MethodsX 7 (2020) 100892

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/mex

Optimization of Genotype by Sequencing data for phylogenetic purposes

L.O. Loureiro[a,b,*], M.D. Engstrom[a,b], B.K. Lim[b]

[a] University of Toronto, Canada
[b] Royal Ontario Museum, Canada

A B S T R A C T

• Herein we propose a framework for assembling and analyzing Genotype by Sequencing (GBS) data to better understand evolutionary relationships within a group of closely related species using the mastiff bats (*Molossus*) as our model system. Many species within this genus have low-levels of genetic variation within and between morphologically distinct species, and the relationships among them remain unresolved using traditional Sanger sequencing methods. Given that both *de novo* and reference genome pipelines can be used to assemble next generation sequences, and that several tree inference methodologies have been proposed for single nucleotide polymorphism (SNP) data, we test whether different alignments and phylogenetic approaches produce similar results. We also examined how the process of SNP identification and mapping can affect the consistency of the analyses. Different alignments and phylogenetic inferences produced consistent results, supporting the GBS approach for answering evolutionary questions on a macroevolutionary scale when the genetic distance among phenotypically identifiable clades is low. We highlight the importance of exploring the relationships among groups using different assembly assumptions and also distinct phylogenetic inference methods, particularly when addressing phylogenetic questions in genetic and morphologically conservative taxa.

• The method uses the comparison of several filter settings, alignments, and tree inference approaches on Genotype by Sequencing data.
• Consistent results were found among several approaches.
• The methodology successfully recovered well supported species boundaries and phylogenetic relationships among species of mastiff bats not hypothesized by previous methods.

A R T I C L E I N F O
Method name: Genotype by Sequencing phylogenetic analyses optimization
Keywords: Genotype by Sequencing, Evolutionary relationships, Bats, Molossidae
Article history: Received 25 November 2019; Accepted 3 April 2020; Available online 20 April 2020

DOI of original article: 10.1016/j.ympev.2019.106690
* Corresponding author at: University of Toronto, Canada.
E-mail address: livia.loureiro@sickkids.ca (L.O. Loureiro).

https://doi.org/10.1016/j.mex.2020.100892

Specification Table

Subject Area:	Biochemistry, Genetics and Molecular Biology
More specific subject area:	*Evolution, phylogeny, genetics*
Method name:	Genotype by Sequencing phylogenetic analyses optimization
Name and reference of original method:	*NA*
Resource availability:	*NA*

*Method details

Background

Advances in genomics technology have allowed the generation of large numbers of molecular markers across the genome, which increases sample sizes and provides additional data to help resolve interpretation of the ecology and evolution of traditionally poorly understood species groups [1]. One of these methods is Genotyping by Sequencing (GBS), which involves sequencing genomic regions flanking restriction sites. Using GBS, many sequences of short length are obtained, vastly increasing the size of the overall data set in comparison to traditional Sanger methods. This technique provides sequence data for thousands of single nucleotide polymorphisms (SNPs), allowing the detection of small, but consistent genetic variation among genetically similar groups not revealed by standard gene sequencing approaches. GBS has been successfully used in studies of population genetics [2], phylogenetic analysis [3,4] and phylogeography [5,6].

In this context, we propose a framework for assembling and analyzing GBS data to better understand evolutionary relationships among species of mastiff bats (*Molossus*), a genus with a complex taxonomic history and low levels of genetic variation [7]. Herein, we test how four different filtering settings affect the accuracy and consistency of our data. Given that both *de novo* and reference genome pipelines are often used to assemble next generation sequencing (NGS) data, and that several tree inference methodologies have been proposed for SNP data, we also test if different alignments and phylogenetic approaches produce similar results. These data offer a useful framework for other comparative studies of ecology and evolution using the GBS approach.

Methodology

We obtained tissues from a total of 189 specimens including all the currently recognized species of *Molossus* [8] and representatives of two other species of molossid bats, *Promops centralis* and *Eumops auripendulus*, that were used as outgroups following Ammerman et al. [9] and Gregorin and Cirranello (2016) [10]. Genomic DNA was isolated using a Qiagen DNeasy extraction kit (Qiagen, Inc. Valencia, CA, USA) following the manufacturer's instructions. Genomic DNA quality was checked by visual inspection on agarose gels and quantified using a Nanodrop spectrophotometer (Nanodrop Technologies). Thirty microlitres of high quality (>100 ng/ul) DNA per sample was used for the library preparation. We submitted the samples to the Cornell Institute of Genomic Diversity (IGD) to obtain SNP datasets through the GBS approach following the protocol described by Elshire et al. [11]. All libraries were sequenced on an Illumina HiSeq 2000.

