Molecular Phylogenetics and Evolution 129 (2018) 304–314

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

A simple strategy for recovering ultraconserved elements, exons, and introns from low coverage shotgun sequencing of museum specimens: Placement of the partridge genus *Tropicoperdix* within the galliformes

De Chen[a,b,*], Edward L. Braun[b], Michael Forthman[c], Rebecca T. Kimball[b], Zhengwang Zhang[a]

[a] MOE Key Laboratory for Biodiversity Science and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing 100875, China
[b] Department of Biology, University of Florida, Gainesville, FL 32611, USA
[c] Department of Entomology & Nematology, University of Florida, Gainesville, FL 32611, USA

ARTICLE INFO

Keywords:
Sequence capture
Reduced complexity assembly
Read mapping
Mitochondrial genome
Phasianidae

ABSTRACT

Next-generation DNA sequencing (NGS) offers a promising way to obtain massive numbers of orthologous loci to understand phylogenetic relationships among organisms. Of particular interest are old museum specimens and other samples with degraded DNA, where traditional sequencing methods have proven to be challenging. Low coverage shotgun sequencing and sequence capture are two widely used NGS approaches for degraded DNA. Sequence capture can yield sequence data for large numbers of orthologous loci, but it can only be used to sequence genomic regions near conserved sequences that can be used as probes. Low coverage shotgun sequencing has the potential to yield different data types throughout the genome. However, many studies using this method have often generated mitochondrial sequences, and few nuclear sequences, suggesting orthologous nuclear sequences are likely harder to recover. To determine the phylogenetic position of the galliform genus *Tropicoperdix*, whose phylogenetic position is currently uncertain, we explored two strategies to maximize data extraction from low coverage shotgun sequencing from approximately 100-year-old museum specimens from two species of *Tropicoperdix*. One approach, a simple read mapping strategy, outperformed the other (a reduced complexity assembly approach), and allowed us to obtain a large number of ultraconserved element (UCE) loci, relatively conserved exons, more variable introns, as well as mitochondrial genomes. Additionally, we demonstrated some simple approaches to explore possible artifacts that may result from the use of degraded DNA. Our data placed *Tropicoperdix* within a clade that includes many taxa characterized with ornamental eyespots (peafowl, argus pheasants, and peacock pheasants), and established relationships among species within the genus. Therefore, our study demonstrated that low coverage shotgun sequencing can easily be leveraged to yield substantial amounts and varying types of data, which opens the door for many research questions that might require information from different data types from museum specimens.

1. Introduction

Next-generation DNA sequencing (NGS) offers promising approaches to discover, sequence, and genotype thousands of genetic markers that enable the study of important questions in ecology, evolution, and conservation (Davey et al., 2011). Of particular interest are NGS approaches that are able to extract a massive amount of orthologous loci from millions of museum specimens available in collections worldwide (Rowe et al., 2011), which have the potential to provide a vast repository of important biological data (Graham et al., 2004; Rocha et al., 2014).

Conceptually, *de novo* whole genome sequencing represents the simplest method to leverage NGS technologies to extract genomic data that can be used for many different purposes, such as addressing phylogenetic questions (Jarvis et al., 2014). However, *de novo* whole genome sequencing and assembly is relatively costly, time-consuming, and computationally difficult. Typically, this involves the construction of multiple sequencing libraries with different insert sizes followed by sequencing to $> 20\times$ coverage. Yet *de novo* whole genome sequencing remains challenging for highly degraded and fragmented antique DNA from museum specimens (e.g., Hung et al., 2014; Murray et al., 2017; Staats et al., 2013) because there is often very limited tissue available

* Corresponding author at: MOE Key Laboratory for Biodiversity Science and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing 100875, China.
E-mail address: chende@bnu.edu.cn (D. Chen).

https://doi.org/10.1016/j.ympev.2018.09.005
Received 26 February 2018; Received in revised form 23 July 2018; Accepted 6 September 2018
Available online 07 September 2018
1055-7903/

from those samples. Moreover, the highly fragmented nature of antique DNA makes it impossible to construct libraries with long DNA inserts, which are used to facilitate assembly. Therefore, although NGS costs are changing rapidly, these considerations suggest that *de novo* whole genome sequencing will remain prohibitive for museum specimens.

The relatively high costs of *de novo* whole genome sequencing has made systematists more interested in reduced-representation NGS approaches, like sequence capture (Faircloth et al., 2012; Lemmon et al., 2012). Sequence capture can be used to obtain large numbers of sequences from orthologous loci from many different samples (e.g., Branstetter et al., 2017; McCormack et al., 2013) and it can be especially useful for museum specimens (e.g., Bi et al., 2013; McCormack et al., 2016; Wood et al., 2018). Sequence capture involves additional steps to enrich the NGS libraries for each specimen using probes that will hybridize with preselected genomic regions of interest. For probes to be useful in a variety of taxa they must be relatively conserved, so commonly used probe sets target some combination of coding exons and non-coding ultraconserved element (UCE) regions. The basis for the cost advantage of sequence capture is the greatly reduced sequencing data, and only a single library is typically used for sequence capture, which can also reduce costs.

A third approach, low coverage shotgun sequencing (also called "genome skimming", Straub et al., 2012) is a fairly straightforward method, usually involving only one library preparation step. Unlike *de novo* whole genome sequencing, this approach is suitable when there are limited amounts of DNA, such as degraded DNA from museum specimens. However, with very few exceptions (e.g., Bruxaux et al., 2018), most studies have only obtained mitochondrial genomes and/or a small number of nuclear loci from low coverage shotgun sequencing of museum specimens (e.g., Besnard et al., 2016; Hung et al., 2013; Kanda et al., 2016). This is likely due to the high copy number of the mitochondrial genome in eukaryotic cells in many taxa, making coverage of mitochondrial sequences much greater than nuclear sequences which can allow assembly of mitochondria from low coverage shotgun sequencing, even from degraded DNA (Besnard et al., 2016). Low coverage shotgun sequencing represents a compromise between sequence capture and *de novo* whole genome sequencing; it can sample the genome more broadly than sequence capture and it is also much less costly than *de novo* whole genome sequencing and assembly.

As genomes of non-model organisms have become increasingly available, several studies have found that phylogenetic analyses of different data types yield different topologies for the same taxa (e.g., Jarvis et al., 2014). Differences between introns and coding exons have attracted substantial attention (Chen et al., 2017; Reddy et al., 2017) but it is clear that there are also differences among other data types (Edwards et al., 2017; Wang et al., 2017). The design of sequence capture probes may not feasible for more variable regions of the genome (e.g., introns). This makes comparisons among data types very difficult for sequence capture studies but straightforward for sequencing approaches that sample across the genome, whether it involves deep sequencing of multiple libraries or shallow sequencing of a single library.

There are several factors to consider in the decision between low coverage shotgun sequencing and sequence capture, particularly for studies involving museum specimens. The primary considerations are costs, labor, and the types of data that can be generated. Both methods require library construction but sequence capture also requires probes and other reagents as well as additional labor for sample enrichment. These additional costs are offset by the limited amount of data targeted by sequence capture (e.g., common UCE probe sets sample < 0.5% of an avian genome). At present, the much smaller target DNA for sequence capture makes the costs of sequence capture lower than those for low coverage shotgun sequencing; however sequence capture requires deeper sequencing so the cost differential is smaller than one might naively expect based on the amount of DNA targeted by the sequence capture probes. As sequencing costs continue to decline (at a greater rate than probes and related costs) the cost savings associated with sequence capture will be reduced or possibly even eliminated. The other consideration is that sequence capture is limited to data types that can be targeted using conserved probes; in contrast, low coverage shotgun sequencing allows one to sample more broadly across the genome and obtain many different data types. Therefore, it is worthwhile to test whether low coverage shotgun sequencing can yield a large amount of sequence for different types of data from museum specimens.

