity by Reference** Increased number of sites will result in increased heterozygosity as sample/reference genetic distance decreases.	**Phylogenetics by Reference** Increased number of sites will result in longer branch lengths as sample/reference genetic distance decreases.
Method	**Assembly Stats by Method** More reads will assemble with local alignment methods (BWA-MEM and Bowtie 2—local), but this will come at the cost of accuracy.	**Heterozygosity by Method** Increased number of sites and increased rate of error for local alignment methods will result in higher heterozygosity.	**Phylogenetics by Method** Increased number of SNPs with a local alignment method will result in longer branch lengths.
Reference+Method	**Assembly Stats by Reference + Method** More reads will map with local alignment and a close reference; fewer reads will map with a distant reference and global alignment.	**Heterozygosity by Reference + Method** More stringent mapping methods will result in larger decrease in heterozygosity with reference difference.	**Phylogenetics by Reference + Method** When mapping method is stringent the branch length will decrease with reference distance, but when mapping method is relaxed branch length will increase due to mismapping.

BWA-MEM use the Burrows-Wheeler transform to align short reads to a reference genome, but they differ in search approaches and scoring weights (for match versus mismatch and insertions, for example) to rapidly identify the best mapping (Table 1; Langmead and Salzberg 2012; Li and Durbin 2009; Wilton and Szalay 2022). Within Bowtie 2 there are two options for stringency: the default global alignment method, which allows for gaps but maps each sequence read end-to-end against the reference genome; and a local alignment method, which requires only part of each sequence read to map to the reference and trims the remaining ends. BWA-MEM is also a local alignment method. This allows us to explore two different scoring algorithms (Bowtie 2 vs. BWA-MEM) and two different levels of stringency (local vs. global).

We expect that increased genetic distance between the reference genome and sample will result in lower sequence recovery, due to difficulties in mapping, and lower estimates of heterozygosity with stringent mapping methods, due to reduced power to recover both alleles when a distant reference is used (see hypotheses and expectations in Table 2). We also expect that the local (either BWA-MEM or Bowtie 2 --local) mapping methods will counteract this bias, as local methods reduce the stringency of mapping and thus should provide more effective mapping to distant references.

2 | Methods

2.1 | Samples

We chose a sample to represent each of four widespread ENA white oak species: *Q. stellata*, *Q. muehlenbergii*, *Q. alba*, and *Q. macrocarpa*. Because of the history of introgression between *Q. gambelii* and *Q. macrocarpa* in South Dakota (Maze 1968) and the hybridisation history of some *Q. macrocarpa* with both *Q. gambelii* and *Q. lobata* (Crowl et al. 2020; McVay et al. 2017), we included three *Q. macrocarpa* samples: a northern sample (Bayfield

County, Wisconsin, USA), a southern sample (Fort Bend County, Texas, USA), and a South Dakota sample (Lawrence County, South Dakota, USA), as well as a *Q. lobata* (SRR14632932) sample. We also included an individual that was genotyped as ~75% *Q. bicolor* and ~25% *Q. alba* (Mohn et al. unpublished) to assess how a known backcross sample would behave. As an outgroup, we included a *Q. rubra* (SRR24768723) sample.

For newly generated samples (all except *Q. lobata* and *Q. rubra*, which came from the NCBI SRA), DNA was extracted from frozen shade leaves of mature trees. Prior to extraction, leaves were surface-sterilised to remove surface fungi and bacteria (protocol in the Supporting Information S1). Leaf material was submerged in 75% ethanol for 1 min, 3.25% sodium hypochlorite for 3 min, and 75% ethanol for 30 s to sterilise the surface. Leaves were then rinsed with sterile deionised water three times, changing the water in the baths between samples. After sterilisation, samples were stored in a −80 degree freezer until extraction. Samples were homogenised using a FastPrep machine with stainless steel beads with an extraction buffer and extracted with the Qiagen DNeasy kit (modified protocol based on Hipp and Weber 2008; protocol in the Supporting Information S2).

The Sequencing and Genomics Technologies Core Facility (SGT) at Duke University built sequencing libraries from submitted DNA using the Illumina DNA Prep Kit (Catalogue #20060059) according to the manufacturer's instructions. Libraries were pooled at an equimolar ratio and sequenced on two Illumina NovaSeq X Plus 10B lanes to produce 150 bp PE reads.

2.2 | Read Cleaning

Poly-X tails were trimmed and a sliding window of 15 base pairs on both the 3′ and 5′ prime ends was used to remove low-quality ends in FastP version 0.23.4 (Chen 2023). Reads with less than 30 bp remaining were then removed. While mapping to a reference genome should be adequate to filter out mitochondrial and chloroplast reads, chloroplast reads were filtered and retained as we intended to analyse them separately for research external to this project. This was done by mapping each sample's reads to chloroplast references with Bowtie 2 (v2.4.4-rhe18; Langmead and Salzberg 2012) using the Phylogenomic Dataset Construction script filter_organelle_reads.py (Morales-Briones et al. 2021; Yang and Smith 2014) with five white oak chloroplasts as the references (NC_081467.1, MK105457.1, MK105459.1, OQ835463.1, MK105467.1). Reads that failed to map concordantly to the chloroplasts were utilised for downstream analyses in the current study. As mitochondrial reads are filtered out by mapping to the reference genome, and as they are not a focus of phylogeographic or population study in oaks, we did not filter them out separately.

2.3 | References, Mapping Reads, and Calling SNPs

After the cleaning, each sample was mapped to the four reference genomes: *Quercus rubra* (JGI V2.0; Kapoor et al. 2023), *Q. lobata* (NCBI GCA_001633185), *Q. mongolica* (https://doi.org/10.6084/m9.figshare.11888118.v2; Ai et al. 2022), and *Q. alba* (version 4.3 haplotype A; Larson et al. 2025). Each reference represents different genetic proximities to ENA white oak species. *Quercus rubra* is an outgroup, equally distantly related to ingroup samples. *Quercus alba* is the same species as one of the ENA ingroup samples included in our study. *Quercus lobata* is a Californian species, but has evidence of introgression with ENA *Q. macrocarpa* (McVay et al. 2017). Eurasian white oaks (the 'roburoid' clade of *Q.* sect. *Quercus*) are closely related to ENA white oaks with no recent history of introgression, but are more closely related to *Q. alba* than the other ENA samples (Hipp et al. 2020). Of the available Eurasian white oak genomes, *Q. mongolica* was chosen as it is more syntenic and has higher gene recovery than the *Q. robur* genome (Fig S8 in Ai et al. 2022).

Mapping was performed with three different methods representing two different read trimming approaches and two different alignment scoring weights for mismatch and gap alignment (Table 1). Each combination of sample and reference was mapped three times: once with the Bowtie 2 (version2.4.4-rhe18; Langmead and Salzberg 2012; Langmead et al. 2019) global alignment default settings (--end-to-end --no-unal), once with the Bowtie 2 local alignment default settings (--local --no-unal) and once with the default settings of BWA-MEM (Li 2013). These methods are not exhaustive but represent widely used tools and read mapping approaches that can provide insight into the impact of read mapping methods on research findings. The files were then converted to bam files and sorted using Samtools 1.14-rhe18 (Danecek et al. 2021).

bcftools mpileup and bcftools call (v. 1.4; Danecek et al. 2021) with a read depth cut-off of 4 (−g 4) and the multi-allele caller (−m) were used to call variants. Variants were called separately for each of the 12 combinations of four references and three alignment methods. Within each combination, the variants of all samples were called simultaneously, yielding a single gVCF file for each combination.