Two different approaches were used to process the data and test the accuracy and precision of the results. Raw sequence files from Illumina were converted into individual genotypes using the Discovery and the Universal Network-Enabled Analysis Kit (UNEAK) pipelines, available as part of the TASSEL 3.0 software [12] (Fig. 1). Both pipelines trim the sequences to a length of 64 basepairs (bp) after the barcode and discard shorter reads. Identical reads are clustered into tags, and all identical tags are merged. The Discovery pipeline uses a reference genome to align the tags [13], and for this alignment we used the genome of *Myotis brandtii* in the related family Vespertilionidae [14,15]. Unfortunately, there were no genomes available for the family Molossidae in the time of the analyses. The UNEAK pipeline uses a *de novo* approach, and the alignment to the reference genome is replaced with a pairwise alignment of tags, in which tag pairs with a 1 bp mismatch are considered as candidate SNPs [16]. The *de novo* alignment assumes homogeneity of rates across

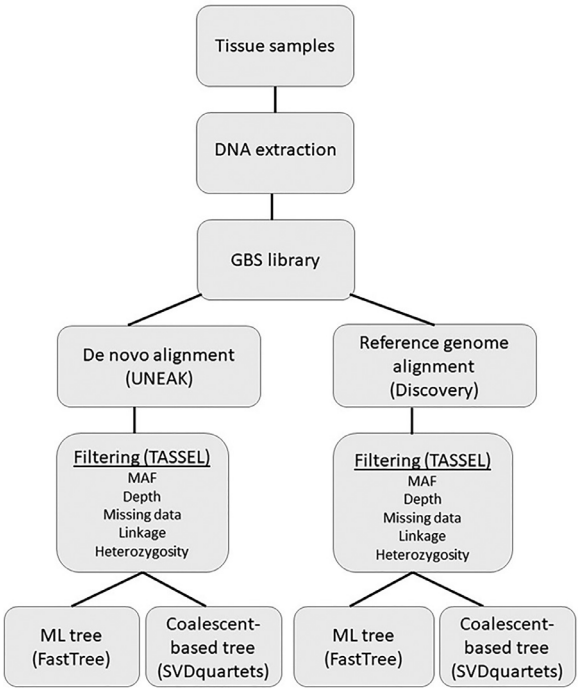

Fig. 1. Flow chart of the different methodologies used in the analyses. Acronyms are discussed in the text.

sequenced tags [17] and although it has been proven to produce highly supported trees (Rojas et al., 2003; Wagner et al., 2012), there is still a potential source of inaccuracy in the homology assumption that is determined based on the similarity between tags. The sequences produced by this method are short (64 bp in the GBS approach), and without a reference genome, sequences from different individuals might be mistakenly assigned as homologous given their base similarity. The use of the reference genome is expected to decrease errors in the inference of orthologous sequence because it allows genomic mapping of loci, and partitioning analysis by linkage groups, coding and non-coding positions, and other genetic subregions, all of which could potentially improve phylogenetic accuracy [17,18]. Unfortunately, a clear limitation in our study is that no genome was available for any species of Molossidae in the time of the analyses. To address this bias, we examined the data generated using a reference genome and a *de novo* alignment separately.

We assessed the quality of the paired GBS-tags using FastQC. The likelihoods of the possible genotypes were estimated for both pipelines, and the Genotype quality (GQ), which is the difference between the most likely and the second most likely genotype, was calculated also using FastQC. The data were then imported into TASSEL where samples were demultiplexed and filtered. The amount of missing data acceptable for phylogenetic estimates still lacks a consensus. The identification of minimum values for these filters is still debatable and may ultimately depend on the dataset in hand and the question that needs to be answered [18,19,22]. Some studies have been very conservative, removing all loci that are missing for any taxon in the dataset (McCormack et al., 2012; [20]), but others included loci that were not present in a large proportion of samples [5,6]. The amount of missing data allowed in the SNPs matrix can affect further analyses, however, some authors argue that larger matrices, even those containing a large amount of missing data, may provide greater phylogenetic accuracy than smaller ones [3,5,19]. Therefore, we tested if the amount of missing data removed would affect our final analyses.