The goal of this study was to identify a practical way to maximize data extraction from low coverage shotgun sequencing of museum specimens beyond the recovery of just the mitochondrial genomes, and to determine whether these data were able to resolve the position of a hard-to-place taxon, the galliform genus *Tropicoperdix*. *Tropicoperdix* was erected by Blyth (1859) but was later subsumed within the genus *Arborophila*, the hill partridges (Davison, 1982); most major checklists (e.g., del Hoyo and Collar, 2014; Dickinson and Remsen, 2013) adopted the change at that time. However, Chen et al. (2015) used data from two mitochondrial gene regions and five nuclear introns to demonstrate that *Tropicoperdix* is distinct from *Arborophila* (the two genera are placed within different subfamilies of the Phasianidae) but their data was unable to resolve the exact position of *Tropicoperdix* within Phasianidae. To address this, we first used data from low coverage shotgun sequencing of two old *Tropicoperdix* museum specimens to explore two approaches for extracting UCE data (we focused on UCE data since we had data from many other species for comparison): reduced complexity assembly and direct read mapping. We also compared the effectiveness of read mapping when using reference sequences from a closely related taxon to the use of a more distantly related taxon. After identifying the most effective approach to extract UCE data, we tested whether we could easily obtain a large number of sequences for other more variable data types, including exons and introns. With these data, we explored possible artifacts that may come from using low coverage shotgun sequencing from highly degraded and fragmented antique DNA. Finally, we tested whether our extracted data would permit us to place the exact position of *Tropicoperdix* within Phasianidae with confidence.

2. Materials and methods

2.1. NGS raw read filtering

We used the low coverage shotgun sequencing data of *Tropicoperdix merlini* and *T. charltonii* generated from Chen et al. (2015). The toepads were provided by the Zoological Reference Collection (ZRC) of the National University of Singapore, *T. merlini* (ZRC 3.1478) was collected in Vietnam, 1924 and *T. charltonii* (ZRC 3.1512) was collected in Borneo, 1914. 134 ng of total DNA (0.67 ng/μl) from *T. merlini* and 2.8 ng (0.014 ng/μl) from *T. charltonii* were extracted using Qiagen DNeasy Blood & Tissue Kit (see details in Chen et al., 2015). The average fragment length was smaller than 100 base pairs (bp) in *T. merlini*, and it could not be determined in *T. charltonii* (Fig. S1). Single-end 100 bp sequence reads were generated using an Illumina HiSeq 2000 Genome Analyzer (85,299,440 reads from *T. merlini* and 72,360,071 reads from *T. charltonii*). We then removed low-quality reads and trimmed adapter sequences using Trimmomatic 0.36 (Bolger et al., 2014) with default settings to a minimum length of 20 bp. Finally, we removed duplicate reads (6,256,466 for *T. merlini* and 14,216,639 for *T. charltonii*) to yield the final sets of clean reads (78,870,187 and 57,646,083, respectively).

We also used sequence capture of UCEs for a congener, *T. chloropus*, where high quality, intact DNA was available (sample from University of Kansas Biodiversity Institute, catalog # 119693, tissue # 25432). RAPiD Genomics (Gainesville, FL, USA) performed the sequence capture of 5060 UCE loci using 5472 probes (Faircloth et al., 2012), and subsequent sequencing. The raw reads were filtered in the same way as described above, then the UCE loci were assembled and extracted

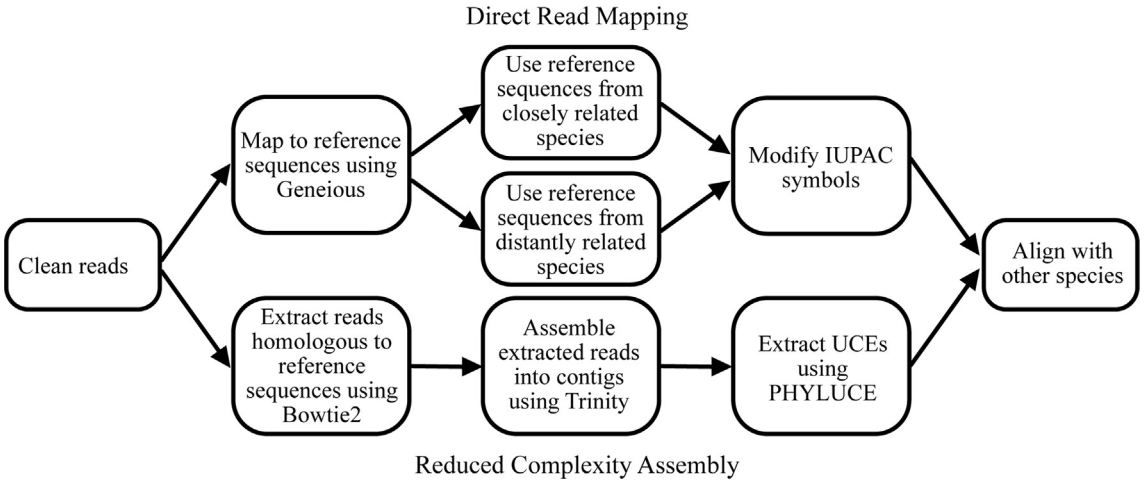

Fig. 1. Flow chart showing reduced complexity assembly (RCA) and read mapping strategies for the recovery of ultraconserved elements (UCEs).

following the procedures detailed in Hosner et al. (2017).

2.2. Comparison between reduced complexity assembly and direct read mapping of UCEs

We compared the performance of different methods for data extraction using UCE since we had data from more species and loci for this marker type. The standard approach often used in the analysis of UCEs is assembly of all data followed by extracting the UCE loci by matching contigs to the sequences of UCE probes (this is the approach implemented in PHYLUCE, (Faircloth, 2016), a commonly used software in dealing with UCE analyses). Our initial trial assembled all clean reads using ABySS (Simpson et al., 2009), yielding 89,887 and 42,444 contigs for *T. merlini* and *T. charltonii*, respectively. However, only 7 and 3 contigs from *T. merlini* and *T. charltonii* matched to UCE probe sequences.

Therefore, we used a reduced complexity assembly (hereafter RCA) strategy that only assembles reads that match a reference sequence (Fig. 1, bottom half). We used 4643 UCE loci that were shared between *T. chloropus* and *Gallus gallus* as reference, and these were trimmed to the same length (422 ± 123 bp) in both species. This method has been used to improve the mitochondrial assembly from low coverage shotgun sequencing (Chen et al., 2015). In our case, first, we aligned the clean reads to the reference sequences using Bowtie2 2.3.0 (Langmead and Salzberg, 2012). Then, we extracted the mapped reads and assembled them into contigs using Trinity r20150302 (Grabherr et al., 2011). Finally, we used PHYLUCE 1.5 (Faircloth, 2016) to determine which contigs represent UCE loci by matching contigs to 5472 UCE probes (Faircloth et al., 2012).