2.4 | Statistics

2.4.1 | Assembly Statistics

After mapping, we collected the following output summary statistics from Bowtie 2: total number of reads, number of reads mapped once concordantly, and percent of reads mapped. To enable the comparison of mapping between Bowtie 2 and BWA-MEM assemblies, we calculated mapping statistics on the BAM files using Samtools Stats (version 1.14-rhel8; Danecek et al. 2021). For the total number of reads mapping and the number of reads mapping concordantly, the Samtools Stats results were comparable to the Bowtie 2 results. However, the MQ = 0 value calculated from Samtools Stats was much smaller and showed a different pattern than the multimapping value from Bowtie 2. This is likely due to the fact that the MQ value is dependent on how each assembly software calculates the probability of correct mapping. Therefore, we compared only the multimapping statistic within Bowtie 2.

2.4.2 | VCF Statistics

We also calculated sample statistics from the VCF file. VCFtools (Danecek et al. 2011) first filtered SNPs to those with a max of

2 alleles and 75% or more of the samples present and then calculated heterozygosity, depth, and missing data for each sample. The VCFtools heterozygosity utilises only the SNP dataset which consists of sites that are variant in or between any sample or between the samples and the reference. The output for each sample includes two numbers: the number of sites in the SNP dataset where the sample was not missing (NSITES) and the number of sites in the SNP dataset that are homozygous in the sample (O(HOM)). From this, we calculated heterozygosity as NSITES minus O(HOM) together divided by the total number of recovered sites in the sample after filtering.

For assembly- and VCF-derived statistics, the relatedness between species was classified as "conspecific" if the species of the sample matched the species of the reference, "close" if the reference and the sample were both from *Q.* subg. *Quercus*, or "distant" if either the reference or the sample was from different subgenera. For each statistic, Tukey's test was performed on an ANOVA to summarise differences among analyses that vary in phylogenetic distance to the reference genome and in read-mapping method. While Tukey's test—like ANOVA—is sensitive to the samples included in the analysis and the size of each sample, it provides a heuristic by which to test the relative impact of the reference and mapping method on our results. The ANOVA and Tukey's tests were performed in R (version 4.4.1; R Core Team 2018) and visualised with ggplot2.

2.4.3 | Phylogenetics

The VCF files were converted into fasta format, requiring a minimum of 4 taxa per aligned nucleotide position, using vcf-2phylip.py (version 2.9; Ortiz 2019). IQ-TREE2 (version 2.2.2.7; Nguyen et al. 2015) was used to estimate a phylogeny using a GTR + G model and 1000 rapid bootstraps were calculated. Tip-to-crown node distance was calculated for each *Q.* sect. *Quercus* sample on each tree.

3 | Results

3.1 | Sequencing Summary Statistics

Newly-sequenced ENA white oak samples had between 49.1 and 79.7 million paired reads. *Quercus lobata* had 147.4 million 100-bp paired reads and *Q. rubra* had 87.3 million 150-bp paired reads. After quality trimming was performed and the chloroplast reads were removed, between 46 and 76 million reads remained for each ENA white oak sample, 136 million for *Q. lobata*, and 84 million reads for *Q. rubra*. These non-chloroplast reads were mapped to the reference genomes with Bowtie 2 and BWA.

3.2 | Mapping method and distance to the reference affected read mapping but not sequence recovery

The highest percent of reads mapped using BWA-MEM and conspecific reference/sample pairs (99.1%), but this was not significantly different from BWA-MEM with close reference/sample pairs (98.7%) or Bowtie 2 --local with close or conspecific reference/sample pairs (96.7% and 98.1%, respectively). The lowest percent of reads mapped with Bowtie 2 --end-to-end and distant reference/sample pairs (56.7%). Ranging from 92.3% to 56.7%, Bowtie 2 --end-to-end had significantly lower read mapping percentages than either Bowtie 2 --local or BWA-MEM at each reference/sample distance (Figure 1). As distance between sample and reference increased, the percent of reads mapped by Bowtie 2 --end-to-end decreased significantly (Tukey's test, based on ANOVA $p < 0.05$) from 92.3% to 56.7% of reads mapping. The proportions of reads mapping for Bowtie 2 --local and BWA-MEM were highest and not significantly different in conspecific sample/reference pairs (98.1% and 99.1% respectively) or between conspecific and closely related sample/reference pairs (98.1% vs. 96.7% Bowtie 2 --local and 99.1% vs. 98.7% BWA-MEM). However, when mapping to close or distant references, BWA-MEM (98.7% and 94.4% respectively) mapped significantly more reads than Bowtie 2 --local (96.7% and 87.1% respectively). A significant decrease in the percent of reads mapping was also observed in conspecific and close versus distant reference/sample pairs for Bowtie 2 --local (96.7% vs. 87.1%) and BWA-MEM (98.7% vs. 94.4%).

Bowtie 2 --end-to-end mapped the smallest number of reads across all analyses (Figure 1a) but did not result in fewer base pairs recovered per sample after SNPs with more than 25% of samples missing were filtered (Figure 2b). Distantly related sample/reference pairs had a significantly lower number of base pairs recovered (242–291 Mbp) than closely related and conspecific sample/reference pairs (405–415 Mbp and 366–378 Mbp respectively; Figure 2b). The *Q. rubra* sample, the outgroup among our samples, had significantly fewer base pairs in the combined VCF file and significantly more missing SNPs when mapped to a distant reference (27.4%–42.0%). When mapped back to the *Q. rubra* reference, the *Q. rubra* sample, as the only conspecific or closely related reference, had fewer base pairs recovered than white oak samples mapped back to their conspecific references. Since we filtered SNPs to those shared by 75% of individuals, SNPs recovered exclusively by *Q. rubra* mapped to the *Q. rubra* reference would have been filtered out.

For conspecific and closely related reference genomes, the increase in mapping reads due to local read mapping is similar in number to the increase in reads mapping to multiple locations.

When we use a conspecific reference, the total number of reads mapping increases while the number of reads mapping uniquely decreases. The percent of reads mapping was higher by 5.8% in Bowtie 2 --local than in Bowtie 2 --end-to-end, but the percent of reads mapping concordantly and uniquely was significantly lower (by 14.0%) (Figure 1). Similarly, in closely related sample/reference pairs, the percent concordantly and uniquely mapped reads was 10.2% higher in Bowtie 2 --end-to-end than Bowtie 2 --local while Bowtie 2 --end-to-end mapped 12.0% fewer reads (Figure 1). In distantly related samples, the concordant and uniquely mapped reads were not significantly different between Bowtie 2 methods (Figure 1).