Similarly, the optimum value for minimum allele frequency (MAF) is not universally agreed upon and there is a trade-off between the use of MAF and the loss of rare alleles. The increase of the MAF value may cause an under-calling of heterozygotes with the loss of biological information, instead of the removal of sequencing errors [21]. Kim et al., [22] argued that for rare SNPs (e. g. MAF <0.01) differentiation between sequencing errors and true rare alleles is difficult, and they should be discarded. Lincky and Battey (2019) [23] showed that highly accurate population inferences are reached when rare alleles are included (MAF 2% to 8%), but decay in accuracy when only common alleles were included. Here, we tested how different MAF values varying from 0.01 to 0.06 would affect the number of SNPS in our data sets and the accuracy and consistency of estimation of phylogenetic relationships.

Low depth sequences are sequences recovered by relatively few reads and, if not removed, they might lead to serious bias (2014) [24]. Sequences with very low depths might increase the probability of calling false SNPs due to PCR and sequencing errors and must be removed [25]. However, excess removal might also reduce the amount of informative DNA sequences [26]. In our data set, we first estimated the mean depth for the reference and *de novo* alignments using VCFtools, and then we simulated how the depth could affect the number of SNPs in both alignments by removing sequences with lower and higher depths than the mean in the TASSEL software. For this simulation we looked at the number of SNPs removed and the agreement between final topologies using the optimum MAF value found for each dataset.

GBS sequences are short, which decreases the chance of intragenic recombination. However, if the SNPs are closely linked on the same chromosome, they might not be independent. Some studies suggest mapping the SNPs in a reference genome to confront this issue [27,28]. The lack of an available closely related reference genome to *Molossus* and the low percentage of tags that aligned to the *Myotis* genome (2.5%), precluded use of this method for our dataset. Each tag produced by the GBS method has 64 bp, and therefore we have tried to overcome this problem by removing SNPs that were separated by less than 128 bp in the genome.

To test for divergence in tree inference approaches we removed invariant sites in both alignments and reconstructed the phylogenetic relationships within the genus *Molossus* using the Maximum Likelihood approach (ML) implemented in FastTree [29] with a GTR + gamma model of nucleotide evolution estimated by Partition Finder 1.0.1 [30] (Fig. 1). In the ML method, the alignment for each locus is aggregated in a supermatrix and a species tree is estimated under the assumption that all sites evolve identically and independently [31]. The relationships among clades were also investigated through a coalescence approach, which accounts for differences in genealogical histories of individual loci using the program SVDquartets [32] implemented in PAUP 4.0 [33] (Fig. 1). The SVDquartets is a coalescent model that uses unlinked SNPs to infer the quartet tree for every four species, and then combines all the subtrees into a species tree [32]. Pettengill et al. [18] found that some programs produce more reliable trees for NGS, even when working with the same phylogenetic assumptions (e.g. ML), but the difference in topologies is usually in poorly supported nodes. Four independent runs were conducted to access topological convergence, each including 500 bootstrap replicates and exhaustive quartet sampling.

Method validation

After merging the Illumina paired-end sequences, more than nine million tags remained. The *de novo* pipeline identified 418,810 SNPs after error curation in the standard network error remover of TASSEL. The alignment using the reference genome discarded 96.6% of tags that did not align with the *Myotis* genome and 0.9% of the tags that aligned multiple times, keeping only 2.5% of the short length sequences. After removal of invariant sites, the reference genome alignment produced 55,350 SNPs. Of the 189 specimens included in the Illumina library, 23 were discarded in the *de novo* pipeline and 16 were discarded in the reference-based pipeline because they had low numbers of raw reads and few loci (>90% of missing data), probably because of low DNA quality. The quality of the isolated DNA depends on the quality of the tissue sample. Fresh tissues produce the highest DNA yield and quality and samples should be stored under conditions that preserve DNA integrity. In addition, repeated freezing and thawing of frozen tissues might reduce the size and quality of