We also used another approach, direct read mapping (hereafter mapping) and tested two different sets of reference sequences (Fig. 1, top half): (1) UCEs of *T. chloropus* as a closely related reference; and (2) UCEs of *G. gallus*, which represented a more distantly related reference. We included the more distant reference because reference sequences from a close relative may not be available in many studies. Since we only included UCE reference sequences that were found in both *T. chloropus* and *G. gallus*, and that were trimmed to be the same length (as described above), we can test the impact of relatedness between reference and focal taxa without bias due to a larger number of UCE loci or longer loci from the *G. gallus* genome. We mapped the final clean reads from *T. merlini* and *T. charltonii* to these references using the "Map to Reference" tool in Geneious R9 (Biomatters Ltd, Auckland, New Zealand) using a "strict" custom sensitivity: allow gaps (maximum per read 2%, maximum gap size 5 bp), minimum overlap 25 bp, minimum overlap identity 95%, maximum mismatches per read 5%, and running 5 iterations. After mapping, the consensus sequences for each UCE locus was saved using a 75% masking threshold and sites that received insufficient coverage (< 3×) were coded using the IUPAC ambiguity symbol N. Because the UCE loci generated from RCA did not include ambiguous symbols, we converted all of the bases with IUPAC ambiguity symbols other than N (i.e., R, Y, S, W, K, M, B, D, H, and V) to N, and excluded UCE loci with fewer than 10 unambiguous nucleotides (i.e., bases A, C, G, T).

For both the RCA and mapping approaches, we combined our extracted UCEs (with appropriate changes to the sequence names for each UCE locus using a custom Perl script) with the UCEs from *T. chloropus*, *G. gallus* and 23 species (selected to represent all possible sister taxa of *Tropicoperdix* based on Chen et al. (2015), as well as outgroups) from Hosner et al. (2016b). We then aligned and edge-trimmed all loci to build concatenated alignments in PHYLUCE in which only loci with more than 75% of taxa present were included.

We used two methods to estimate UCE phylogenies using data generated by RCA and mapping respectively. First, we estimated maximum likelihood (ML) trees from concatenated alignments using RAxML 8.2.3 (Stamatakis, 2014) with the GTR + G model and '-f a' option (which generates the optimal tree and conducts 100 rapid bootstrap searches). Second, we used SVDquartets (Chifman and Kubatko, 2014) as implemented in PAUP* 4.0a159 (Swofford, 2003) with 100 bootstraps, each sampling all quartets. SVDquartets is a consistent estimator of the species tree and has already been shown to be an effective reconstruction method on several data sets (Chifman and Kubatko, 2015; Long and Kubatko, 2017).

We compared our RCA to the mapping approach using the following criteria: (1) the number of UCE loci that were recovered; (2) the number of unambiguous nucleotides that were recovered; and (3) the impact of different approaches on phylogenetic estimation. Because the minimal length of unambiguous nucleotides in UCE loci from RCA (34 bp) was longer than that from mapping (10 bp), we also excluded UCE loci from mapping that contained less than 34 bp of unambiguous nucleotides and re-compared the number of recovered UCE loci and unambiguous nucleotides between them.

2.3. Recovery of sequences for other data types using the mapping approach

We extracted exonic data using our mapping approach with reference exons from Prum et al. (2015), who conducted a sequence capture study that largely focused on exonic regions from more than 200 nuclear loci to build a phylogeny of 198 bird species (including nine galliforms). From this, we obtained 222 coding exons from *G. gallus* as our reference. We used Geneious R9 with the "strict" custom sensitivity and the same consensus criteria that we used for UCEs (see above) and excluded exon sequences with fewer than 10 unambiguous

nucleotides. We combined the sequences we recovered from *Tropicoperdix* species with the sequences from the nine galliform species analyzed by Prum et al. (2015) using PHYLUCE (after running the Perl script described above to convert each exon sequence into the necessary format). Then, we built a concatenated alignment after aligning and edge-trimming all loci using PHYLUCE; our final alignment comprised 214 exonic regions for which more than 75% of taxa were present for each exon.

To test our ability to extract intronic data we generated a dataset of 92 introns from 40 loci using a combination of published data (Kimball and Braun, 2014), data mining from sequenced genomes, and PCR amplification and sequencing (primer sequences provided in Table S1). We then used the *G. gallus* sequences for the 92 introns as our reference sequences. For the intronic mapping, we used a "tolerant" custom sensitivity in Geneious R9: allow gaps (maximum per read 5%, maximum gap size 5 bp), minimum overlap 25 bp, minimum overlap identity 85%, maximum mismatches per read 15%, and running 5 iterations. We used the same consensus criteria described above to obtain consensus sequences, excluding intron sequences containing fewer than 10 unambiguous nucleotides. We combined these extracted sequences with intron sequences from 10 other galliform species and used the Perl script described above to convert each sequence into the necessary format. All loci were then aligned and edge-trimmed using PHYLUCE. The final alignment comprised 79 introns for which more than 75% of taxa were present for each intron.

We obtained mitochondrial genomes from GenBank for 17 species that provided a similar taxonomic coverage as our UCE and intron datasets (the exon dataset had very few overlapping taxa). We used the mitochondrial genome of *G. gallus* (GenBank: X52392) as our reference, using the "tolerant" custom sensitivity and the same consensus criteria described above in Geneious R9. We then generated a final data matrix containing the complete mitochondrial genomes except the control region due to the difficulty of aligning the mitochondrial control region across galliforms (Meiklejohn et al., 2014).

2.4. Phylogenetic comparison of different data types recovered from mapping

We wanted to compare phylogenetic trees for all data types (i.e., UCEs, exons, introns, and mitochondrial genomes) from museum specimens based on our mapping approach. However, the large number of UCEs for *T. chloropus* which were obtained using sequence capture of intact DNA could facilitate the placement of other *Tropicoperdix* species with the UCE data (since data from *T. chloropus* was not available for the exon, intron, or mitochondrial datasets). Thus, in addition to the UCE phylogenies described above using all species, we excluded *T. chloropus* from the UCE alignments, and re-estimated the UCE phylogeny using RAxML. Additionally, we estimated a UCE phylogeny with only the lowest quality data, that of *T. charltonii*, by excluding both *T. chloropus* and *T. merlini*. This allowed us to assess whether the lowest quality UCE data alone could place a genus of uncertain affinities, when higher quality data from close relatives was unavailable. For the exon, intron, and mitochondrial datasets, where only *T. merlini* and *T. charltonii* were available, we used RAxML to estimate ML trees and bootstrap support using the same settings as described for the UCE datasets.

2.5. Assessing factors that affect branch lengths when using antique DNA

Sequences from museum specimens tend to exhibit long branches in concatenated ML trees because of higher contamination and sequencing errors (McCormack et al., 2016). Additionally, missing data, which is common in sequences from museum specimens, can also lead to long branches (Darriba et al., 2016). We examined branch lengths using the patristic distance (the sum of the intervening branches) from the common ancestor of *Gallus* and *Tropicoperdix* to the tips of *G. gallus* and each *Tropicoperdix* species for all of our datasets.

We ran additional tests using the UCE data where we had recovered more sequences. To examine the impact of sequence errors on branch lengths for the UCE data, we increased the base-calling coverage requirement from mapping to 5×, which should reduce the likelihood of contamination and sequencing errors. We generated an alignment that included both the 3× and 5× coverage data to improve the alignment quality for the shorter 5× data, and then removed the 3× data to obtain a dataset with the 5× data only. Using this alignment, we estimated the ML tree using RAxML as described above to see if branch lengths for the 5× coverage sequences were shorter than those from the 3× coverage sequences.

To test if branch lengths in the UCE data were affected by missing data, we built an alignment containing no ambiguity symbols or gaps from the 3× coverage sequences ("gap-free alignment"). Since this dataset had limited power, we used it to estimate the branch lengths only by using a fixed topology (that of ML tree based on 3× coverage sequences) in RAxML ("-f e" option) (Stamatakis, 2014).

3. Results

3.1. RCA versus mapping for the recovery and phylogeny of UCEs

We extracted 124,315 (0.16%) reads of *T. merlini*, and 57,196 (0.10%) reads of *T. charltonii* after aligning to the reference for the RCA method. For mapping, after 5 iterations, we mapped 262,624 (0.33%) and 26,199 (0.05%) reads to *T. merlini* and *T. charltonii* respectively.

