Syst. Biol. 70(4):844–854, 2021

DOI:10.1093/sysbio/syaa081
Advance Access publication October 21, 2020

A Cautionary Note on the Use of Genotype Callers in Phylogenomics

PABLO DUCHEN* AND NICOLAS SALAMIN

Department of Computational Biology, University of Lausanne, Quartier Sorge, 1015 Lausanne, Switzerland
**Correspondence: Pablo Duchen, Department of Computational Biology, University of Lausanne, Quartier Sorge, 1015 Lausanne, Switzerland; E-mail: pablo.duchenbocangel@unil.ch*

Received 10 June 2020; reviews returned 2 October 2020; accepted 2 October 2020
Associate Editor: Lars Jermiin

Abstract.—Next-generation-sequencing genotype callers are commonly used in studies to call variants from newly sequenced species. However, due to the current availability of genomic resources, it is still common practice to use only one reference genome for a given genus, or even one reference for an entire clade of a higher taxon. The problem with traditional genotype callers, such as the one from GATK, is that they are optimized for variant calling at the population level. However, when these callers are used at the phylogenetic level, the consequences for downstream analyses can be substantial. Here, we performed simulations to compare the performance between the genotype callers of GATK and ATLAS, and present their differences at various phylogenetic scales. We show that the genotype caller of GATK substantially underestimates the number of variants at the phylogenetic level, but not at the population level. We also found that the accuracy of heterozygote calls declines with increasing distance to the reference genome. We quantified this decline and found that it is very sharp in GATK, while ATLAS maintains high accuracy even at moderately divergent species from the reference. We further suggest that efforts should be taken towards acquiring more reference genomes per species, before pursuing high-scale phylogenomic studies. [ATLAS; efficiency of SNP calling; GATK; heterozygote calling; next-generation sequencing; reference genome; variant calling.]

Next-generation sequencing (NGS) grants access to a wealth of genomic data such as full genomes, transcriptomes, enriched target loci, or RADseq. The increased availability of such data has revolutionized the way we study evolutionary processes at the population and species levels. Consequently, the growth in technologies to improve the sequencing of vast amounts of genomic regions was accompanied by the development of a diverse range of bioinformatic tools to analyze these new data (McCormack et al. 2013).

One of the main constraints of current NGS approaches is the reliance on a reference genome to call genetic variants. Many model organisms (such as humans or the fruit fly *Drosophila melanogaster*) have well-annotated reference genomes (dos Santos et al. 2015). For nonmodel organisms, however, access to species-specific reference genomes is still very often limited. Nielsen et al. (2011) stated that for an appropriate variant-calling accuracy, the amount of sequence identity between the reads and the reference (or the amount of tolerable mismatches between reads and reference) has to be optimized for every species individually. For example, if the reference genome used for variant calling in humans is used for other organisms with different levels of genetic diversity, then there will be a significant loss of sequencing depth, and thus, variability in many regions will be underestimated (Nielsen et al. 2011). The effect of the reference genome used for the analysis of newly sequenced individuals will therefore be different depending on the evolutionary scale being studied. Consequently, variant calling at a population level (with reference genomes close to the target populations) might be more accurate than variant calling at a phylogenetic level (with reference genomes that are distant to the species analyzed). So far, there are no studies that have looked at this specific question.

What is also clear is that, currently, we are still far from having well-annotated reference genomes for all species. For instance, countless studies over the past few years had access to only one single reference genome, if any, despite analyzing multiple species within entire genera. Such is the case for fish (e.g., Chakrabarty et al. 2017; Burress et al. 2018; Hulsey et al. 2017, 2018; Marcionetti et al. 2019), mammals (e.g., Kumar et al. 2017; Lima et al. 2018; Moura et al. 2020), birds (e.g., Ottenburghs et al. 2016), reptiles (e.g., Bragg et al. 2016), amphibians (e.g., Portik et al. 2016), insects (e.g., Yan et al. 2020), and plants (e.g., Nobre et al. 2018; Kreuzer et al. 2019; Helmstetter et al. 2020; Olvera-Mendoza et al. 2020; Brandrud et al. 2020; Wang et al. 2020). The situation can sometimes be even more dramatic with a single reference genome available for entire tribes or subfamilies, encompassing multiple genera (e.g., Hulsey et al. 2017; Wang et al. 2017; Hulsey et al. 2018; Heckenhauer et al. 2019; Loiseau et al. 2019).

The current situation raises the question of how read mapping and variant calling are affected by the distance between the target species (or populations) and the reference genome. As stated by Nielsen et al. (2011), using optimized pipelines developed for one specific species often results in suboptimal variant calling in another species. For instance, Schubert et al. (2014) acknowledge the fact that the widely used Genome Analysis Toolkit GATK (DePristo et al. 2011) is not well suited for nonhuman organisms, since, among other things, it depends on external information such as known sets of variant sites that are often unavailable for nonmodel organisms.

Some studies have already addressed the impact of single references on multispecies NGS analyses. For instance, mapping to a single reference genome within one genus versus using *de novo* assembly can result in fewer variants being called by the reference-based method (Fitz-Gibbon et al. 2017). Second, using few references for multiple species can confound paralogy with orthology, which in turn, affects phylogenetic reconstruction (Chakrabarty et al. 2017). Third, the further the phylogenetic distance from a reference genome the less efficient the NGS pipelines become (because of mapping issues), as it is the case for target enrichment (Bragg et al. 2016) or transcriptome-based loci (Portik et al. 2016).

The efficiency in calling sites that are heterozygous may also be undermined when the reference genome is too distant. Consequently, heterozygote positions are often discarded due to the uncertainty in heterozygote calls, which in turn affects downstream analyses, such as the inference of divergence times (Lischer et al. 2014). Such problems also pertain to ploidy levels above two, for which the development of newer methods for variant calling is necessary (Blischak et al. 2018). Finally, another important obstacle when calling variants in a newly sequenced species is the lack of knowledge on species-specific variable sites, which are used to recalibrate base quality scores (Schubert et al. 2014). Such knowledge is very helpful for judging the accuracy of variant calls, and for obtaining more accurate SNPs.

Overall, the degree of information loss when using one reference genome for multispecies NGS studies is currently unknown. Therefore, we conducted a simulation study to assess the efficiency of variant calling across various phylogenetic scales. We hypothesize a sharp drop in this efficiency as the evolutionary distance to the reference genome increases. The advantage of using simulations is that we know the exact position of variable sites, so that variant calling using traditional pipelines can be tested and the loss of information quantified. More specifically, we simulated diploid reads along with different tree topologies and different evolutionary scales. For every tree and every scaling, we selected one reference genome and performed variant calling with the genotype callers of GATK 4.1 and ATLAS v. 0.9 (Link et al. 2017). We focused on two aspects that affect downstream analyses substantially, namely: the number of called variants, and the accuracy of heterozygote calling.

Methods

Our pipeline followed these general steps: simulations of species trees, rescaling of the trees to match various levels of divergence, simulation of genomic sequences along the trees, and simulation of diploid reads from each genomic sequence. From each tree, we selected one sequence as a reference and performed standard genome assembly and variant calling with GATK and ATLAS. Each of these steps is detailed below.