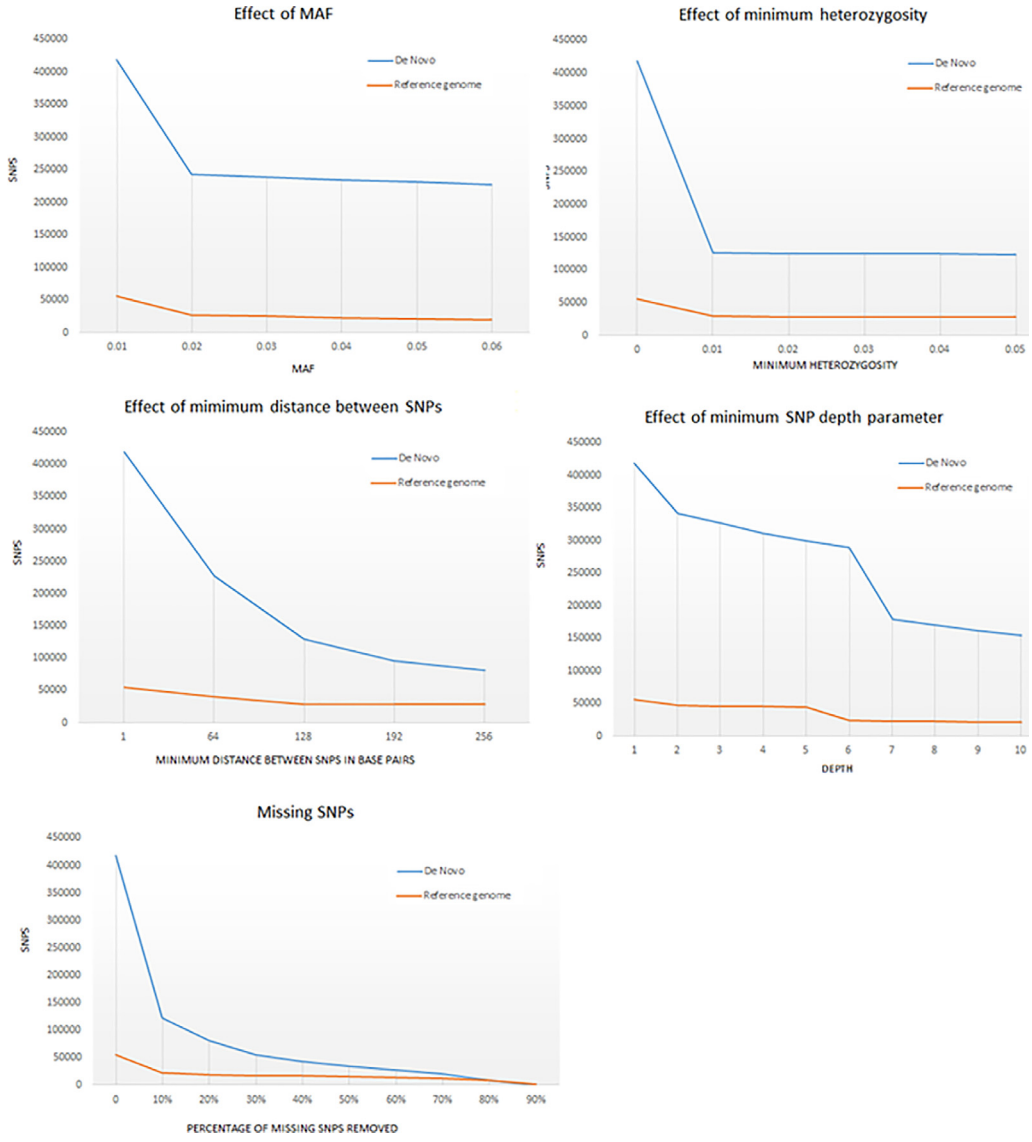

Fig. 2. Comparison of different parameters in the number of SNPs obtained with the *de novo* alignment and the reference genome alignment for *Molossus*.

DNA. For compromised samples, the initial concentration was lower than required by GBS library preparation (>100 ng/ul), and multiple DNA concentration procedures were required to archive the required DNA quality. After this procedure, thirty microlitres of each sample were sent to sequence, but the remaining DNA aliquot were less than thirty microlitres per individual, too low for repeating the GBS library preparation.

Both alignments behaved similarly for all filters. In our dataset, the removal of 10% missing SNPs decreased the total number of markers by less than half. The number of SNPs decreased further when removing 20% to 90% of missing data (Fig. 2). Our final trees did not change when more than 50% of missing data was removed, but loss of support was observed in some clades when allowing for

missing data higher than 50%. Therefore, if the SNPs were missing in 50 % or less of the samples, we kept the marker, but if the marker was present in less than 50% of the samples they were discarded.