We obtained about 3× the number of UCE loci with mapping than RCA for *T. merlini*, which yielded up to 93% of the targeted UCE loci (Table 1). Mapping generally outperformed RCA for *T. charltonii* as well, which overall yielded drastically fewer loci than did *T. merlini* (at most 8% of loci recovered, Table 1). Consistent with these results, the number of resolved unambiguous nucleotides from RCA were fewer than that from mapping, especially for *T. merlini* (Table 1). Moreover, the resulting UCE loci from mapping with the same minimal length of unambiguous nucleotides as RCA (34 bp) still resulted in more than 2× the yield using mapping than from RCA in *T. merlini*; and slightly

Table 1
Yields of UCE loci and unambiguous nucleotides for *Tropicoperdix merlini* and *T. charltonii* using reduced complexity assembly (RCA) and read mapping (mapping) (see details in Fig. 1).

	Taxa	Loci	% of reference loci	Nucleotides	% of reference nucleotides
Close reference	*T. chloropus*	4643	–	1,963,658	–
Distant reference	*G. gallus*	4643	–	1,962,780	–
RCA (minimal length of nucleotides is 34 bp)	*T. merlini*	1572	34%	202,051	10%
	T. charltonii	211	4.5%	12,715	0.6%
Mapping to close reference (minimal length of nucleotides set to 34 bp)	*T. merlini*	4108	88%	712,956	36%
	T. charltonii	230	5%	12,210	0.6%
Mapping to close reference (minimal length of nucleotides set to 10 bp)	*T. merlini*	4321	93%	717,868	37%
	T. charltonii	392	8%	16,020	1%
Mapping to distant reference (minimal length of nucleotides set to 10 bp)	*T. merlini*	4216	91%	632,306	32%
	T. charltonii	345	7%	14,009	1%

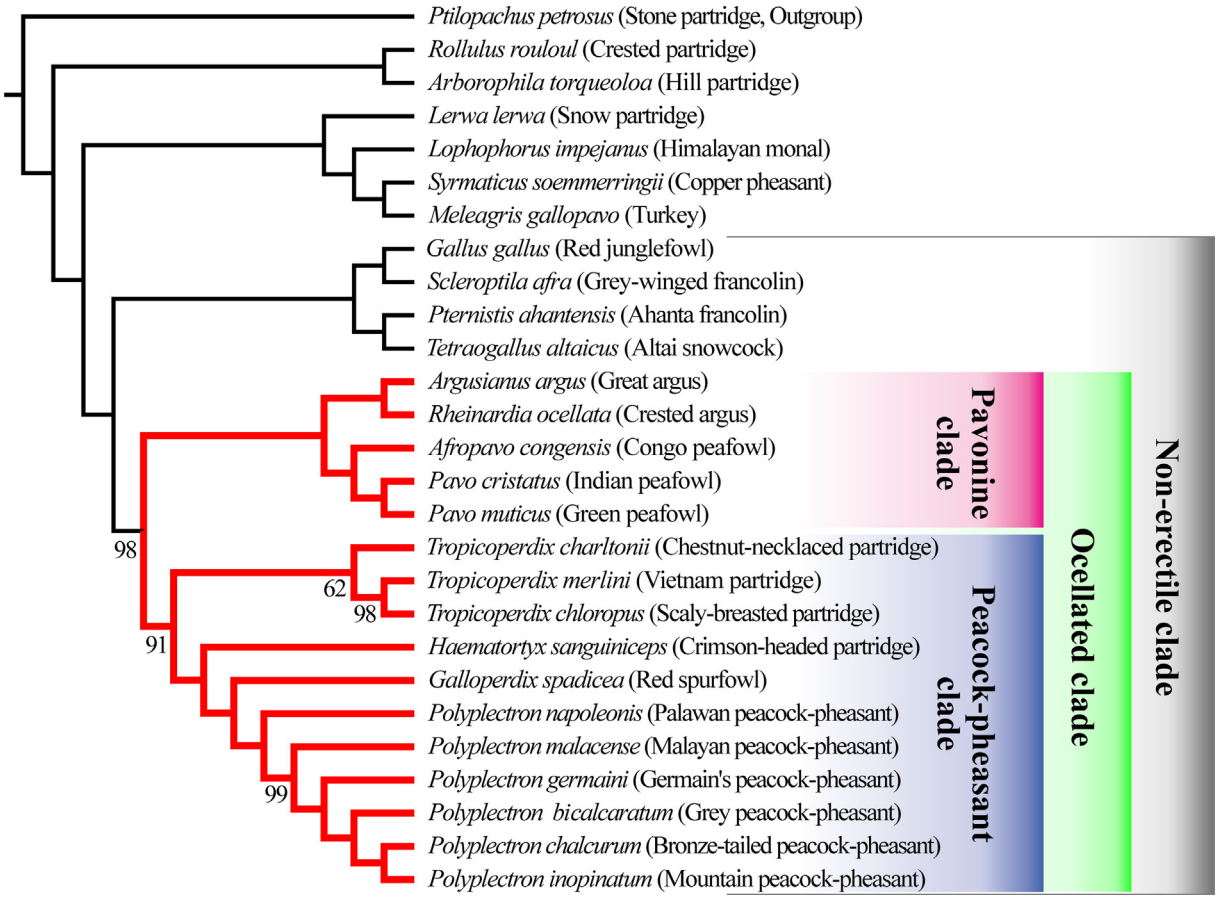

Fig. 2. Species tree estimated using SVDquartets and UCE data. *Tropicoperdix merlini* and *T. charltonii* UCEs were extracted by mapping to the closely-related reference (*T. chloropus*). Clade names used in the text are indicated to the right of the figure and the ocellated clade is emphasized using thick lines (red in the online version). Bootstrap support is indicated below each node when it is less than 100%; all other nodes have 100% bootstrap support. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

outperformed RCA for *T. charltonii* in this instance (Table 1).

The UCE mapping data were sufficient to confidently place *Tropicoperdix* within a clade that included peacock pheasants, and two other partridge genera. Herein we refer to this clade the "peacock pheasant clade" after its best known member (Fig. 2); this clade is nested within a larger clade that also includes peafowl and argus pheasants that was named the "ocellated clade" by Hosner et al. (2016b). Use of SVDquartets yielded a species tree with strong (> 90%) bootstrap support for virtually of the nodes (Fig. 2). The concatenated ML tree showed the same topology, but with 100% bootstrap support for all nodes (Fig. 3). This UCE dataset also allowed us to evaluate the relationships among all three species within *Tropicoperdix*, which were identical to a previous estimate of phylogeny from mitochondrial genes and an expectation based on the morphological similarities among *Tropicoperdix* species (Chen et al., 2015).

The concatenated ML tree based on the RCA data showed the same topology and 100% bootstrap support for all nodes (Fig. 3). However, the branch lengths of *T. merlini* and *T. charltonii* were much longer with the RCA data than the mapping data (Fig. 3), a finding that suggests RCA leads to more errors and/or more missing data than mapping. Supporting this, we were able to identify examples of likely assembly errors in the RCA data (Fig. S2). Use of SVDquartets with the RCA dataset (Fig. S3) resulted in much lower support than with the mapping data (Fig. 2), especially within the genus *Tropicoperdix*, although the same topology was recovered, the bootstrap support was lower than 50%. This suggests that likely errors and/or missing data in the RCA data may have a greater impact on SVDquartets. Overall, it seems clear that our mapping strategy yielded better data than the RCA approach.