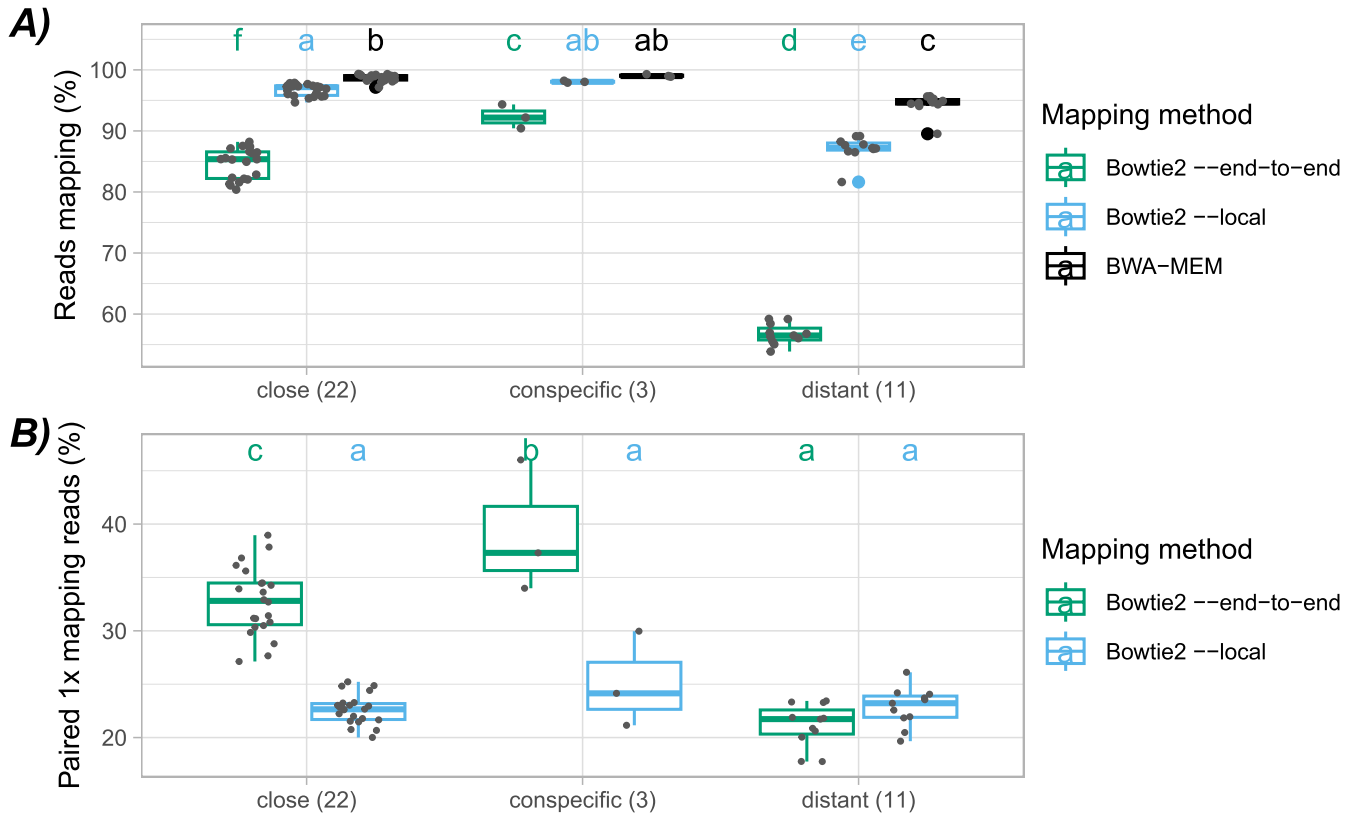

FIGURE 1 | (A) The percent reads mapping and (B) percent concordantly uniquely mapping reads of all the reads. These statistics were calculated from the raw BAM files and grouped by method (colour) and distance between reference and sample (x-axis). Sample size of each box plot in brackets.

3.3 | Mapping Method Differentially Influences Heterozygosity as Distance Between Reference Genome and Samples Increased

All methods mapped fewer reads as the distance between the reference and sample increased; heterozygosity across the whole genome, however, significantly increased with Bowtie 2 --local and BWA-MEM but slightly (not significantly) decreased with the more stringent Bowtie 2 --end-to-end mapping as the distance between the reference and sample increased (Figure 2c). The mean percent heterozygosity of samples mapped to a conspecific reference was lower in Bowtie 2 --end-to-end than Bowtie 2 --local or BWA (Bowtie 2 --end-to-end: 0.93% heterozygous sites, Bowtie 2 --local: 1.81%, BWA: 2.27%). Additionally, with Bowtie 2 -end-to-end, percent heterozygosity was not significantly different between conspecific sample/reference pairs (0.93%), closely related sample/reference pairs (0.90%), or distant sample/reference pairs (0.56%; Figure 2). Unlike Bowtie 2 -end-to-end, for Bowtie 2 --local and BWA-MEM percent heterozygosity was significantly higher in the distant (difference of 1.15% and 3.19%, respectively) and closely related (difference 0.46%, which was not significant, and 1.11%, respectively) references relative to the conspecific reference (Figure 2c). Analysis using BWA-MEM also yielded significantly higher percent heterozygosity than Bowtie 2 --local.

3.4 | Tree Topology and Branch Length are Impacted by the Reference Genome

Branch lengths varied between analyses conducted using different references. In all trees with a conspecific reference/species pair present—which is all analyses except the ones using *Q. mongolica* as a reference—the sample that was conspecific with the reference genome had an increased branch length (Figure 3). The tip-to-crown distance of trees inferred using either closely related or distantly related references was significantly higher for BWA and Bowtie 2 --local than for Bowtie 2 --end-to-end (Figure 4). While Bowtie 2 --local had significantly different tip-to-crown distances depending on the phylogenetic distance between the reference and sample, BWA showed no significant difference in distance from crown to tip for conspecific versus close versus distantly related reference/sample pairs (Figure 4).

In phylogenetic analysis of datasets generated using Bowtie 2 --end-to-end, the topology also varied with the reference. In all trees regardless of reference, *Q. macrocarpa* specimens were monophyletic, with *Q. lobata* sister to the clade of *Q. macrocarpa* samples. All other relationships varied. The phylogenies based on *Q. alba* and *Q. mongolica* references both had *Q. stellata* sister to *Q. muehlenbergii*, *Q. alba* sister to the *Q. bicolor—Q. alba* backcross, and these two clades (*Q. stellata and Q. muehlenbergii*; *Q. alba* and the *Q. bicolor—Q. alba* backcross clades) sister to the *Q. lobata* and *Q. macrocarpa* clade. By contrast, phylogenies based on the *Q. rubra* and *Q. lobata* references both had the *Q. bicolor—Q. alba* backcross sister to *Q. lobata* and *Q. macrocarpa*. However, in the phylogeny based on the *Q. lobata* reference, *Q. alba* was sister to *Q. stellata*, and these two together were sister to the remaining white oak samples. The phylogeny based on the *Q. rubra* reference had *Q. stellata*, *Q. muehlenbergii*, *Q. alba*, and the *Q. bicolor—Q. alba* backcross paraphyletic to *Q. lobata* and *Q. macrocarpa* (which were sister, as in all trees). For each reference, the topologies did not change between using Bowtie