Minor alleles are expected to occur in a smaller percentage of the dataset, usually in less than 10% of the population [21,22,34,35]. Our simulations showed that the number of SNPs removed with MAF lower than 2% did not affect our the number of SNPs significantly. For either alignments, only a small number of SNPs was discarded (<2% for both pipelines) with MAF >0.02, which is consistent with true polymorphic sites. However, large parts of the dataset were excluded with the removal of MAF < 0.02 (58% with the *de novo* and 49% with the reference genome approaches) (Fig. 2), which is more likely to represent sequence errors.

Variation in DNA quality can affect genotype accuracy. Low DNA quality, or concentration, often affect SNPs identification, and the genotype heterozygosity rate can be used as a measure of the accuracy of the samples [36]. In our dataset, the same pattern found for MAF occurred for the heterozygosity filter, but the heterozygosity filter had a higher impact in the *de novo* alignment in comparison with the reference genome data set. When set to 0.01, this filter discarded approximately two-thirds of the SNPS in the *de novo* alignment and half of those in the reference genome alignment (Fig. 2).

More than half of the SNPS were closely aligned in the genome, which could indicate linked loci. We removed SNPs that were 128 base pairs apart, twice the length of a single tag, to decrease linked SNPs in the dataset (Fig. 2). Linkage analyses indicate that only nine pairs of linked SNPs remained after the removal of SNPs 128 bp apart in the de novo alignment and only two pairs in the reference genome alignment, which were removed manually for further analyses. Although the removal of SNPs with distances higher than 128 bp apart discarded a larger percentage of potentially linked SNPs, the increase of this filter value also removed more than 34,000 non linked SNPs in the *de novo* alignment and more than 300 in the reference genome alignment. The removal of linked SNPs did not affect the other parameters of the data sets (e. g. mean depth) and species relationships compared to the linked dataset. However, the relationships within some populations changed slightly, which might indicate a higher effect on population structure analyses.

The average depth in the *de novo* and reference genome alignments were seven and six respectively, which is consistent with our analysis that shows a plateau in the number of SNPs after setting missing data depth less than six and seven for each respective data set (Fig. 2). The depths were calculated using the best parameter for MAF (0.02) in VCFtools. For both *de novo* and reference genome pipelines, the removal of sequences with less than the mean depth caused a considerable decrease in SNPs (44% in the *de novo* and 43% in the reference genome pipeline), and this high proportion of low depth tags suggests a high degree of uncertainty about the homology of those sequences [18]. However, when sequences with depth higher than the average were set as missing data, the number of SNPs included was not greatly affected but the removal of those sequences led to the loss of rare variants. Indeed, when values larger than the mean depth were removed, de novo and reference genome topologies lost support for some clades, but branching sequences were not changed. In the final analyses, only tags with lower depth than the mean were removed.

According to the previous results, the sequences were filtered for missing data < 50%, minor allele frequency (MAF) >0.02, heterozygosity >0.01, and depth coverage lower than seven and six for the *de novo* and reference alignment, respectively. After data filtering, the *de novo* pipeline yielded 71,801 SNPs and the reference genome pipeline yielded 27,323 SNPs. To remove unlinked and uncertain SNPs, sites with more than 50% of missing data were discarded as were sites less than 128 bp apart. The final data set with unlinked SNPs had 29,448 SNPs for the *de novo* pipeline and 15,569 SNPs for the reference genome dataset. In addition, we also used more and less conservative filtering values to test for differences in accuracy and consistency in tree topologies.