3.2. Mapping to closely versus distantly related reference sequences

We recovered a larger number of UCE loci and more unambiguous nucleotides when mapping to the more closely related reference (Table 1). However, the differences in the number of loci and unambiguous nucleotides were modest. Indeed, 97% of the *T. merlini* UCE loci recovered using the close reference were also recovered using the distant reference. For *T. charltonii*, this number was reduced to 88% (Table 1). In both cases, the number of unambiguous nucleotides recovered using the distant reference was only reduced to 87–88% of the recovery from the close reference (Table 1). An ML tree that combined sequences recovered from the distant relative to those recovered from the close relative in the same alignment showed the same topology within *Tropicoperdix*, though the bootstrap support within this clade decreased (Fig. S4). The branch length of *T. merlini* estimated from the distant relative was virtually identical to that from the close relative, but the branch length of *T. charltonii* estimated from the distant relative was longer than that from the close relative (Fig. S4).

3.3. Recovery of exons, introns, and mitochondrial genomes by mapping

We obtained 100% of the reference exons from *T. merlini* and 33% of exons for *T. charltonii* by mapping to *G. gallus*. The number of unambiguous nucleotides for *T. merlini* was about 61% that of the reference exons, whereas it was only about 1% of the reference sequences for *T. charltonii* (Table 2). For introns, we recovered 95% of loci for *T. merlini* and 13% for *T. charltonii* by mapping to the *G. gallus* reference; the resolved unambiguous nucleotides were about 39% (*T. merlini*) and

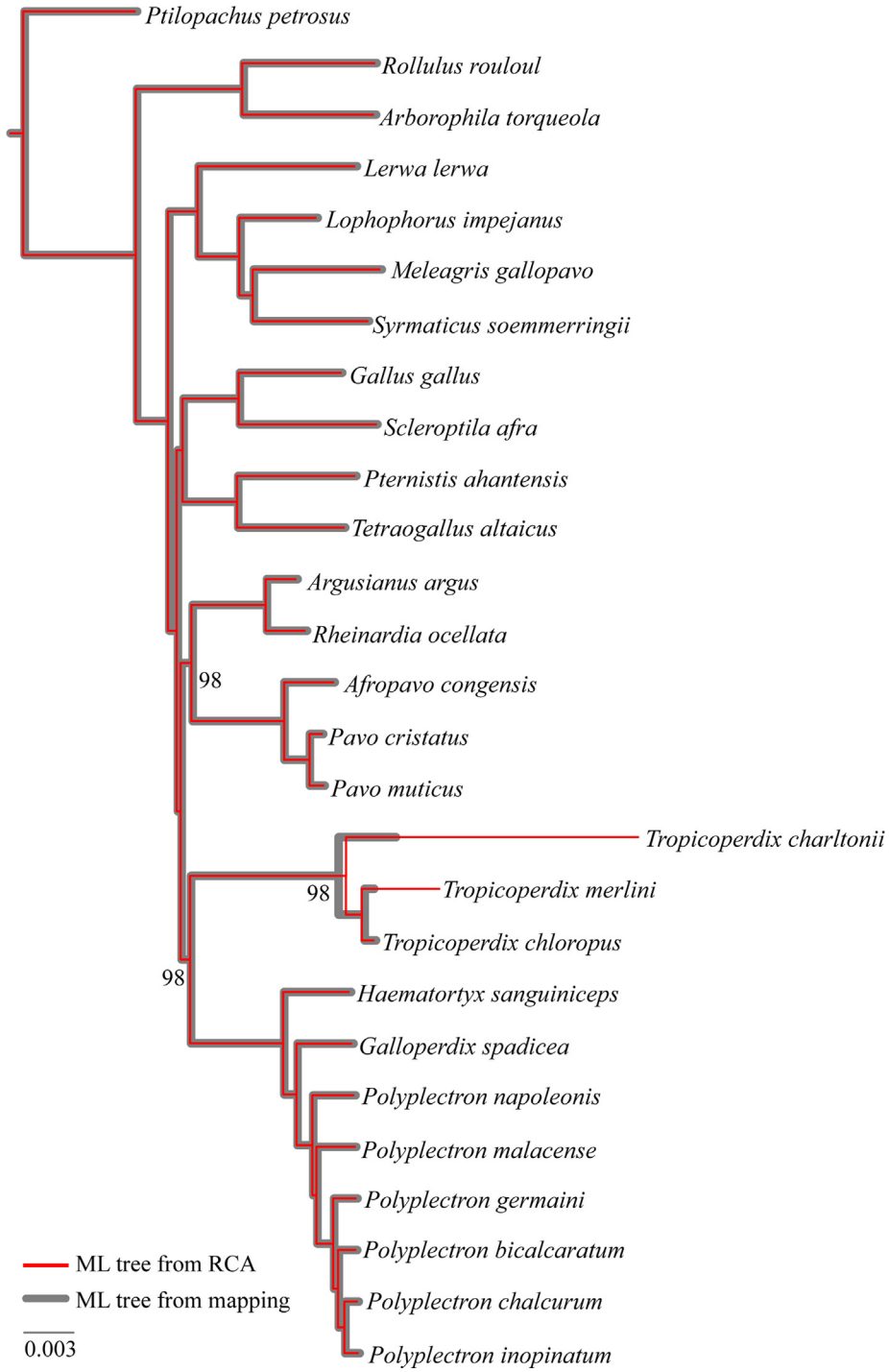

Fig. 3. Maximum likelihood (ML) trees for UCEs, with *T. merlini* and *T. charltonii* extracted using RCA (thin lines, red in the online version) and mapping to the closely-related reference (*T. chloropus*) (thick lines). All nodes had 100% bootstrap support for both analyses. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2% (*T. charltonii*) that of the reference introns (Table 2). Finally, we recovered the complete mitochondrial genomes for *T. merlini* and *T. charltonii* by mapping to the *G. gallus* mitochondrial genome (Table 2).

3.4. Tree topologies for each data type

To fairly compare phylogenetic trees for different data types, we excluded *T. chloropus* from the UCE alignments. The resulting UCE tree placed the other *Tropicoperdix* species within the peacock pheasant clade with 100% bootstrap support (Fig. S5). Further excluding *T. merlini* and leaving only the lowest quality *T. charltonii* data still placed *Tropicoperdix* within the peacock pheasant clade, albeit with much lower support (64%; Fig. S6). The exon tree had 100% bootstrap support at every node (Fig. 4a). In contrast, using the smaller intron and mitochondrial datasets led to several nodes (Fig. 4b and c) that were poorly supported. There were also some poorly supported topological differences among the trees (though the limited overlapping taxa prevented extensive comparison, especially for the exon tree). The mitochondrial tree failed to place *Tropicoperdix* in the peacock pheasant clade, which was recovered by UCEs and introns. With introns, support

Table 2
Yields of exons, introns and mitochondrial genomes by mapping to *G. gallus*.

	Taxa	Loci	% of reference loci	Nucleotides	% of reference nucleotides
Exon	*G. gallus*	222	–	294,210	–
	T. merlini	222	100%	179,135	61%
	T. charltonii	74	33%	3452	1%
Intron	*G. gallus*	92	–	43,647	–
	T. merlini	87	95%	17,198	39%
	T. charltonii	12	13%	760	2%
Mitochondrial genome	*G. gallus*	1	–	16,775	–
	T. merlini	1	–	16,699	–
	T. charltonii	1	–	16,699	–

for the peacock pheasant clade was extremely low and the recovered relationship outside the peacock pheasant clade differed from the UCE tree (Figs. 3 and 4b).

3.5. Assessing factors that affect branch lengths of *T. charltonii*

T. charltonii, which had lowest quality DNA (Fig. S1) and yielded the least data (Tables 1 and 2), exhibited longer branches than *T. merlini* in all analyses and datasets except the mitochondrial dataset (Figs. 3 and 4, Table 3).