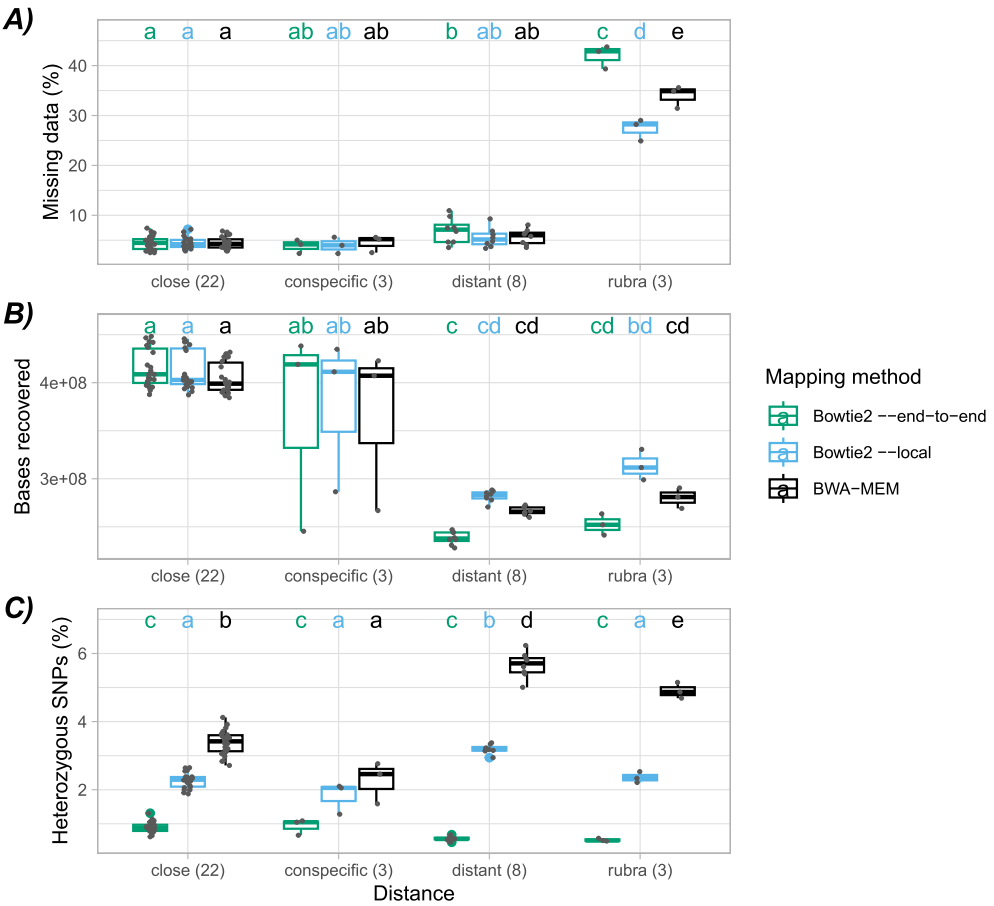

FIGURE 2 | (A) The percent missing data, (B) the number of bases recovered after filtering in each sample, and (C) the heterozygous SNPs divided by the number of bases recovered calculated by VCFtools from the VCF file. These are grouped by mapping method (colour) and distance to reference (x-axis). The *Q. rubra* samples mapped to a distant reference were graphed separately from other distant references as the filtering disproportionately impacted them (see Results and Discussion). Sample size of each box plot in brackets.

2 --end-to-end and Bowtie 2 --local except for the phylogenies based on the *Q. rubra* reference, in which the positions of *Q. alba* and *Q. muehlenbergii* swapped between the two mapping methods.

BWA read mapping with all *Q.* sect. *Quercus* references (i.e., all except *Q. rubra*) produced the same phylogenetic topology as the Bowtie 2 read mapping to *Q. alba* and *Q. mongolica*, with one exception: using BWA, *Q. macrocarpa* (SD) was sister to *Q. macrocarpa* (S) instead of sister to both *Q. macrocarpa* (S) and (N). The phylogeny based on the *Q. rubra* reference and BWA mapping was similar to the Bowtie 2 --end-to-end topology with *Q. rubra* as a reference, with a similar exception: *Q. macrocarpa* (SD) was sister to *Q. macrocarpa* (S) instead of sister to both *Q. macrocarpa* (S) and (N), and *Q. alba* and *Q. muehlenbergii* switched positions. Additionally, the bootstrap was low (13) for the position of *Q. alba*.

4 | Discussion

Multispecies analyses of any kind based on genome resequencing data present a comparative genomic challenge: there are often several potential reference genomes to use for read mapping and SNP calling, and any single reference may be conspecific, closely related, or distantly related to different samples in the study. This gradient of genetic similarity between reference and samples introduces the potential for read-mapping biases and thus raises questions about which reference will introduce the least bias and which read mapping method will minimise any bias that cannot be eliminated by reference choice. Previous studies comparing mapping methods have considered mapping speed and accuracy based only on samples or simulated data very similar to the reference (Keel and Snelling 2018; Langmead and Salzberg 2012; Li and Durbin 2009), or just acknowledged the relatedness of mapping methods on the genetic proximity of references and sample data (Canzar and Salzberg 2017). Our study demonstrates that, in fact, the decision on how to call variants for comparative genomic studies is more complicated than simply identifying which method or reference maps the most reads: in our study, the widely used BWA-MEM consistently mapped more reads than either Bowtie 2 --local or –end-to-end but yielded datasets with increasing heterozygosity as the genetic distance between sample and reference increased. Surprisingly, the high heterozygosity did not correlate with branch length bias in phylogenetic analyses. We found the most stringent method—Bowtie 2 --end-to-end—resulted in minimal reference bias in heterozygosity but mapped far fewer reads than the less stringent local method. With a reference that is equally