Simulation of Trees and Scaling

We performed two separate sets of tree simulations and scalings. As a first set, we simulated two contrasting topologies with arbitrary rescalings of each topology (*Base cases* step). The motivation for the base-cases analysis was to explore the performance of genotype callers at varying phylogenetic scales while keeping the topology constant. For the second set of simulations (*Generalization* step), we simulated 10 random topologies and drew random rescaling values from a uniform distribution. The motivation for this latter analysis was to check the robustness of our results with several randomly generated topologies.

Base cases.— As stated above, we used two types of trees (each with 20 tips) representing two contrasting topologies (Fig. 1). The purpose of using these two topologies was to make an initial assessment of the potential effect that the topology and the position of the reference genome can have on genotype calling. The first tree, referred to as "Tree A", was a standard birth–death tree simulated with the R package *TreeSim* by Stadler (2011) (Fig. 1, left panel). The second tree represents a scenario with a very recent burst of speciation (as opposite to Tree A). This tree was simulated with *ms* (Hudson 2002) and is referred to as "Tree B" (Fig. 1, right panel). Although *ms* is traditionally used for coalescent simulations, we used it here only to simulate this particular topology (the command line is shown in the Supplementary Material, Section A.1 available on Dryad at https://doi.org/10.5061/dryad.fn2z34ts3). We scaled the trees to three arbitrary total depths of 0.13, 0.065, and 0.013 measured in units of substitutions per site. Although these values were arbitrarily chosen, they do correspond to realistic values observed in phylogenetic trees at the genus level. For instance, the first scaling corresponds to the actual divergence observed in the phylogenetic tree of clownfishes (Litsios et al. 2014). The second and third scaling values correspond to half and one-tenth of the original value of the first scaling, representing smaller clades within a phylogenetic tree. We called these rescaled values Large, Medium, and Small, respectively. Thus, the simulations in the *base cases* step included six types of trees: tree A Large, tree A Medium, tree A Small, tree B Large, tree B Medium, and tree B Small (Table 1, first column). Finally, to check the performance of genotype callers on a level of divergence that can be considered as including several distinct populations represented by single individuals, we added a fourth scaling of trees A and B to reach a much smaller divergence with a total tree depth of 0.0013.

Generalization.— The second series of simulations was performed by generating (using again the R package *TreeSim*) 10 random topologies of 20 species each, with a lineage birth rate drawn from the uniform distribution $U(0.5,1)$ and an extinction rate drawn from $U(0.05,0.5)$. The trees were then rescaled by drawing a scaling factor from a uniform distribution $U(0.01,0.3)$. Such

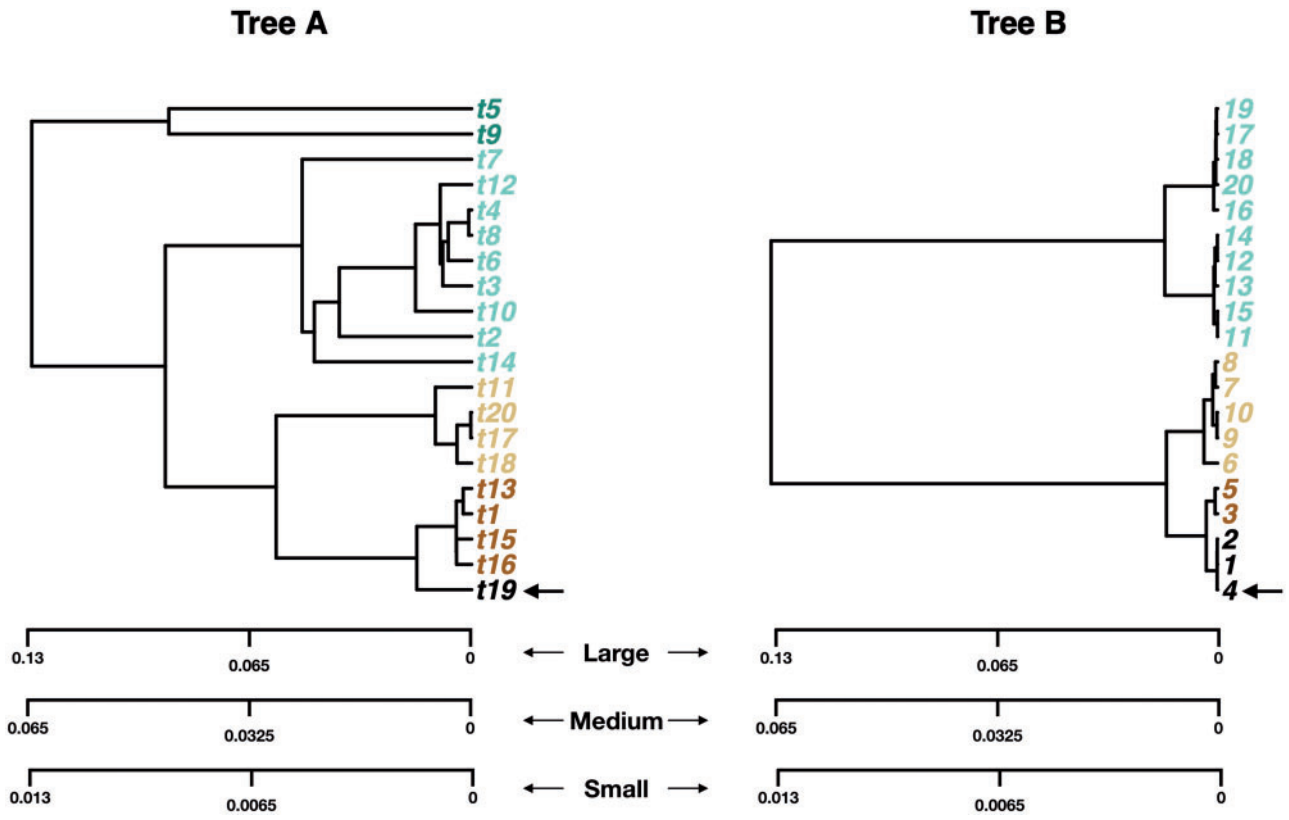

FIGURE 1. *Base-cases* example trees: a birth–death tree A (left), and a recent-burst tree B (right). For each tree we generated three arbitrary rescalings: Large, Medium, and Small, each representing various types of divergence found in phylogenetic studies. The chosen reference genome is shown with an arrow. The colors of the tip labels are chosen to represent the increasing distance to the reference (these colors will be used again in Figs. 4 and 5). The command lines and parameters used to generate these trees are described in the Supplementary Material, Section A.1 available on Dryad.

TABLE 1. Some stats from the sequence simulations, mapping of reads, and genotype calling subroutines: Total number of SNPs for each alignment of each tree, mean (and standard deviation, sd) read-mapping quality (PHRED scores) for each tree (a per-species distribution of read-mapping qualities is shown in SI-Fig. B.1, and SI-B.2), and total variant-calling computation time for all 20 species of each tree.

Tree	Number of SNPs	Mean and (sd) PHRED	Computation time (GATK)	Computation time (ATLAS)
A-large	193,884	57.04 (6.84)	~6h	~5min
A-medium	97,087	59.85 (1.47)	~5h	~4min
A-small	19,287	60.00 (0.00)	~30min	~4min
B-large	143,715	57.27 (6.83)	~4h	~3min
B-medium	72,113	59.74 (1.99)	~3.5h	~3min
B-small	14,256	60.00 (0.00)	~30min	~2min
A-population	1978	60.00 (0.00)	~10min	~2min
B-population	1429	60.00 (0.00)	~10min	~2min

distribution spans divergence values both lower and higher than the values used in the *base cases* section.