When using the more stringent mapping criterion (minimum 5× coverage instead of 3×) that should reduce error for the UCEs, we obtained 3187 loci for *T. merlini* and 43 loci for *T. charltonii*, which included about 40% and 10%, respectively, of the unambiguous nucleotides that were recovered with the 3× coverage (Table S2). Phylogenetic analysis of the alignment using the 5× coverage data recovered an identical tree topology to that generated using the 3× coverage UCEs (Fig. S7). We expected that, if sequencing errors explain the effect on branch lengths, the 5× coverage tree should have shorter branches than the 3× coverage tree due to the use of more accurate UCE sequences. However, *T. merlini* showed virtually identical branch lengths and the estimates of branch lengths for *T. charltonii* actually increased when the 5× coverage dataset was used (Table 3).

The branch lengths of *T. charltonii* and *T. merlini* from the gap-free alignment were shorter than those estimated from the 3× alignment (Table 3). However, *T. charltonii* still exhibited a longer branch than *T. merlini* and *T. chloropus* (Table 3, Fig. S8), although the amount of data in this alignment was very limited (3551 bp).

4. Discussion

We successfully obtained four different types of sequence data using low coverage shotgun sequencing from approximately 100-year-old museum specimens using a simple mapping strategy, two of which corresponded to conserved regions (UCEs and exons), one that corresponded to more variable nuclear regions (introns), and the fourth was the complete mitochondrial genomes. Our results, therefore, indicate that it is relatively straightforward to extract a larger amount of sequence data from low coverage shotgun sequencing than has often been done in many previous studies (e.g., Besnard et al., 2016; Hung et al., 2013; Kanda et al., 2016). It is also possible to obtain multiple types of sequence data (including UCEs, coding exons, and introns), permitting detailed comparisons of analyses using those types of data.

4.1. Phylogenetic position of Tropicoperdix

Tropicoperdix was for many years placed in the genus *Arborophila*. Since no sequence data was available from this genus until recently, it was misplaced in meta analyses that use taxonomies (e.g., Jetz et al., 2012) and absent from "big trees" based on empirical sequence data generated prior to 2015 (e.g., Burleigh et al., 2015). Most recent estimates of galliform phylogeny (Hosner et al., 2016b; Kimball and Braun, 2014; Wang et al., 2013) break Phasianidae into three major clades: (1) Arborophilinae (hill partridges and allies); (2) the "non-erectile clade" (comprising junglefowl, Old World quail, many partridges and francolins, and the peafowl and allies; see Fig. 2); and (3) the "erectile clade" (sister to the non-erectile clade and comprising turkey, grouse, true pheasants, true partridges, tragopans, and their allies; see Kimball and Braun (2008) for details. Chen et al. (2015), who provided the first molecular data for *Tropicoperdix*, convincingly demonstrated that *Tropicoperdix* is a member of the non-erectile clade. Instead of weakly placing *Tropicoperdix* in the Pavonine clade (which includes *Argusianus*, *Afropavo*, and *Pavo*; see Fig. 2) (Chen et al., 2015), our UCE dataset strongly indicated *Tropicoperdix* should be placed in the peacock-pheasant clade, the sister group of the Pavonine clade that includes *Haematortyx*, *Galloperdix*, and *Polyplectron* (Fig. 2).

4.2. Phylogenies estimated using different data types

Studies using relatively small numbers of loci often result in some poorly supported relationships or conflicting results among studies, as has been found previously in Phasianidae (e.g. Kimball and Braun, 2014; Wang et al., 2013). Here, the datasets with smaller numbers of loci and sites showed low support (Fig. 4b, c), and for the mitochondrial data, placed *Tropicoperdix* in a different position from the other datasets (Fig. 4c). The UCE tree, in contrast, recovered high bootstrap support for all nodes (Figs. 2 and 3), even when only *T. merlini* and *T. charltonii* were included (Fig. S5), though more modest support is found with the lowest quality *T. charltonii* data (Fig. S6). This corroborates the expectation that UCEs have the potential to resolve difficult evolutionary radiations that have remained unresolved with more traditional datasets (but see Meiklejohn et al. (2016) for a study illustrating the potential limits of UCE data). It is also important to note that mitochondrial DNA, as well as introns and exons, remain the dominant types of data available from prior studies, and attempts to synthesize these data into large trees (e.g., Burleigh et al., 2015; Pyron et al., 2013) are likely to remain valuable. Extracting introns, exons, and mitochondrial data allows inclusion of data from museum specimens to be integrated into these existing datasets. Furthermore, it seems reasonable to recommend supplementing the recovery of UCE sequences with these other data types to improve phylogenies (Hosner et al., 2016a; Persons et al., 2016). These results emphasized the importance for other studies seeking to clarify the phylogenetic positions of focal species through low coverage shotgun sequencing to use more than just mitochondrial and/or a few nuclear loci (e.g., Besnard et al., 2016; Kanda et al., 2016).

Phylogenomic studies have shown that different nuclear data types (e.g., coding versus UCEs, introns, or conserved non-exonic elements) can yield different topologies (Edwards et al., 2017; Jarvis et al., 2014; Reddy et al., 2017). Although we did not observe any clear data type effects, this may be because our datasets lacked substantial overlap for comparison. For other studies, however, detecting potential data type differences, and thus understanding additional sources of uncertainty in the underlying evolutionary patterns, requires analyzing multiple data types (Reddy et al., 2017). Sequence capture largely restricts data recovery to those regions that are targeted by conserved probes, and so may primarily yield UCEs (e.g., Faircloth et al., 2012; McCormack et al., 2016) and/or conserved exonic regions (e.g., Prum et al., 2015). However, low coverage shotgun sequencing data allowed us to pull out substantial numbers of UCEs, exons, and introns using a simple mapping strategy that could even be used to extract other data types.

4.3. Effects of DNA quantity and quality on sequence recovery

The DNA quantity (134 ng) and quality (average fragment size lower than 100 bp, Fig. S1) of *T. merlini* was lower than that of most bird and mammal museum specimens used in sequence capture studies