Validation results

The ML and SVDquartets trees with both the *de novo* and the reference genome alignments are congruent when the optimum filtering settings are used, recovering the same clades and the same relationships among species [7,8]. However, when more conservative filter values, compared to those optimum values found in previous analysis, are used (e. g. MAF>0.02; heterozygosity >0.01), clades

within *Molossus* start to lose support; although, the relationships among them do not change. This outcome is consistent with loss of true rare polymorphic sites, which should not be removed from the dataset. The removal of markers with less than 50% of missing data did not change supports in the phylogenetic trees. However, if markers with more than 50% of missing data are removed and less conservative filtering setting values are used (MAF<0.02; heterozygosity <0.01), the topologies between de novo and reference genome alignment lose agreement and relationship among some clades within *Molossus* are no longer supported by both alignments. These results are consistent with incorporation of sequencing errors. Using the optimum filtering values (Fig. 2), the best scoring ML and SVDquartets trees contained most nodes with 100% bootstrap support, and a few nodes with lower support, which were always above 80% [7,8]. Although both alignment approaches produced the same species-level trees, there were minor shifts in relationships within well supported clades at a population level [7,8]. Although the two inference methods used in this study have different assumptions and work with different algorithms, we still recovered the same topology at the species level, supporting the assessment of phylogenetic relationships [7,8].

The greater power of next-generation sequencing approaches compared to Sanger methods for answering phylogenetic questions is well established [37,38], but there are only a few studies comparing the concordance between *de novo* and reference-based alignments. In our study, relationships within terminal clades with low bootstrap support were affected by the choice of the pipeline, which could have resulted from the difference in number of SNPs retained from both alignments. However, the use of the reference genome in the alignment does not seem to be essential for recovering a robust overall phylogenetic tree, since both phylogenies, *de novo* and with the reference genome, were similar at the species level when optimal filtering settings values were used.

We highlight the importance of exploring the relationships among groups using different assembly assumptions and also distinct phylogenetic inference methods, particularly when addressing phylogenetically conservative groups. All models of molecular evolution are a simplification of the actual evolutionary process, and the inappropriate choice of filters during alignment or tree inference can lead to systematic bias in the phylogenetic reconstruction [39,40]. Therefore, we emphasize the value of carefully optimizing SNPs filters to minimize the effect that missing data, independent SNPs, and incorrect inference of homology could have on the results.

Acknowledgments

This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) (9 99999.011880/2013-09). Neotropical fieldwork has been primarily funded by the Royal Ontario Museum with additional financial support in Ecuador by Ecuambiente Consulting Group and in Guyana by Conservation International and funding through the Academy of Natural Sciences, Philadelphia. We thank the following curators and collection support staff that provided access or loaned specimens: R. Gregorin (UFLA), F. A. Perini (UFMG), B. D. Patterson (FNMH), C. J. Conroy (MVZ), M. Campbell (MSB), B. S. Coyner (Sam Noble Museum), N. B. Simmons (AMNH), H. J. Garner (TTU), C. Lopez-Gonzalez (Instituto Politécnico Nacional, Mexico City), J. Juste (CSIC), A. L. Gardner (NMNH/USMN), M. de Vivo and J. G. Barros (MZUSP), C. G. Costa (MCN- PUC Minas), G. Graciolli and M. Bordignon (UFMS), E. Morielle-Versute (UNESP), L. Peracchi (UFRRJ), and J. A. Oliveira (MNRJ). We also thank Oliver Haddrath for providing constructive feedback on this manuscript.

Declaration of Competing Interest

Authors declare no conflict of interest.

References

[1] H.C. Rowe, S. Renaut, A. Guggisberg, RAD in the realm of next-generation sequencing technologies, Molec. Ecol. 20 (2011) 3499–3502, doi:10.1111/j.1365-294X.2011.05197.x.

[2] P.A. Hohenlohe, P.C. Phillips, W.A. Cresko, Using population genomics to detect selection in natural populations: key concepts and methodological considerations, Int. J. Plant Sci. 9 (2010) 1059–1071.