Simulation of DNA Sequences

For each of the trees generated in the previous section, we simulated 1 million base-pair (bp) long sequences with the software *seq-gen* (Rambaut and Grass 1997) under the HKY model (Hasegawa et al. 1985). Each tree (with 20 tips) will thus contain 20 simulated sequences at the tips to account for the stochasticity that is inherent to models of sequence evolution. Concerning the expected percentage of invariant sites, we started with an arbitrary value of 80% for the large divergence scaling of trees A and B, which resulted in about 20% of variant sites in the final alignment. This percentage is in line with what can be observed in phylogenetic trees at the genus level (e.g., Duchen and Renner 2010; Litsios et al. 2014). For the Medium and Small trees (with half and one-tenth of the divergence, respectively) the expected value of the percentage of variant sites in the simulations of DNA sequences is 10% for the Medium tree and 2% for the Small tree, assuming that mutation rates and all other simulation parameters are set to the same values (see the Supplementary Material available on

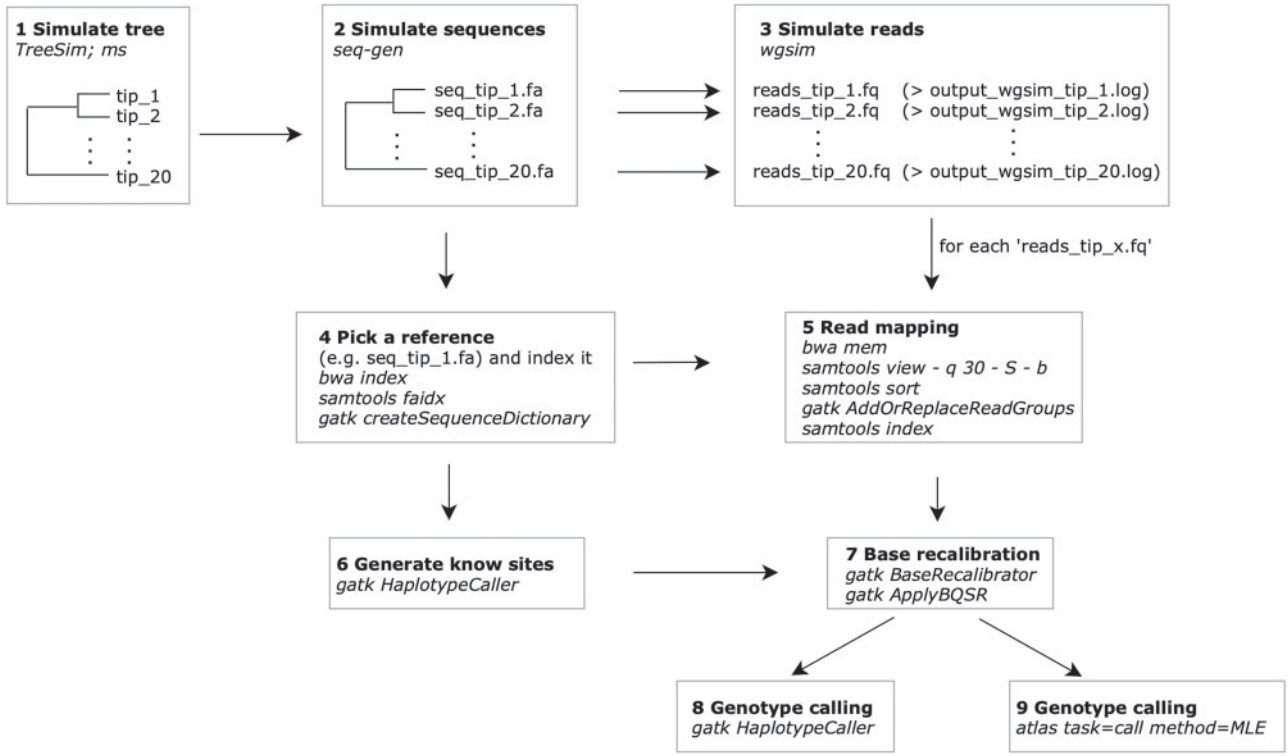

FIGURE 2. Summary of the main steps taken in this study. The order of steps is given by the numbers in each box. Initial sequence simulation and indexing of the reference genome involve haploid sequences (steps 1, 2, 4, and 6). Simulation of reads, mapping, recalibration, and genotype calling involve diploid sequences (steps 3, 5, 7, 8, and 9). Arrows indicate that some output of a previous step will be used as input for a next step. Programs used, and a summary of the command lines are indicated in italics. A complete description of all command lines is given in the Supplementary Material, Section A available on Dryad.

Dryad for a complete transcript of the pipeline and code). For the fourth rescaling representing population-level divergence, we expected to have around 0.2% of variant sites (provided that the original tree was 100 times more divergent and had 20% of variant sites). A similar approach was used to set up the proportion of invariable sites for the *Generalization* step. The proportion was taken between 2% and 20%, with the latter value assigned to the topology with the highest divergence. The sequences generated by *seq-gen* are all haploid. The generation of diploid reads necessary to simulate data used by genotype callers is described below (Fig. 2).

Simulation of Diploid Reads

For each simulated sequence at the tips of every topology, we simulated 100,000 diploid reads (each read 100 bp long) with the software *wgsim* (Li 2013). Briefly, *wgsim* selects regions (from the input sequence) of a user-specified length (100 bp in our case). With these regions, *wgsim* will substitute, delete, or insert nucleotides with probabilities provided by the user as well. For the case of diploid reads, *wgsim* will substitute a nucleotide with another nucleotide or any other ambiguity IUPAC letter, which codes for ambiguous "heterozygous" nucleotides. All positions where heterozygote calls have been introduced in the simulated reads are recorded in the output of *wgsim*. We used the default parameters, which resulted in a total sequencing depth of 10×, a mutation rate of 0.001, and about 15% of indels introduced. These default parameters generated a total heterozygosity of around 0.06% per simulated species. To account for species-specific heterozygosity and mutation rates, we have also varied the mutation rate from half up to twice the default value, drawing it from $U(0.0005,0.002)$ for each species. We used *wgsim* because it allows us to know the exact position and state (heterozygote/homozygote) of all variants simulated, which is a key aspect that we want to test to understand the performance of genotype callers.

Mapping of Reads and Generation of BAM Files

For every tree and scaling factor, we selected one sequence that was used as the reference for variant calling (Fig. 2). In each simulated data set, the reference sequence was indexed with *bwa index* (Li and Durbin 2010), *samtools faidx* (Li et al. 2009), and *gatk CreateSequenceDictionary* (DePristo et al. 2011). The mapping of the simulated reads was done with *bwa* using the option *mem*. All SAM files generated during this step were converted to BAM format with *samtools view* and we discarded reads with mapping qualities below 30.

We then sorted all BAM files with the option *samtools sort*, added read groups with *gatk AddOrReplaceReadGroups*, and finally reindexed the files with *samtools index*. For tree A and its scalings, we repeated this analysis with two other reference sequences in order to assess the effect of the position of the reference in the phylogenetic tree on genotype calling.