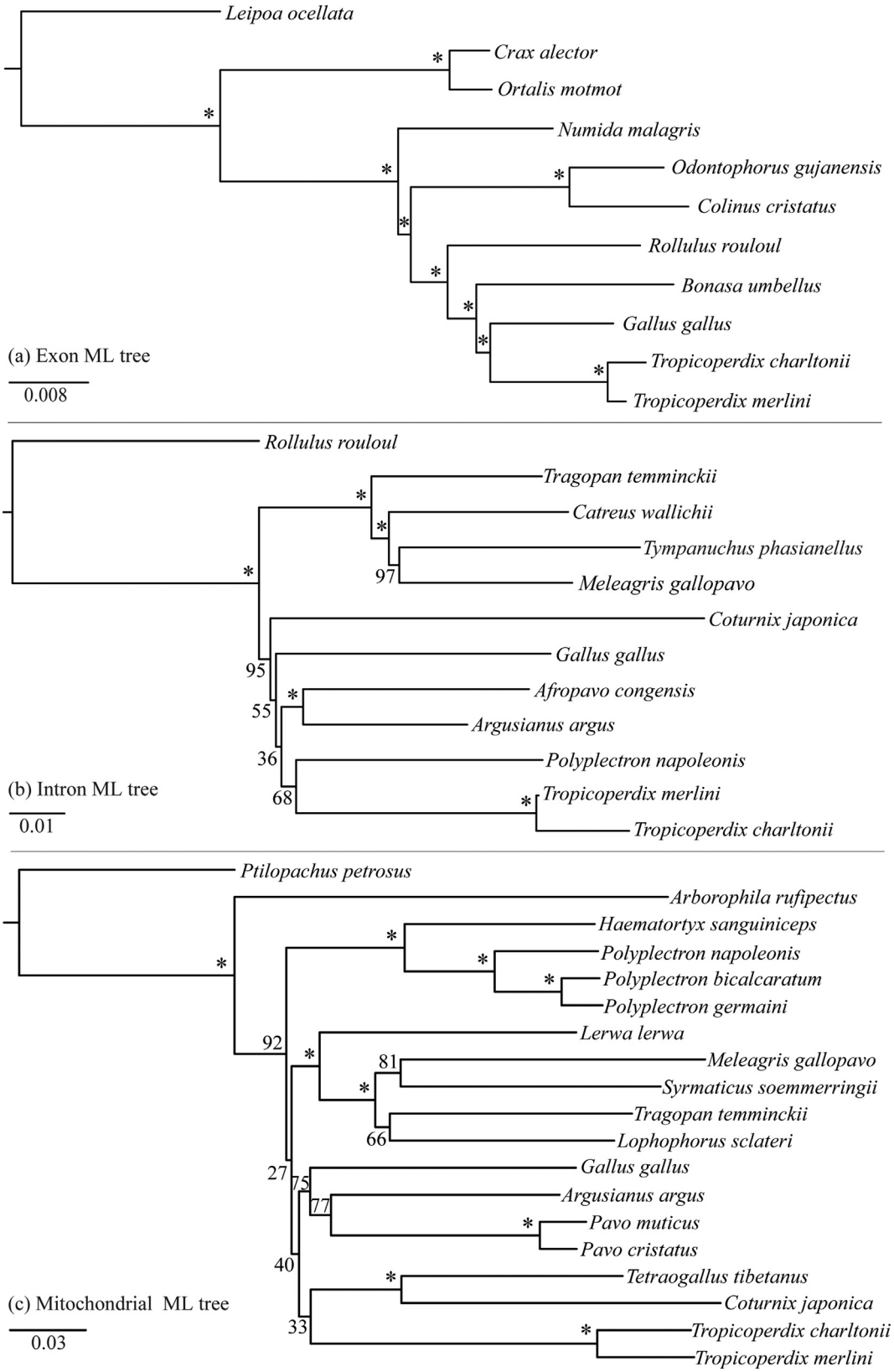

Fig. 4. ML trees from different data types obtained by mapping to *Gallus gallus*: exons (a), introns (b), and the mitochondrial genome (c). Nodes with 100% bootstrap support are indicated with asterisks *.

(e.g., Hawkins et al., 2016; McCormack et al., 2016). Yet, the recovery of UCE data from low coverage shotgun sequencing of *T. merlini* was similar to those studies. Since low coverage genome sequencing requires fewer manipulations than sequence capture, it may be advantageous for limited, low quality DNA that may be prone to contamination. Furthermore, low coverage shotgun sequencing from fresh

Table 3
Branch lengths from different UCE datasets, exon, intron and mitochondrial datasets. UCE mapping datasets were generated from mapping to a close relative. Branch lengths were calculated from the common ancestor of *Gallus* and *Tropicoperdix* to the tips of *G. gallus* and each *Tropicoperdix* species.

Dataset	To *G. gallus*	To *T. chloropus*	To *T. merlini*	To *T. charltonii*
UCE from RCA	0.0101	0.0122	0.0164	0.025
UCE from mapping with 3× coverage criterion	0.0101	0.0121	0.0120	0.0133
UCE from mapping with 5× coverage criterion	0.0102	0.0122	0.0120	0.0198
UCE Gap-free alignment	0.0066	0.0097	0.0085	0.0110
Exon from mapping	0.0125	–	0.0138	0.0158
Intron from mapping	0.0504	–	0.0482	0.0649
Mitochondrial genomes from mapping	0.1096	–	0.1564	0.1551

tissue can also be used to extract many different data types (e.g., Zimmer and Wen, 2015).

However, the extremely low DNA quantity (2.8 ng) and quality (fragment size could not be determined) of *T. charltonii* (Fig. S1) appeared to result in uneven coverage of that genome (see also Bruxaux et al., 2018). While we were able to recover the complete mitochondrial genome for *T. charltonii*, coverage elsewhere was sporadic. In spite of this, we were still able to place this species robustly with all of our datasets. Deeper sequencing might increase the recovery of nuclear regions, but it seems likely that the low amounts of input DNA created biases in library construction and that deeper coverage would not substantially improve our results (Bruxaux et al., 2018). Instead, for such extremely low quantity and quality DNA, sequence capture for targeted nuclear regions may work better (Blaimer et al., 2016; Knapp and Hofreiter, 2010), even if sequence capture is not always ideal (as discussed above).

4.4. Mapping outperforms RCA in the recovery of UCEs

The mapping strategy we used yielded a substantially larger number of UCE loci and unambiguous nucleotides than the RCA strategy (Table 1). The main difference between RCA and mapping in UCE extraction is that with RCA reads are assembled prior to extracting UCEs, while mapping directly mapped reads onto reference sequences without prior assembly. The highly fragmented DNA extracted from museum samples could result in DNA segments to be fragmented into several independent contigs that represent a single UCE locus (or the exons and introns we tested) during assembly. Most of these "subcontigs" would not contain the UCE probe sequence, so only one "subcontig" which contains sufficient UCE probe sequence would be attributed to that UCE locus, other subcontigs would be discarded. In more extreme cases, none of these subcontigs may contain sufficient UCE probe sequence, so that UCE locus would not be recovered. In contrast, mapping may allow all of these subcontigs to be assigned to the correct locus, maximizing the recovery for both loci and nucleotides. Therefore, for low coverage shotgun sequencing from antique DNA, mapping is not only better than *de novo* assembly (Kanda et al., 2016; Sproul and Maddison, 2017), it is also better than RCA for the recovery of usable data.

In addition to recovering fewer loci and nucleotides, the branch lengths of the ML tree for *T. merlini* and *T. charltonii* from RCA were longer than those from mapping (Fig. 3, Table 3). Although missing data effects could contribute to this, we found some nucleotides appeared to have been called incorrectly using RCA but were coded as ambiguous (N's) with mapping (see Fig. S2 for an example). Although post-assembly filtering, such as Pilon (Walker et al., 2014), can remove those incorrect nucleotides, it is time consuming and could increase levels of missing data, which might further bias branch lengths (see details below). Thus, mapping can easily increase the accuracy of the recovered nucleotides, a common problem in analyses of antique DNA (McCormack et al., 2016).

4.5. Choice of reference sequences

Although we recovered more data using a closely related reference sequence, we still successfully recovered many UCEs, introns, and exons using *G. gallus* as a reference, which is a more distant relative of *Tropicoperdix*. There are an increasing number of genome sequences for diverse organisms have been produced, and it seems likely that it will soon be possible to obtain genomes as close to any taxa as *G. gallus* is to *Tropicoperdix*. Indeed, efforts such as the G10K (Haussler et al., 2009) and B10K (Zhang, 2015) projects are already producing potential reference genomes for many vertebrates. The UCE tree topology within *Tropicoperdix* was not affected by the use of a closely related versus more distantly related reference, and the branch lengths of *T. merlini* were virtually identical between data extracted using the different references (Fig. S4). However, this was not true for *T. charltonii*, which had a longer branch when using a more distant relative as a reference (Fig. S4), suggesting that more closely related references should be used, when possible, to extract data from extremely low quality DNA.