[3] B.E.R. Rubin, R.H. Ree, C.S. Moreau, Inferring phylogenies from RAD sequence data, PLoS One 7 (1–12) (2012), doi:10.1371/journal.pone.0033394.
[4] M.A Cronin, A. Cánovas, D.L. Bannasch, A.M. Oberbauer, J.F. MeDrano, E. Ostrander, Single nucleotide polymorphism (SNP) variation of wolves (*Canis lupus*) in Southeast Alaska and comparison with wolves, dogs, and Coyotes in North America., J. Hered. 106 (2015) 26–36, doi:10.1093/jhered/esu075.
[5] C.E. Wagner, I. Keller, S. Wittwer, O.M. Selz, S. Mwaiko, L. Greuter, A. Sivasundar, O. Seehausen, Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation, Mol. Ecol. 22 (2013) 787–798, doi:10.1111/mec.12023.
[6] K.J. Emerson, C.R. Merz, J.M. Catchen, P.A. Hohenlohe, W.A. Cresko, W.E. Bradshaw, C.M. Holzapfel, Resolving postglacial phylogeography using high-throughput sequencing, Proc. Nat. Acad. Sci. 107 (37) (2010) 16196–16200, doi:10.1073/pnas.1006538107.
[7] L.O. Loureiro, M.D. Engstrom, B.K. Lim, Single Nucleotide Polymorphisms (SNPs) provide unprecedented resolution of species boundaries, phylogenetic relationships, and genetic diversity in the mastiff bats (*Molossus*), Molec. Phylog. Evol. 143 (2019) 106690, doi:10.1016/j.ympev.2019.106690.
[8] L.O. Loureiro, M.D. Engstrom, B.K. Lin, Genotype by Sequencing data in the evolutionary relationships of the mastiff bat (Chiroptera, Molossidae, *Molossus*), Data Brief (2020).
[9] L.K. Ammerman, D.N. Lee, T.M. Tipps, First molecular phylogenetic insights into the evolution of free-tailed bats in the subfamily Molossinae (Molossidae, Chiroptera), J. Mammal. 93 (2012) 12–28, doi:10.1644/11-mamm-a-103.1.
[10] R. Gregorin, A. Cirranello, Phylogeny of Molossidae Gervais (Mammalia: Chiroptera) inferred by morphological data, Cladistics 32 (2016) 2–35, doi:10.1111/cla.12117.
[11] R.J. Elshire, J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, S.E. Mitchell, A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species, PLoS One 6 (2011) 1–10, doi:10.1371/journal.pone.0019379.
[12] P.J. Bradbury, Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, E.S. Buckler, TASSEL: Software for association mapping of complex traits in diverse samples, Bioinformatics 23 (2007) 2633–2635, doi:10.1093/bioinformatics/btm308.
[13] J.C. Glaubitz, T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline, PLOS ONE 9 (2014) e90346.
[14] J.J. Shi, D.L. Rabosky, Speciation dynamics during the global radiation of extant bats, Evolution 69 (2015) 1528–1545.
[15] E.C. Teeling, G. Jones, S.J. Rossiter, et al., Phylogeny, genes, and Hearing: Implications for the evolution of echolocation in bats, in: B. Fenton, et al. (Eds.), Bat Bioacoustics, Springer Science + Business Media, New York, 2016, pp. 25–64.
[16] F. Lu, R. Elshire, E.S. Buckler, D.E. Costich, J. Glaubitz, J.H. Cherney, M.D. Casler, A.E. Lipka, Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol, PLoS Genet 9 (2013) e1003215, doi:10.1371/journal.pgen.1003215.
[17] B. Rannala, Z. Yang, Phylogenetic Inference Using Whole Genomes, Annu. Rev. Genomics Hum. Genet. 9 (2008) 217–231, doi:10.1146/annurev.genom.9.081307.164407.
[18] J.B. Pettengill, Y. Luo, S. Davis, Y. Chen, N. Gonzalez-Escalona, A. Ottesen, H. Rand, M.W. Allard, E. Strain, An evaluation of alternative methods for constructing phylogenies from whole genome sequence data: a case study with *Salmonella*, Peer J. (2014) 2e620, doi:10.7717/peerj.620.
[19] H. Huang, L. Knowles, Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of rad sequences, Syst. Biol. 65 (2016) 357–365, doi:10.1093/sysbio/syu046.
[20] A.J Zellmer, M.M. Hanes, S.M. Hird, B.C. Carstens, Deep phylogeographic structure and environmental differentiation in the carnivorous plant *Sarracenia alata*, Syst. Biol. 61 (2012) 763–777, doi:10.1093/sysbio/sys048.
[21] S. Ni, M. Stonekin, Improvement in detection of minor alleles in next generation sequencing by base quality recalibration, BMC Genom. 17 (2016) 1–12, doi:10.1186/s12864-016-2463-2.
[22] J. Kim, S. Park, J. Yang, et al., SNPs in axon guidance pathway genes and susceptibility for Parkinson's disease in the Korean population, J. Hum Genet. 56 (2011) 125–129, doi:10.1038/jhg.2010.130.
[23] E. Lincky, C.J. Battey, Minor allele frequency thresholds strongly affect population structure inference with genomic datasets, Mol. Ecol. Resour 19 (3) (2019) 639–647, doi:10.1111/1755-0998.12995.
[24] K. Song, J. Ren, G. Reinert, M. Deng, M.S. Waterman, F. Sun, New developments of alignment-free sequence comparison: Measures, statistics and next-generation sequencing, Brief. Bioinf. 15 (2014) 343–353, doi:10.1093/bib/bbt067.
[25] A.T.O. Melo, R.I. Bartaula, Hale GBS-SNP-CROP: A reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data, BMC Bioinf. 17 (2016) 1–15, doi:10.1186/s12859-016-0879-y.
[26] H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, Bioinformatics 27 (2011) 2987–2993, doi:10.1093/bioinformatics/btr509.
[27] T. Kawakami, L. Smeds, N. Backström, A. Husby, A. Qvarnström, C.F. Mugal, P. Olason, H. Ellegren, A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution, Mol. Ecol. 23 (2014) 4035–4058, doi:10.1111/mec.12810.
[28] C. Blair, C.R. Campbell, A.D. Yoder, Assessing the utility of whole genome amplified DNA for next-generation molecular ecology, Mol. Ecol. Resour. 15 (2015) 1079–1090, doi:10.1111/1755-0998.12376.
[29] M.N. Price, P.S. Dehal, A.P. Arkin, Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix, Mol. Biol. Evol. 26 (2009) 1641–1650, doi:10.1093/molbev/msp077.
[30] R. Lanfear, B. Calcott, S.W.Y. Ho, S. Guindon, Partition finder: combined selection of partitioning schemes and substitution models for phylogenetic analyses, Mol. Biol. Evol. 29 (6) (2012) 1695–1701, doi:10.1093/molbev/mss020.
[31] J. Chou, A. Gupta, S. Yaduvanshi, R. Davidson, M. Nute, S. Mirarab, T. Warnow, A comparative study of SVDquartets and other coalescent-based species tree estimation methods, BMC Genom. 16 (2015) 1–11, doi:10.1186/1471-2164-16-S10-S2.
[32] J. Chifman, L. Kubatko, Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites, J. Theor. Biol. 374 (2015) 35–47, doi:10.1016/j.jtbi.2015.03.006.
[33] D.L. Swofford, PAUP*. Phylogenetic Analysis Using Parsimony (* and other methods), Sinauer Associates, Sunderland, 2003 Version 4.