Base Recalibration and Genotype Calling

Base recalibration makes use of known variable positions. As recommended by the GATK Best Practices protocol (Van der Auwera et al. 2013), if these positions are unknown, one possible approach is to run the genotype caller of GATK once (or a few times) and use the output to recalibrate the quality scores. In our simulations, we know the positions of the variable sites but we did not use them for the recalibration, since our goal was precisely to assess the performance of genotype callers when there is no previous information on the positions of variable sites, which is the case in most newly sequenced nonmodel organisms. We therefore ran the genotype caller of GATK once to generate a table of inferred variable sites and the base recalibration was then accomplished with the *BaseRecalibrator* and *ApplyBQSR* subroutines from GATK.

We used the *HaplotypeCaller* subroutine from GATK (DePristo et al. 2011), and the *MLE* subroutine of the software ATLAS (Link et al. 2017) to call genotypes, using the exact same recalibrated BAM files described above as input in both cases. The main difference between ATLAS and GATK is that the former computes the genotype likelihoods of all 10 possible genotypes at every given SNP, while the latter estimates the most likely minor allele before computing the three possible genotype likelihoods (see DePristo et al. (2011) and Link et al. (2017) for the description of both methods). Both genotype callers generate VCF files (Danecek et al. 2011) as output. For each simulated SNP, we recorded whether the genotype returned by both callers matched the reference genome (i.e., "0/0" in the genotype field of the VCF file) or not (i.e., "1/1"), and whether the site was a heterozygote (i.e., "0/1").

Performance of Genotype Callers

Total number of called variants.— We measured the performance of the two genotype callers GATK and ATLAS by first extracting the total number of called variants in every sequence of the simulated trees. This was done by counting in each VCF file the number of times a non "0/0" variant was called. We then compared this number to the true number of variants that is known from the simulations. This true number of variants can be estimated by counting the number of differences between each sequence and the corresponding reference, plus the number of simulated heterozygote positions, which is part of the output of the diploid-read simulator *wgsim* (see section *Simulation of Diploid Reads*).

We focused on the number of variants for several reasons. First, the number of variants is an important measure of diversity, which is why it is expected for genotype callers to correctly retrieve it, surpassing the potential effect of sequencing errors. Additionally, at the phylogenetic level, an accurate calling of homozygotes (that are different from the reference) can be an indicator of fixed substitutions that are, in turn, important to define the divergence between species and to measure the evolution of molecular markers.

Accuracy in Heterozygote Variant Calling

We extracted from the VCF files generated by GATK and ATLAS the positions of all heterozygote calls with the functions *extract.gt* and *is.het* from the R package *vcfR* (Knaus and Grünwald 2017). We first tested if the total number of called heterozygotes corresponded to the true number of simulated heterozygotes. More precisely, we calculated the ratio between the number of called heterozygotes (by both GATK and ATLAS) and the true number of heterozygote positions obtained during the simulations. We called this ratio the *Number of Called/True heterozygotes*. We expected this ratio to be 1 for a perfect accuracy. A ratio above 1 would mean that there are more heterozygotes being called than the number simulated (representing false positives), and a ratio below 1 would mean that the number of heterozygote positions is underestimated by the genotype caller. We also recorded if the positions of the called heterozygotes were correctly inferred by comparing them with the true positions simulated. This measure was estimated by counting the number of times a heterozygote position was correctly called by GATK and ATLAS and divided it by the total number of true heterozygote positions known from the simulated reads. We called this the *Accuracy of heterozygote variant calling*. We expected an accuracy of 1 for a perfect position match between called and true heterozygote positions.

Calculating Phylogenies from Genotype-Calling Output

We converted the VCF output (Tree A, Large scale, and population-level scale) from both GATK and ATLAS to FASTA sequences and kept all called variants to build alignments. During conversion, ambiguous heterozygote calls were coded with their respective IUPAC codes, and indels were excluded to avoid alignment issues. The phylogenetic trees were then calculated with RAxML-ng (Kozlov et al. 2019) using the GTJC model, which allows for 10 states (4 nucleotides plus 6 ambiguity codes). We used the sequence "t9" to root the tree. For both contrasting scalings (Large and population level) the following trees were inferred: 1) a tree with the original sequences simulated by *seq-gen*, 2) a tree with the converted output of ATLAS, and 3) a tree with the converted output of GATK. Topological distances between the inferred trees and the

true tree were calculated using the method of Kuhner and Felsenstein (1994).

Results

Simulation of Sequences and Reads

We initially simulated two topologies to illustrate the effect of the divergence between a reference genome and some newly produced genomic data. These topologies were produced with a birth–death process (tree A) or a coalescent-based process (tree B), representing two contrasting tree shapes. We then rescaled each topology to three different depths: Large (tree depth 0.13), Medium (tree depth 0.065), and Small (tree depth 0.013). We also added one more scaling of trees A and B to represent population-level divergences (tree depth 0.0013). We then simulated 10 additional topologies with birth and death rates taken from a uniform distribution and rescaled each topology to depths between 0.01 and 0.3 (Supplementary Fig. SI-B.4 available on Dryad, refer to section *Generalization* for details). From the output of *wgsim*, between 50% and 60% of all mutations resulted in heterozygous sites, meaning that the total default heterozygosity introduced by *wgsim* per simulated species is around 0.06% when the mutation rate is set to the program's default value. When introducing species-specific mutation rates the heterozygosity varied between half and twice this default value (i.e., between 300 and 1200 heterozygote positions per simulated sequence). The total number of segregating sites per alignment varied depending on the scale of the tree (Table 1).

Mapping of Reads

The reads were mapped with different qualities depending on the distance to the corresponding reference (recall that we chose one reference per tree). For instance, reads belonging to the reference sequence or to tips adjacent to the reference had PHRED scores of 60. Mapping qualities then decreased with increasing distance to the reference (Supplementary Fig. SI-B.1 available on Dryad, scale Large). Topologies of Medium and Small scales maintained an overall high read-mapping quality (Supplementary Fig. SI-B.1 available on Dryad, scales Medium and Small). The same tendency was observed in tree B, with the only difference that we see a sharper difference in mapping quality when changing clades (Supplementary Fig. SI-B.2 available on Dryad).

Genotype Calling

Number of called variants.— We counted the number of called variants, that is, all calls that are different from the reference (calls not of the form "0/0" in the genotype field of the output VCF file of each species). For tree A with the Small scale, we found that both GATK and ATLAS accurately recovered the number of variants (Fig. 3; third column). However, for the Medium and Large scales GATK substantially underestimated the number of variants (Fig. 3; first and second columns). The caller ATLAS, on the other hand, did a perfect job for the Medium scales, but still underestimated the number of variants for the Large scale (particularly the tips that are far from the reference, Fig. 3, red lines), although this underestimation was not as sharp as the one seen for GATK. We observed a similar tendency for tree B (Fig. 3, second row). A similar pattern was also observed when we simulated random topologies with random scalings: when the divergence from the reference is low then both GATK and ATLAS recover the true number of variants, while GATK underestimated the number of variants with increasing divergence. In contrast, ATLAS was able to outperform GATK (Supplementary Fig. SI-B.5 available on Dryad). We found similar results when using species-specific mutation rates (Supplementary Fig. SI-B.11 available on Dryad), and when using references that branched out deeper in tree A (Supplementary Fig. SI-B.8 available on Dryad).