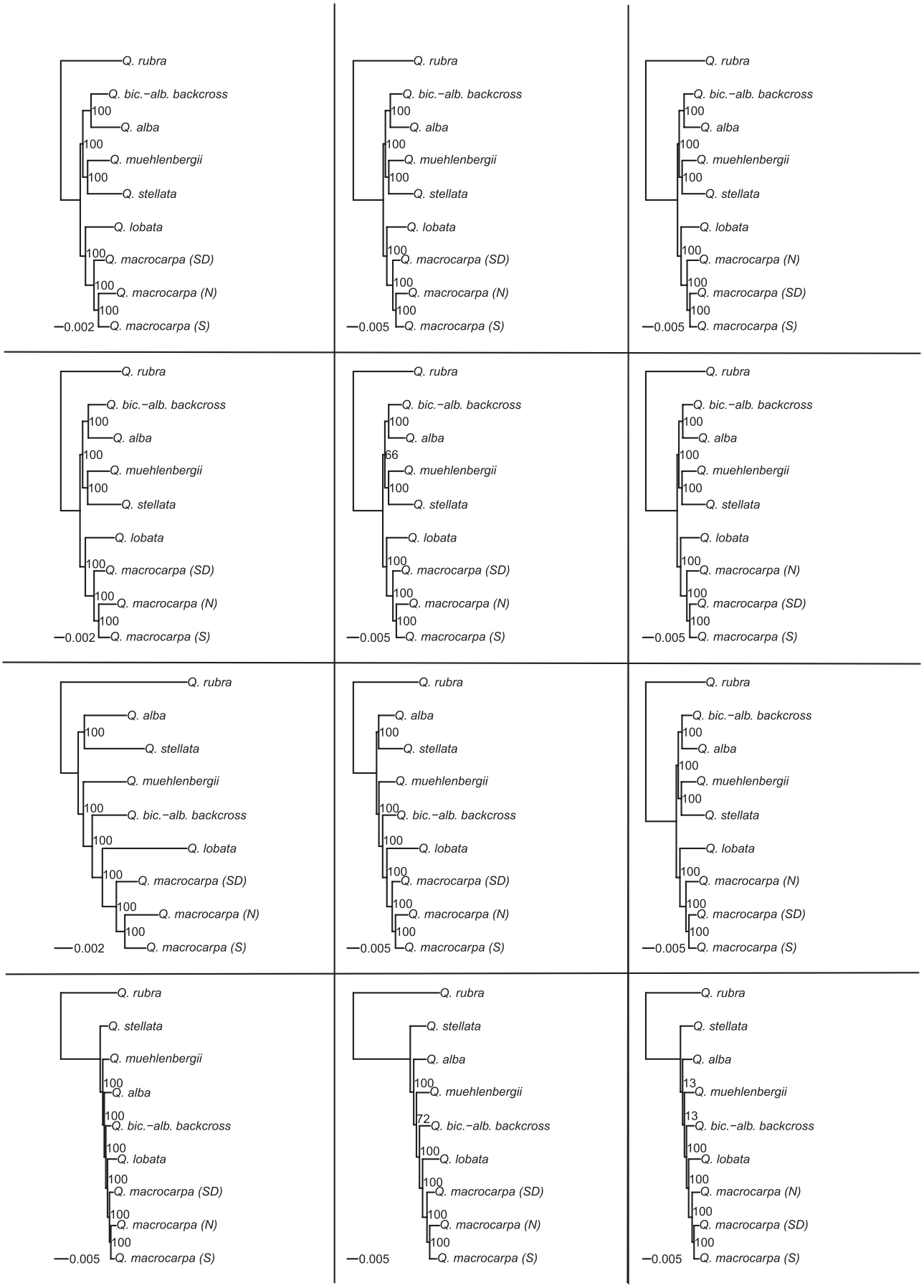

FIGURE 3 | Legend on next page.

FIGURE 3 | Phylogenies vary in branch length and topology based on different mapping methods and different references. Scale bars represent branch lengths in modelled substitutions per nucleotide. Abbreviations for *Q. macrocarpa* populations: N = northern; S = southern; SD = South Dakota (Black Hills).

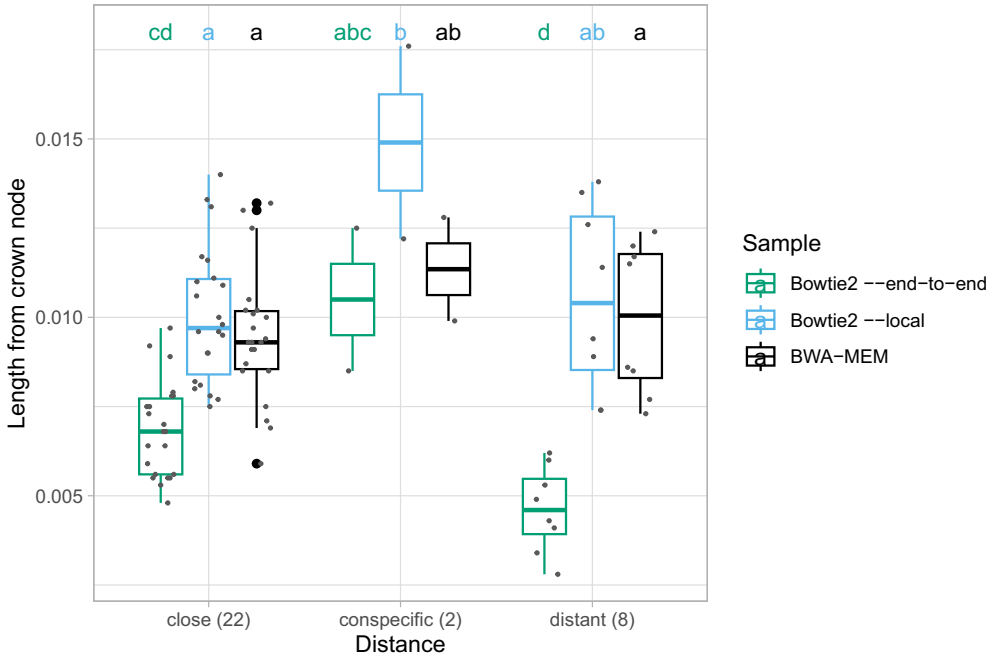

FIGURE 4 | The tip-to-crown distance fluctuates based on reference. Sample size of each box plot in brackets.

distantly related to all samples—*Quercus rubra* in our study, which is not within the ingroup *Q.* sect. *Quercus* clade—the number of base pairs recovered dropped by over a quarter across all methods, the heterozygosity decreased slightly (not significantly) with Bowtie 2 --end-to-end, and the tree topology became very imbalanced. A decrease in heterozygosity is expected if mapping stringency comes at the cost of the total number of reads mapped. Thus, comparative genomic studies require a joint consideration of reference genome and read mapping method.

While we might have anticipated the interaction of mapping stringency and reference on heterozygosity—that more stringent methods might result in lower heterozygosity estimates on distant reference-sample pairs (Table 2)—the degree to which the heterozygosity increased with relaxed mapping methods and more distant references was surprising. As expected with the stringent global mapping method, when a reference genome is more distant, variants map more poorly, and less heterozygosity is detected. However, while local mapping methods map more sequence reads, they do so at the cost of accuracy. Mapping error increases as the distance between the reference genome and sample increases. As a consequence, we observed increased heterozygosity with both close and distant reference genomes compared to conspecific references with BWA.

In our study, the most efficient and least biased reference was a close relative (from within our ingroup clade) that was not conspecific with any of our study samples. This reference genome—from *Quercus mongolica*, a white oak from within our study clade but allopatric for the past ca. 15–20 million years from all included species—introduced less bias than a reference from one of the sampled species. When the reference genome came from a species in our study, more reads mapped for the conspecific samples, leading to a slightly higher estimated heterozygosity and longer phylogenetic tip lengths for the conspecific sample; this in turn could affect inferences about the timing of divergence in a phylogenetic study. Use of such an outgroup reference—*Quercus rubra* in our study—yielded low read mapping and, in BWA or Bowtie 2 --local, significantly higher heterozygosity than that estimated from the conspecific sample/reference pair, and tree topology change with increased imbalance. In our study, the ingroup reference not conspecific with any study samples—*Q. mongolica*—falls in the Goldilocks zone: superior to use over an outgroup reference as the genetic distance between the outgroup reference and our ingroup samples strongly reduced read-mapping efficiency, and over a conspecific reference due to minimal biases introduced.