[34] F. Marroni, S. Pinosio, E. Di Centa, I. Jurman, W. Boerjan, N. Felice, F. Cattonaro, M. Morgante, Large-scale detection of rare variants via pooled multiplexed next-generation sequencing: Towards next-generation, Ecotilling. Plant J. 67 (2011) 736–745, doi:10.1111/j.1365-313X.2011.04627.x.
[35] H. Siu, Y. Zhu, L. Jin, M. Xiong, Implication of next-generation sequencing on association studies, BMC Genom. 12 (2011), doi:10.1186/1471-2164-12-322.
[36] C.A. Anderson, F.H. Pettersson, G.M. Clarke, L.R. Cardon, A.P. Morris, K.T. Zondervan, Data quality control in genetic case-control association studies, Nat Protoc 5 (9) (2010) 1564–1573, doi:10.1038/nprot.2010.116.
[37] M.G. Harvey, B.T. Smith, T.C. Glenn, B.C. Faircloth, R.T. Brumfield, Sequence capture versus restriction site associated DNA sequencing for shallow systematics, Syst. Biol. 65 (2016) 910–924, doi:10.1093/sysbio/syw036.
[38] B.M. Anderson, K.R. Thiele, S.L. Krauss, M.D. Barrett, Genotyping-by-sequencing in a species complex of Australian hummock grasses (*Triodia*): Methodological insights and phylogenetic resolution, PLoS ONE (2017), doi:10.1371/journal.pone.0171053.
[39] O. Jeffroy, H. Brinkmann, F. Delsuc, H. Philippe, Phylogenomics: the beginning of incongruence? Trends Genet. 22 (2006) 225–231, doi:10.1016/j.tig.2006.02.003.
[40] M.J. Phillips, F. Delsuc, D. Penny, Genome-scale phylogeny and the detection of systematic biases, Mol. Biol. Evol. 21 (2004) 1455–1458, doi:10.1093/molbev/msh137.