Calling of Heterozygotes

After running the *HaplotypeCaller* and *MLE* subroutines of GATK and ATLAS, respectively, we found little difference in their performance for the simulations at the population level. In this case, they both called the true heterozygote variants accurately, although there was a 20% overestimation of the heterozygote calls by GATK (Supplementary Fig. SI-B.3 available on Dryad). However, we found the opposite result when dealing with the simulations at the phylogenetic level. The accuracy in heterozygote calling declined with increasing distance to the corresponding reference genome, and this reduction was particularly strong for GATK when compared to ATLAS. This pattern was observed in both types of topologies (Figs. 4 and 5, second column), in all 10 additional simulated phylogenies under various scalings (Supplementary Fig. SI-B.6 available on Dryad), in the simulations with the different reference sequences (Supplementary Figs. SI-B.9 and SI-B.10 available on Dryad, second column), and in the simulations with species-specific mutation rates (Supplementary Fig. SI-B.12 available on Dryad).

We also found that GATK tended to overestimate the number of heterozygotes when the genomes were close to the reference, but this tendency decreased dramatically with increasing phylogenetic distance to the reference (Figs. 4 and 5, second row). In other words, at a phylogenetic scale, GATK will substantially underestimate the true number of heterozygous positions along the genome. The genotype caller of ATLAS, on the other hand, maintained the ratio of called vs. true heterozygote variants close to 1 most

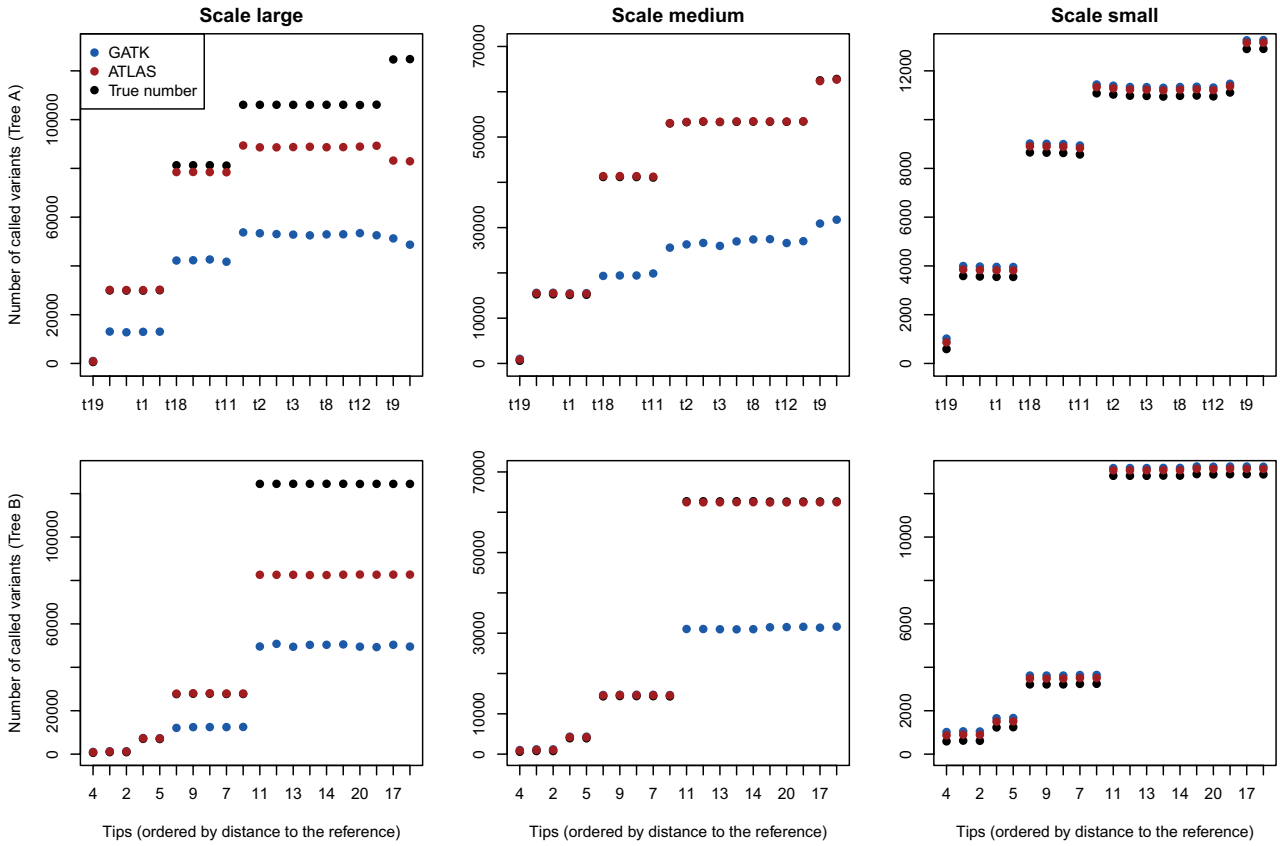

FIGURE 3. Number of called variants for the tips of tree A (first row), and tree B (second row). The tips on the x-axis are ordered according to their distance to the reference (the reference being always at the extreme left).

of the time, except for the most distant species to the reference, and only for the large scale trees (Figs. 4 and 5, third column). A similar pattern was observed with the 10 additional simulated phylogenies under various scalings (Supplementary Fig. SI-B.7 available on Dryad), and in the simulations with the different references (Supplementary Figs. SI-B.9 and SI-B.10 available on Dryad, third column).

Inferring Phylogenetic Trees

As an example of the impact of using only one reference in genotype calling at a phylogenetic scale, we inferred phylogenetic trees using the output of GATK and ATLAS on the simulated data sets produced with tree A. First, we picked the large scale version of tree A because it represents the largest differences between the tips and because it is the case where we see the most differences in terms of genotype-calling accuracy. As a contrast, we also picked the version of tree A with the smallest scale (population level). In both cases, we found that the inferred trees differ from the true tree and from the inferred tree based on the original simulated *seq-gen* sequences (Supplementary Table SI-C.1, Fig. SI-B.13 available on Dryad). The tree inferred from the genotype calls of ATLAS (as opposite to GATK) is closer to the tree inferred from the *seq-gen* alignment (Supplementary Table SI-C.1 available on Dryad). To make these comparisons valid, all alignments used include the heterozygous sites introduced by *wgsim*. Adding heterozygous sites makes a measurable impact in tree inference, as evidenced by the large topological distance with the original tree simulated by *TreeSim* (Supplementary Table SI-C.1 available on Dryad).