A few studies have considered genetic relatedness in read mapping (Galla et al. 2019; Patton et al. 2019; Rick et al. 2023; Thorburn et al. 2023). Similar to our study, Thorburn et al. (2023) and Galla et al. (2019) found that having a more closely related reference resulted in better read mapping. Using genome resequencing data, Galla et al. (2019) concluded that a non-conspecific reference was adequate for population genetic statistics as long as it was related enough (con-familial or more closely related in their study). However, the percent of reads mapping to our congeneric outgroup reference was similar to the percent of reads mapping to the reference genome they found to

be too phylogenetically distant, which was from the same order as their ingroup samples (Galla et al. 2019). Therefore, "related enough" may represent different taxonomic levels depending on the study organisms.

The most stringent mapping method, Bowtie 2 --end-to-end, is both efficient and apparently unbiased in base pair recovery with closely related reference/sample pairs. While using Bowtie 2 --end-to-end resulted in fewer reads mapped than Bowtie 2 --local and BWA-MEM, it resulted in no significant decrease in recovered base pairs, and in more read pairs being mapped uniquely and concordantly in the genome than Bowtie 2 --local, provided the reference is conspecific with or closely related to the sample. Additionally and more importantly, Bowtie 2 --end-to-end yielded an estimate of heterozygosity similar across genome/sample genetic distances enabling downstream comparison. On the other hand, heterozygosity estimates were higher in Bowtie 2 --local and BWA-MEM and increased as the phylogenetic distance between the sample and reference increased especially for BWA-MEM. This is likely due to mismapping and multimapping (reads mapping to multiple places in the genome, as seen with Bowtie 2 --local versus Bowtie 2 --end-to-end), causing higher heterozygosity. Similar to our heterozygosity results, Thorburn et al. (2023) found higher nucleotide diversity within samples mapped to a more distant reference than the close reference using BWA-MEM. Unfortunately, the differences in heterozygosity results between methods limit the ability to compare across studies that use different methods.

When only a distantly related reference is available, using Bowtie 2 --end-to-end resulted in fewer base pairs recovered, slightly lower heterozygosity, and short tip-to-crown node branch lengths. Bowtie 2 --local recovered with a distantly related reference recovered the most base pairs–more base pairs than either BWA-MEM or Bowtie 2 --end-to-end. While Bowtie 2 --local significantly over-estimates heterozygosity using a distantly related reference, it is less biased than BWA-MEM. This may be due to the difference in scoring weights between Bowtie 2 and BWA-MEM (table 1; Canzar and Salzberg 2017; Wilton and Szalay 2022). Since BWA-MEM defaults weight mismatches less negatively than gaps while Bowtie 2 weights gaps slightly less negatively than mismatches (table 1; Wilton and Szalay 2022), the preference of mismatches over gaps by BWA-MEM may lead to inflated heterozygosity estimates. All three methods—regardless of the impact on heterozygosity—recovered more imbalanced tree topologies with a distant reference than with a close reference (Figure 3). Altogether, this suggests that none of these methods is adequate for mapping to a distant reference.

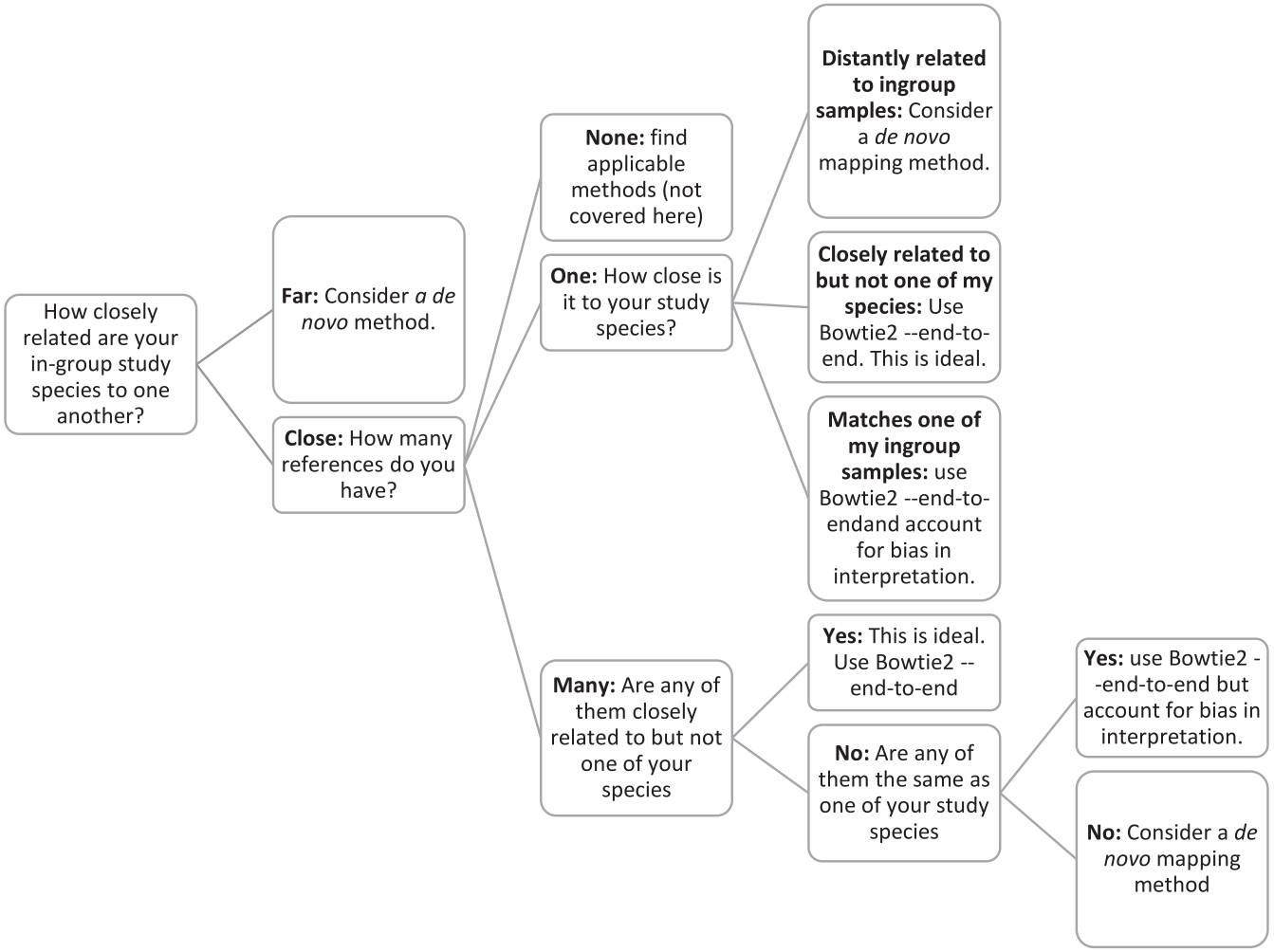

FIGURE 5 | A decision tree for choosing which reference and method to use for analyses.