DISCUSSION

In this study, we tested the performance of state-of-the-art genotype callers on simulated data sets aimed at representing evolutionary divergences, either within or above the species level. Genotype calling applied on reads simulated at a population level reached an overall good accuracy, although GATK overestimated the number of heterozygotes by as much as 20% (Supplementary Fig. SI-B.3 available on Dryad, last panel). Such overestimation of heterozygotes by GATK has also been reported in previous studies (Hwang et al. 2015; Link et al. 2017). In contrast, ATLAS was able to call heterozygotes with a 100% accuracy in all cases at the population level (Supplementary Fig. SI-B.3 available on Dryad, last panel). When the divergence

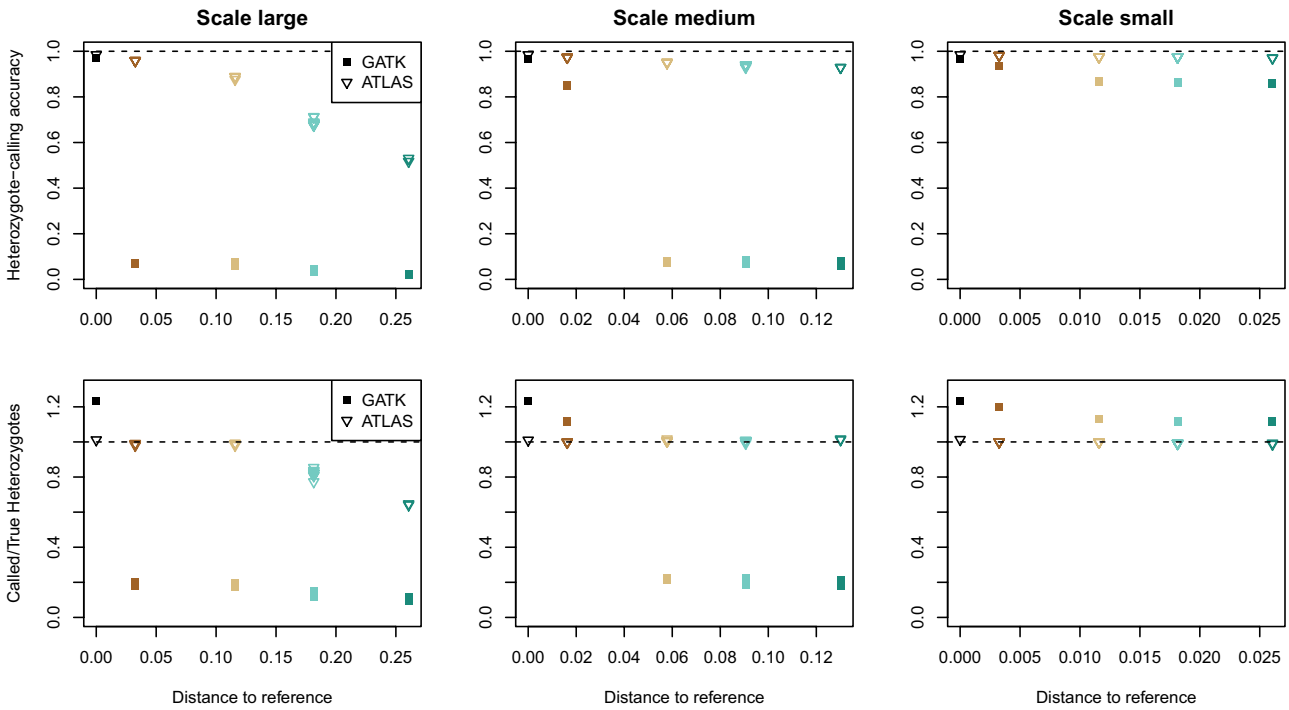

FIGURE 4. Accuracy of heterozygote calling for tree A at three different phylogenetic scales: Large (first column), Medium (middle column), and Small (last column). The phylogenetic distance from each tip of the phylogeny to the reference is plotted against the heterozygote-calling accuracy (first row), and against the called versus true heterozygotes (second row). The colors of each point or symbol correspond to the tip colors shown in Fig. 1.

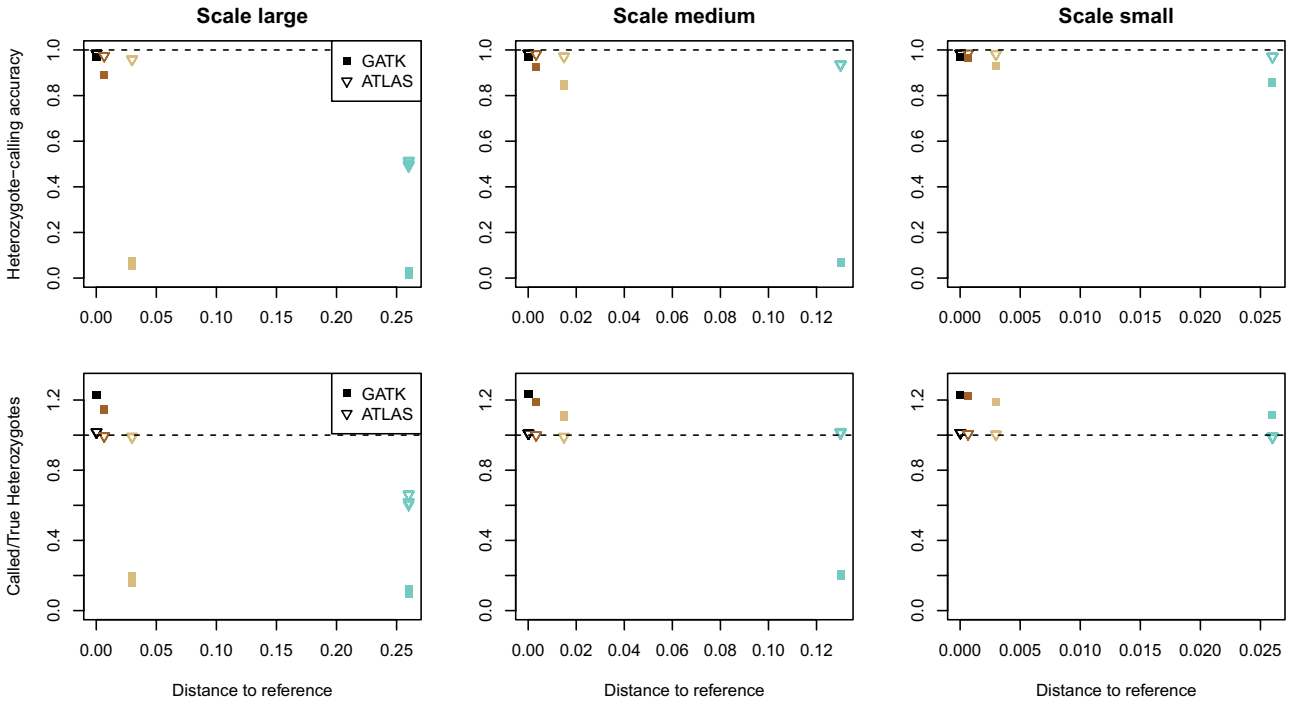

FIGURE 5. Accuracy of heterozygote calling for tree B at three different phylogenetic scales: Large (first column), Medium (middle column), and Small (last column). The phylogenetic distance from each tip of the phylogeny to the reference is plotted against the heterozygote-calling accuracy (first row), and against the called versus true heterozygotes (second row). The colors of each point or symbol correspond to the tip colors shown in Fig. 1.

between the simulated sequences increased to a level representative of shallow phylogenomic studies, we found the performance of the GATK suboptimal. For instance, we observed a large underestimation in the total number of called variants (Fig. 3), and low accuracy when calling heterozygote positions (Figs. 4 and 5). While this issue is particularly prevalent in GATK, the caller ATLAS is less affected. The main difference between the two callers lies in their assumptions when estimating the possible genotype likelihoods necessary to call variants. ATLAS considers all ten possible genotype likelihoods, which might be better suited when divergent sequences are considered. GATK, on the other hand, infers one minor allele to be used in the genotype likelihood calculations. As expected, the underperformance of genotype calling is stronger when the phylogenetic distance between the species-specific reads and the reference genome (against which these reads were mapped) becomes larger (see also Bragg et al. 2016; Portik et al. 2016; Fitz-Gibbon et al. 2017). Additionally, GATK has an inherent filter such that many low-quality calls are not emitted. In contrast, ATLAS emits all calls and lets the user decide on the filtering, which might also explain the better performance of ATLAS (in our case we left the default parameters, see Supplementary Material Section A.5 available on Dryad). We assume this might also explain the lower number of variants output by GATK as opposed to ATLAS.

Depending on the research question, underestimating the number of variants can have different consequences, but failing to accurately call heterozygote positions can have more detrimental effects (Lischer et al. 2014), which is why we focused on this particular parameter in the present study. This problem can also be exacerbated for higher ploidy levels, for which newer methods for variant calling under such circumstances have been suggested (Blischak et al. 2018). Additionally, many studies have acknowledged the problems of lacking species-specific references, and they opted for the solution to build pseudoreferences or synthetic references for every species instead (e.g., Bateman et al. 2018; Grummer et al. 2018; Skipwith et al. 2019). However, such studies focused more on target/sequence capture that are usually designed for interspecific studies and where the call of heterozygotes is of less interest. We focused instead on studies that make use of entire genomes for phylogenomic or population genomic analyses across multiple species, with well-annotated reference genomes.

The tree topology also seems to play a role in determining the accuracy of genotype calling. Here, we started with two contrasting topologies: a "typical" birth–death topology and a topology representing high speciation rates in the recent past (Fig. 1). In both cases, and at all scalings tested, the decrease of accuracy with increasing distance to the reference is not linear but jumps as it goes through the different clades of the tree (Fig. 5, second and third columns). In other words, the further the most recent common ancestor between the reference and a given species is, the sharper the decrease in genotype-calling accuracy that we found.

There are several reasons that might explain the low performance of genotype calling when the divergence to the reference genome increases. First, the read-mapping quality reduces with increasing distance to the reference because of the lower similarity between the reads and the reference sequences (Supplementary Figs. SI-B.1 and SI-B.2 available on Dryad). Second, the algorithm of GATK requires a priori information about known variable sites to recalibrate base quality scores (DePristo et al. 2011). However, this information is often unavailable for studies involving several species (Schubert et al. 2014). If this is the case, we can then expect that the species close to the reference might benefit from a better recalibration than the species far from the reference. Contrary to GATK, ATLAS has the alternative of base quality score recalibration based on invariant sites. Unless phylogenetically known invariant sites are provided, ATLAS will assume that all provided sites are invariant. The benefit of this type of recalibration model is that no assumption needs to be made about the underlying allele, just about the status as invariant. This fact might also explain the shorter computation time needed for ATLAS (Table 1).

In this study, we based our first scaling on the divergence that can be observed at the genus level for the clownfish phylogeny (Litsios et al. 2014). We are aware that, for other organisms, divergence at the genus level will be different, which is why we tested several additional scalings (Supplementary Figs. 1 and SI-B.4 available on Dryad). We did not need to test larger divergence values because the performance of GATK on the trees simulated here already showed a rapid decrease with increasing distance to the reference. Given this trend, we expect this underperformance to be even greater for larger phylogenetic trees with greater divergence values among its species.

Nevertheless, we would like to make clear that current NGS software (such as GATK) does provide very good tools for NGS processing. It is only the genotype-calling subroutine of GATK that needs to be used carefully, and applied only to population-level studies with a priori knowledge of variable sites, or use the ATLAS genotype caller instead. We conclude that, for multispecies NGS studies, primary efforts should be taken towards building more reference genomes, rather than sequencing more species without a reference.

Supplementary Material

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.fn2z34ts3.

Acknowledgments

We would like to thank the Associate Editor and three anonymous reviewers for the valuable comments that greatly improved the quality of this article. N.S. received

funding from the Swiss National Science Foundation (310030-185223) and from the University of Lausanne. We thank the computing infrastructure of the University of Lausanne for access to computing clusters. Finally, thanks to Emily Williams for insightful suggestions that improved the writing of this article.

References

Bateman R.M., Sramkó G., Paun O. 2018. Integrating restriction site-associated DNA sequencing (RAD-seq) with morphological cladistic analysis clarifies evolutionary relationships among major species groups of bee orchids. Ann. Bot. 121:85–105.

Blischak P.D., Kubatko L.S., Wolfe A.D. 2018. SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. Bioinformatics 34:407–415.

Bragg J.G., Potter S., Bi K., Moritz C. 2016. Exon capture phylogenomics: efficacy across scales of divergence. Mol. Ecol. Resour. 16:1059–1068.

Brandrud M.K., Baar J., Lorenzo M.T., Athanasiadis A., Bateman R.M., Chase M.W., Hedrén M., Paun O. 2020. Phylogenomic relationships of diploids and the origins of allotetraploids in *Dactylorhiza* (Orchidaceae). Syst. Biol. 69:91–109.

Burress E., Alda F., Duarte A., Loureiro M., Armbruster J., Chakrabarty P. 2018. Phylogenomics of pike cichlids (Cichlidae: Crenicichla): the rapid ecological speciation of an incipient species flock. J. Evol. Biol. 31:14–30.

Chakrabarty P., Faircloth B.C., Alda F., Ludt W.B., Mcmahan C.D., Near T.J., Dornburg A., Albert J.S., Arroyave J., Stiassny M.L., et al. 2017. Phylogenomic systematics of ostariophysan fishes: ultraconserved elements support the surprising non-monophyly of Characiformes. Syst. Biol. 66:881–895.

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., De-Pristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R., 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

DePristo M.A., Banks E., Poplin R., Garimella K.V., Maguire J.R., Hartl C., Philippakis A.A., Del Angel G., Rivas M.A., Hanna M., McKenna A., Fennell T.J., Kernytsky A.M., Sivachenko A.Y., Cibul-skis K., Gabriel S.B., Altshuler D., Daly M.J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491.

dos Santos G., Schroeder A.J., Goodman J.L., Strelets V.B., Crosby M.A., Thurmond J., Emmert D.B., Gelbart W.M., FlyBase Consortium. 2015. FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res. 43:D690–D697.

Duchen P., Renner S.S. 2010. The evolution of *Cayaponia* (Cucurbitaceae): Repeated shifts from bat to bee pollination and long-distance dispersal to Africa 2–5 million years ago. Am. J. Bot. 97:1129–1141.

Fitz-Gibbon S., Hipp A.L., Pham K.K., Manos P.S., Sork V.L. 2017. Phylogenomic inferences from reference-mapped and de novo assembled short-read sequence data using RADseq sequencing of California white oaks (*Quercus* section *Quercus*). Genome 60:743–755.

Grummer J.A., Morando M.M., Avila L.J., Sites Jr J.W., Leaché A.D. 2018. Phylogenomic evidence for a recent and rapid radiation of lizards in the Patagonian *Liolaemus fitzingerii* species group. Mol. Phylogenet. Evol. 125:243–254.