While short-read data continues to be prevalent due to its affordability, it may be replaced with long-read data as the latter becomes cheaper and more accurate. Because of the increased length of overlap between long reads and an appropriate reference genome, there will likely be increased accuracy in mapping long reads. However, properly mapping real structural variants without increased mismapping may remain challenging. In the foreseeable future, however, short-read data are going to be an essential tool for comparative studies, and the recommendations presented here provide important guidelines for using such data appropriately and effectively.

4.1 | Phylogenetic Implications of Histories of Introgression

In the phylogenetic trees the ~75% bicolor—25% alba backcross was either sister to *Q. alba* or *Q. lobata* + *Q. macrocarpa* and had low bootstrap support in the trees with *Q. rubra* as reference and a local mapping method. This variation likely reflects the sample's real history with *Q. bicolor*, which is either sister to *Q. macrocarpa* (Hipp et al. 2020) or sister to *Q. macrocarpa* + *Q. lobata* (Larson et al. 2025). Thus, some of the variation in tree topology between different references and methods may be due to hybridisation and introgression in species histories. Utilising multiple individuals per species may assist with teasing apart species-wide histories vs. local or sample-specific histories of hybridisation. This is especially important in the Eastern North American white oaks, a known syngameon with extensive interspecific introgression (Hardin 1975; Hipp et al. 2019; Ribicoff et al. 2025).

One interesting outcome involves the multiple populations of *Q. macrocarpa*. The South Dakota (Black Hills) population of *Q. macrocarpa* is likely introgressed with *Q. gambelii* (Maze 1968), so we would expect *Q. macrocarpa* (S) and (N) to group together to the exclusion of *Q. macrocarpa* (SD). It is striking that the BWA-MEM analyses consistently violate this expectation while Bowtie 2 does not (neither in local nor global mode), and that the reference genome has no effect on the topology of populations within *Q. macrocarpa*. We are not sure how to interpret this, particularly in light of the fact that we have not taken into account introgression in our analyses in this study. But given what appears to be relatively high mapping inaccuracies using BWA-MEM compared with even Bowtie 2-local, we consider the BWA-MEM mapping more problematic. In other cases, different references highlight different aspects of the introgression histories. For example, the variation in the phylogenetic placement of *Q. alba* and *Q. muehlenbergii* as well as low bootstrap supporting their placement in some trees leads us to suspect that variation in topology reflects real histories of gene flow and is not due solely to errors in read mapping or variant calling. However, greater than 40% of SNPs in *Q. alba* are shared polymorphisms with other ENA white oak species (Larson et al. 2025), suggesting high levels of incomplete lineage sorting and pointing to the need to account for both lineage sorting and introgression in reconstructing the fine-scale evolutionary history of these species.

5 | Conclusions

Phylogenomic and population genomic analysis of multiple species should choose both the reference and the read mapping method carefully and consider the implications in their analysis (Figure 5). For multispecies analysis, a closely related, non-conspecific reference and Bowtie 2 -end-to-end should be used. However, when a close reference is not available, a conspecific reference will provide the next most accurate results, but bias should be considered in the data interpretation. If any of the ingroup samples are distantly related, Bowtie 2 --local can increase data recovery and minimise bias.

Author Contributions

Rebekah A. Mohn led the design, analysis, and writing of the manuscript. Mira Garner collected and extracted DNA from newly sequenced individuals. Paul S. Manos contributed the sequencing and the computational resources to complete this project. Andrew L. Hipp contributed significantly to shaping the design, and to the writing of the manuscript. All authors contributed to revisions.

Acknowledgements

The authors would like to thank Kasey Pham, Desanka Lazic, Drew Larson, Marlene Hahn, and Lindsey Worcester for feedback on ideas and the paper. The authors would also like to thank Jeannine Cavender-Bares and Leah Samuels for assistance collecting specimens. This work was supported by the National Science Foundation under award DEB-2129281 (ALH) and DEB-2129237 (PSM).

Disclosure

Benefit-Sharing Statement: Publicly sharing our data and results provides benefits to landowners and conservation decisions as understanding the impact of mapping methods on results is critical for conservation analysis.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Raw sequence reads for all individuals are deposited in the SRA (BioProject PRJNA1200080). Scripts, VCF files, and summary statistics are available on Dryad (DOI: 10.5061/dryad.fttdz0964).

References

Ai, W., Y. Liu, M. Mei, et al. 2022. "A Chromosome-Scale Genome Assembly of the Mongolian Oak (*Quercus Mongolica*)." *Molecular Ecology Resources* 22, no. 6: 2396–2410. https://doi.org/10.1111/1755-0998.13616.

Canzar, S., and S. L. Salzberg. 2017. "Short Read Mapping: An Algorithmic Tour." *Proceedings of the IEEE* 105, no. 3: 436–458. https://doi.org/10.1109/JPROC.2015.2455551.

Chen, S. 2023. "Ultrafast One-Pass FASTQ Data Preprocessing, Quality Control, and Deduplication Using Fastp." *iMeta* 2, no. 2: e107. https://doi.org/10.1002/imt2.107.

Crowl, A. A., P. S. Manos, J. D. McVay, A. R. Lemmon, E. M. Lemmon, and A. L. Hipp. 2020. "Uncovering the Genomic Signature of Ancient Introgression Between White Oak Lineages (*Quercus*)." *New Phytologist* 226, no. 4: 1158–1170. https://doi.org/10.1111/nph.15842.

Danecek, P., A. Auton, G. Abecasis, et al. 2011. "The Variant Call Format and VCFtools." *Bioinformatics* 27, no. 15: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

Danecek, P., J. K. Bonfield, J. Liddle, et al. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10, no. 2: giab008. https://doi.org/10.1093/gigascience/giab008.

Denk, T., and G. W. Grimm. 2010. "The Oaks of Western Eurasia: Traditional Classifications and Evidence From Two Nuclear Markers." *Taxon* 59: 351–366. https://www.jstor.org/stable/25677595.

Eizenga, J. M., A. M. Novak, J. A. Sibbesen, et al. 2020. "Pangenome Graphs." *Annual Review of Genomics and Human Genetics* 21: 139–162. https://doi.org/10.1146/annurev-genom-120219-080406.

Galla, S. J., N. J. Forsdick, L. Brown, et al. 2019. "Reference Genomes From Distantly Related Species Can be Used for Discovery of Single Nucleotide Polymorphisms to Inform Conservation Management." *Genes* 10, no. 1: 9. https://doi.org/10.3390/genes10010009.

Hardin, J. W. 1975. "Hybridization and Introgression in *Quercus Alba*." *Journal of the Arnold Arboretum* 56, no. 3: 336–363.