Hasegawa M., Kishino H., Yano T.-A. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Heckenhauer J., Paun O., Chase M.W., Ashton P.S., Kamariah A., Samuel R. 2019. Molecular phylogenomics of the tribe Shoreeae (Dipterocarpaceae) using whole plastid genomes. Ann. Bot. 123:857–865.

Helmstetter A.J., Kamga S.M., Bethune K., Lautenschläger T., Zizka A., Bacon C.D., Wieringa J.J., Stauffer F., Antonelli A., Sonké B., et al. 2020. Unraveling the phylogenomic relationships of the most diverse African palm genus *Raphia* (Calamoideae, Arecaceae). Plants 9:549.

Hudson R. R. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Hulsey C.D., Zheng J., Faircloth B.C., Meyer A., Alfaro M.E. 2017. Phylogenomic analysis of Lake Malawi cichlid fishes: further evidence that the three-stage model of diversification does not fit. Mol. Phylogenet. Evol. 114:40–48.

Hulsey C.D., Zheng J., Holzman R., Alfaro M.E., Olave M., Meyer A. 2018. Phylogenomics of a putatively convergent novelty: did hypertrophied lips evolve once or repeatedly in Lake Malawi cichlid fishes? BMC Evol. Biol. 18:179.

Hwang S., Kim E., Lee I., Marcotte E.M. 2015. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci. Rep. 5:17875.

Knaus B.J., Grünwald N.J. 2017. vcfR: a package to manipulate and visualize variant call format data in R. Mol. Ecol. Res. 17:44–53.

Kozlov A.M., Darriba D., Flouri T., Morel B., Stamatakis A. 2019. RAxML-ng: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35:4453–4455.

Kreuzer M., Howard C., Pendry C.A., Adhikari B., Hawkins J.A. 2019. Phylogenomic approaches to DNA barcoding of herbal medicines: developing clade-specific diagnostic characters for Berberis. Front. Plant Sci. 10:586.

Kuhner M. K., Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.

Kumar V., Lammers F., Bidon T., Pfenninger M., Kolter L., Nilsson M.A., Janke A. 2017. The evolutionary history of bears is characterized by gene flow across species. Sci. Rep. 7:46487.

Li H. 2013. wgsim: read simulator for next generation sequencing. Available from: http://github.com/lh3/wgsim (accessed in 2019 and 2020).

Li H., Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26:589–595.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.

Lima M.G., Silva-Júnior J.d.S.E., Èerny D., Buckner J.C., Aleixo A., Chang J., Zheng J., Alfaro M.E., Martins A., Di Fiore A., et al. 2018. A phylogenomic perspective on the robust capuchin monkey (*Sapajus*) radiation: first evidence for extensive population admixture across South America. Mol. Phylogenet. Evol. 124:137–150.

Link V., Kousathanas A., Veeramah K., Sell C., Scheu A., Wegmann D. 2017. ATLAS: analysis tools for low-depth and ancient samples. bioRxiv 105346.

Lischer H.E., Excoffier L., Heckel G. 2014. Ignoring heterozygous sites biases phylogenomic estimates of divergence times: implications for the evolutionary history of *Microtus* voles. Mol. Biol. Evol. 31:817–831.

Litsios G., Pearman P.B., Lanterbecq D., Tolou N., Salamin N. 2014. The radiation of the clownfishes has two geographical replicates. J. Biogeogr. 41:2140–2149.

Loiseau O., Olivares I., Paris M., de La Harpe M., Weigand A., Koubinova D., Rolland J., Bacon C.D., Balslev H., Borchsenius F., et al. 2019. Targeted capture of hundreds of nuclear genes unravels phylogenetic relationships of the diverse Neotropical palm tribe Geonomateae. Front. Plant Sci. 10:864.

Marcionetti A., Rossier V., Roux N., Salis P., Laudet V., Salamin N. 2019. Insights into the genomics of clownfish adaptive radiation: genetic basis of the mutualism with sea anemones. Genome Biol. Evol. 11:869–882.

McCormack J.E., Hird S.M., Zellmer A.J., Carstens B.C., Brumfield R.T. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol. 66:526–538.

Moura A.E., Shreves K., Pilot M., Andrews K.R., Moore D.M., Kishida T., M ¨ oller L., Natoli A., Gaspari S., McGowen M., et al. 2020. Phylogenomics of the genus *Tursiops* and closely related Delphininae reveals extensive reticulation among lineages and provides inference about eco-evolutionary drivers. Mol. Phylogenet. Evol. 146:106756.

Nielsen R., Paul J.S., Albrechtsen A., Song Y.S. 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12:443–451.

Nobre L.L.d.M., dos Santos J.D.O., Leite R., Almeida C. 2018. Phylogenomic and single nucleotide polymorphism analyses revealed the hybrid origin of *Spondias bahiensis* (family Anacardiaceae): de novo genome sequencing and comparative genomics. Genet. Mol. Biol. 41:878–883.

Olvera-Mendoza E.I., Godden G.T., Montero-Castro J.C., Porter J.M., Lara-Cabrera S.I. 2020. Chloroplast and nuclear ribosomal cistron phylogenomics in a group of closely related sections in *Salvia* subg. *Calosphace*. Braz. J. Bot. 43:177–191.

Ottenburghs J., Megens H.-J., Kraus R.H., Madsen O., van Hooft P., van Wieren S.E., Crooijmans R.P., Ydenberg R.C., Groenen M.A., Prins H.H. 2016. A tree of geese: a phylogenomic perspective on the evolutionary history of True Geese. Mol. Phylogenet. Evol. 101:303–313.

Portik D.M., Smith L.L., Bi K. 2016. An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). Mol. Ecol. Resour. 16:1069–1083.

Rambaut A., Grass N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics 13:235–238.

Schubert M., Ermini L., Der Sarkissian C., Jónsson H., Ginolhac A., Schaefer R., Martin M.D., Ferna ìndez R., Kircher M., McCue M., et al. 2014. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat. Protocols 9:1056.

Skipwith P.L., Bi K., Oliver P.M. 2019. Relicts and radiations: phylogenomics of an Australasian lizard clade with east Gondwanan origins (Gekkota: Diplodactyloidea). Mol. Phylogenet. Evol. 140:106589.

Stadler T. 2011. Simulating trees with a fixed number of extant species. Syst. Biol. 60:676–684.

Van der Auwera G.A., Carneiro M.O., Hartl C., Poplin R., Del Angel G., Levy-Moonshine A., Jordan T., Shakir K., Roazen D., Thibault J., et al. 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr. Protocols Bioinformatics 43:11–10.

Wang M., Zhang L., Zhang Z., Li M., Wang D., Zhang X., Xi Z., Keefover-Ring K., Smart L.B., DiFazio S.P., et al. 2020. Phylogenomics of the genus *Populus* reveals extensive interspecific gene flow and balancing selection. New Phytol. 225:1370–1382.

Wang X., Ye X., Zhao L., Li D., Guo Z., Zhuang H. 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (Poaceae: Bambusoideae). Sci. Rep. 7:1–11.

Yan Z., Martin S.H., Gotzek D., Arsenault S.V., Duchen P., Helleu Q., Riba-Grognuz O., Hunt B.G., Salamin N., Shoemaker D., et al. 2020. Evolution of a supergene that regulates a trans-species social polymorphism. Nat. Ecol. Evol. 4:240–249.