Hipp, A. L., P. S. Manos, M. Hahn, et al. 2020. "Genomic Landscape of the Global Oak Phylogeny." *New Phytologist* 226, no. 4: 1198–1212. https://doi.org/10.1111/nph.16162.

Hipp, A. L., and J. A. Weber. 2008. "Taxonomy of Hill's Oak (*Quercus Ellipsoidalis*: Fagaceae): Evidence From AFLP Data." *Systematic Botany* 33, no. March: 148–158. https://doi.org/10.1600/036364408783887320.

Hipp, A. L., A. T. Whittemore, M. Garner, et al. 2019. "Genomic Identity of White Oak Species in an Eastern North American Syngameon." *Annals of the Missouri Botanical Garden* 104, no. 3: 455–477. https://doi.org/10.3417/2019434.

Huang, H., and L. L. Knowles. 2016. "Unforeseen Consequences of Excluding Missing Data From Next-Generation Sequences: Simulation Study of RAD Sequences." *Systematic Biology* 65, no. 3: 357–365. https://doi.org/10.1093/sysbio/syu046.

Kapoor, B., J. Jenkins, J. Schmutz, et al. 2023. "A Haplotype-Resolved Chromosome-Scale Genome for *Quercus Rubra* L. Provides Insights Into the Genetics of Adaptive Traits for Red Oak Species." *G3 (Bethesda, Md.)* 13, no. 11: jkad209. https://doi.org/10.1093/g3journal/jkad209.

Keel, B. N., and W. M. Snelling. 2018. "Comparison of Burrows-Wheeler Transform-Based Mapping Algorithms Used in High-Throughput Whole-Genome Sequencing: Application to Illumina Data for Livestock Genomes1." *Frontiers in Genetics* 9: 35. https://doi.org/10.3389/fgene.2018.00035.

Langmead, B., and S. L. Salzberg. 2012. "Fast Gapped-Read Alignment With Bowtie 2." *Nature Methods* 9, no. 4: 357–359. https://doi.org/10.1038/nmeth.1923.

Langmead, B., C. Wilks, V. Antonescu, and R. Charles. 2019. "Scaling Read Aligners to Hundreds of Threads on General-Purpose Processors." *Bioinformatics* 35, no. 3: 421–432. https://doi.org/10.1093/bioinformatics/bty648.

Larson, D. A., M. E. Staton, B. Kapoor, et al. 2025. "A Haplotype-Resolved Reference Genome of *Quercus Alba* Sheds Light on the Evolutionary History of Oaks." *New Phytologist* 246: 331–348. https://onlinelibrary.wiley.com/doi/abs/10.1111/nph.20463.

Li, H. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs With BWA-MEM." *arXiv Preprint*. Retrieved from. http://arxiv.org/abs/1303.3997.

Li, H., and R. Durbin. 2009. "Fast and Accurate Short Read Alignment With Burrows–Wheeler Transform." *Bioinformatics* 25, no. 14: 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Linck, E., and C. J. Battey. 2019. "Minor Allele Frequency Thresholds Strongly Affect Population Structure Inference With Genomic Data Sets." *Molecular Ecology Resources* 19, no. 3: 639–647. https://doi.org/10.1111/1755-0998.12995.

Manel, S., C. Perrier, M. Pratlong, et al. 2016. "Genomic Resources and Their Influence on the Detection of the Signal of Positive Selection in Genome Scans." *Molecular Ecology* 25, no. 1: 170–184. https://doi.org/10.1111/mec.13468.

Maze, J. 1968. "Past Hybridization Between *Quercus Macrocarpa* and *Quercus gambelii*." *Brittonia* 20, no. 4: 321. https://doi.org/10.2307/2805689.

McVay, J. D., D. Hauser, A. L. Hipp, and P. S. Manos. 2017. "Phylogenomics Reveals a Complex Evolutionary History of Lobed-Leaf White Oaks in Western North America." *Genome* 60, no. 9: 733–742. https://doi.org/10.1139/gen-2016-0206.

Morales-Briones, D. F., G. Kadereit, D. T. Tefarikis, et al. 2021. "Disentangling Sources of Gene Tree Discordance in Phylogenomic Data Sets: Testing Ancient Hybridizations in Amaranthaceae s.l." *Systematic Biology* 70, no. 2: 219–235. https://doi.org/10.1093/sysbio/syaa066.

Nguyen, L. T., H. A. Schmidt, A. Von Haeseler, and B. Q. Minh. 2015. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." *Molecular Biology and Evolution* 32, no. 1: 268–274. https://doi.org/10.1093/molbev/msu300.

Ortiz, E. M. 2019. Vcf2phylip v2.0: Convert a VCF Matrix into Several Matrix Formats for Phylogenetic Analysis. https://doi.org/10.5281/zenodo.2540861. Zenodo.

Patton, A. H., M. J. Margres, A. R. Stahlke, et al. 2019. "Contemporary Demographic Reconstruction Methods Are Robust to Genome Assembly Quality: A Case Study in Tasmanian Devils." *Molecular Biology and Evolution* 36, no. 12: 2906–2921. https://doi.org/10.1093/molbev/msz191.

Prasad, A., E. D. Lorenzen, and M. V. Westbury. 2022. "Evaluating the Role of Reference-Genome Phylogenetic Distance on Evolutionary Inference." *Molecular Ecology Resources* 22, no. 1: 45–55. https://doi.org/10.1111/1755-0998.13457.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Ribicoff, G., M. Garner, K. Pham, et al. 2025. "Introgression, Phylogeography, and Genomic Species Cohesion in the Eastern North American White Oak Syngameon." *Molecular Ecology* 34: e17822. https://doi.org/10.1111/mec.17822.

Rick, J. A., C. D. Brock, A. L. Lewanski, J. Golcher-Benavides, and C. E. Wagner. 2023. "Reference Genome Choice and Filtering Thresholds Jointly Influence Phylogenomic Analyses." *Systematic Biology* 73: syad065. https://doi.org/10.1093/sysbio/syad065.

Shafer, A. B. A., C. R. Peart, S. Tusso, et al. 2017. "Bioinformatic Processing of RAD-Seq Data Dramatically Impacts Downstream Population Genetic Inference." *Methods in Ecology and Evolution* 8, no. 8: 907–917. https://doi.org/10.1111/2041-210X.12700.

Thorburn, D.-M. J., K. Sagonas, M. Binzer-Panchal, et al. 2023. "Origin Matters: Using a Local Reference Genome Improves Measures in Population Genomics." *Molecular Ecology Resources* 23, no. 7: 1706–1723. https://doi.org/10.1111/1755-0998.13838.

Wilton, R., and A. S. Szalay. 2022. "Performance Optimization in DNA Short-Read Alignment." *Bioinformatics* 38, no. 8: 2081–2087. https://doi.org/10.1093/bioinformatics/btac066.

Yang, Y., and S. A. Smith. 2014. "Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics." *Molecular Biology and Evolution* 31, no. 11: 3081–3092. https://doi.org/10.1093/molbev/msu245.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Supporting Information S1:** Leaf tissue sterilisation protocol. **Supporting Information S2:** QUERCUS DNeasy extraction protocol.