4.6. Are the observed branch length differences artifactual?

Our results consistently showed that *T. charltonii* had a much longer branch length than *T. merlini* (and, where sampled, *T. chloropus*) for all of our nuclear data (Table 3). Such differences could be explained by one of several different (but not mutually exclusive) hypotheses. It is possible that *T. charltonii* evolves at a higher rate, such that the longer branches accurately reflect the underlying pattern of divergence in that species. However, given that *T. charltonii* was also the lowest quality genome, it is possible that the long branches in *T. charltonii* could be driven by a larger number of sequencing errors (McCormack et al., 2016) or a larger amount of missing data (Darriba et al., 2016). In fact, it is clear that sequence errors and/or missing data do have an impact given that we observed highly inflated branch lengths when the RCA data were used (Fig. 3) and somewhat longer branches when the data were generated by mapping to a more distant reference (Fig. S4). The question here is to evaluate factors that affect branch lengths of *T. charltonii* in our best method for sequence recovery (i.e., mapping to closely related references).

Our results suggested that simple error is unlikely to be driving the branch length differences, since the branch length of *T. charltonii* was longer (rather than shorter) when using the 5× coverage dataset that should have fewer errors (Fig. S7). Since the 5× alignment had more missing data than the 3× alignment, especially for *T. charltonii* (90% fewer nucleotides, Table S2), and the branch length of *T. charltonii* was relatively longer in the 5× versus the 3× coverage (Table 3), this does suggest that missing data may be driving the increased branch length of *T. charltonii*. The gap-free alignment directly tested the missing data hypothesis. As expected, the branch lengths of *T. charltonii* and *T. merlini* were shorter than those estimated from the 3× alignment (Table 3), suggesting that missing data did impact branch length estimates. However, the branch lengths of *T. charltonii* were still longer than *T. merlini* and *T. chloropus* in the gap-free alignment (Table 3), which could indicate processes other than missing data may have an influence. But it is worth noting that elimination of all sites with missing data (retaining only 3551 bp), may have reduced the power to accurately estimate branch lengths.

When considering other data types, the observed branch length differences for exons was quite similar to those for UCEs, but there was a larger effect for the intronic data (i.e., the relative branch length for *T. charltonii* was even longer, Table 3). In contrast, we did not observe any obvious branch length differences between *T. charltonii* and *T. merlini* in

our analysis of the mitochondrial dataset (Fig. 4 and Table 3). Since the mitochondrial dataset had no missing data, unlike the exon and intron datasets, these results are also consistent with an impact of missing data. Thus, while not absolutely definitive, overall these results suggested that missing data may be a major factor resulting in biased branch length estimates and needs to be considered in studies where some taxa may have a lot of missing data. Nonetheless, the observed branch length differences are modest (particularly for UCEs and exons) when using read mapping with 3× coverage and the topology appeared to be robust, suggesting that we can obtain reliable data for phylogenetic analyses using this strategy.

5. Conclusions

In this study, we explored the potential utility of low coverage shotgun sequencing from museum specimens. Our results demonstrated that low coverage shotgun sequencing data can easily be leveraged to yield substantial amounts of different types of data throughout the genome, though not all approaches to extract data work equally well. In addition, we demonstrated some simple tests to look for branch length effects due to errors and missing data that can be applied in other studies. This allowed us to place the genus *Tropicoperdix* with confidence and it opens the door for many research questions that might require information from different data types from museum specimens where only degraded and/or limited amounts of DNA are available.

Data accessibility

Sequence read files were uploaded to NCBI GenBank accessible through NCBI Project number PRJNA481848. Sequence alignments and tree files were deposited in Mendeley, DOI: http://doi.org/10.17632/7cw62snbt4.1.

Author contributions

All authors designed research; D.C., R.T.K. and Z. Z. performed molecular work; D.C., E.L.B., M.F. and R.T.K. analyzed data; and all authors wrote the manuscript.

Funding

This study was funded by the National Natural Science Foundation of China (31601839), the Chinese Fundamental Research Funds for the Central Universities (2017NT09), the National Key Program of Research and Development, Ministry of Science and Technology of China (2016YFC0503200), and the U.S. National Science Foundation (DEB-1118823 and DEB-1655683 to R.T.K. and E.L.B).

Acknowledgments

We thank Drs. George Tiley and Peter Hosner for their help in data analyses. We also thank Dr. Geoffrey Davison and Lee Kong Chian Natural History Museum for their help with the collection of samples. Two anonymous reviewers helped improve the manuscript.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2018.09.005.

References

Besnard, G., Bertrand, J.A.M., Delahaie, B., Bourgeois, Y.X.C., Lhuillier, E., Thébaud, C., 2016. Valuing museum specimens: high-throughput DNA sequencing on historical collections of New Guinea crowned pigeons (*Goura*). Biol. J. Linn. Soc. 117, 71–82.

Bi, K., Linderoth, T., Vanderpool, D., Good, J.M., Nielsen, R., Moritz, C., 2013. Unlocking the vault: next-generation museum population genomics. Mol. Ecol. 22, 6018–6032.

Blaimer, B.B., Lloyd, M.W., Guillory, W.X., Brady, S.G., 2016. Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. PLoS ONE 11, e0161531.

Blyth, E., 1859. Journal of the Asiatic Society of Bengal. Calcutta, India.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Branstetter, M.G., Longino, J.T., Ward, P.S., Faircloth, B.C., 2017. Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. Methods Ecol. Evol. 8, 768–776.

Bruxaux, J., Gabrielli, M., Ashari, H., Prŷs-Jones, R., Joseph, L., Milá, B., Besnard, G., Thébaud, C., 2018. Recovering the evolutionary history of crowned pigeons (Columbidae: *Goura*): implications for the biogeography and conservation of New Guinean lowland birds. Mol. Phylogenet. Evol. 120, 248–258.

Burleigh, J.G., Kimball, R.T., Braun, E.L., 2015. Building the avian tree of life using a large-scale, sparse supermatrix. Mol. Phylogenet. Evol. 84, 53–63.

Chen, D., Liu, Y., Davison, G.W.H., Dong, L., Chang, J., Gao, S., Li, S.-H., Zhang, Z., 2015. Revival of the genus *Tropicoperdix* Blyth 1859 (Phasianidae, Aves) using multilocus sequence data. Zool. J. Linn. Soc. 175, 429–438.

Chen, M.-Y., Liang, D., Zhang, P., 2017. Phylogenomic resolution of the phylogeny of Laurasiatherian mammals: exploring phylogenetic signals within coding and non-coding sequences. Genome Biol. Evol. 9, 1998–2012.

Chifman, J., Kubatko, L., 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30, 3317–3324.

Chifman, J., Kubatko, L., 2015. 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, 35–47.

Darriba, D., Weiß, M., Stamatakis, A., 2016. Prediction of missing sequences and branch lengths in phylogenomic data. Bioinformatics 32, 1331–1337.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., Blaxter, M.L., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12, 499.

Davison, G.W.H., 1982. Systematics within the genus *Arborophila* Hodgson. Federal Museum J. 27, 125–134.

del Hoyo, J., Collar, N.J., 2014. HBW and Birdlife International Illustrated Checklist of the Birds of the World. vol. 1: Non-passerines. Lynx Edicions, Barcelona.

Dickinson, E.C., Remsen, J.V., 2013. The Howard and Moore Complete Checklist of the Birds of the World, fourth ed. Aves Press, Eastbourne, U.K.

Edwards, S.V., Cloutier, A., Baker, A.J., 2017. Conserved nonexonic elements: a novel class of marker for phylogenomics. Syst. Biol. 66, 1028–1044.

Faircloth, B.C., 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics 32, 786–788.

Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61, 717–726.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C., Peterson, A.T., 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol. Evol. 19, 497–503.

Haussler, D., O'Brien, S.J., Ryder, O.A., Barker, F.K., Clamp, M., Crawford, A.J., Hanner, R., Hanotte, O., Johnson, W.E., McGuire, J.A., Miller, W., Murphy, R.W., Murphy, W.J., Sheldon, F.H., Sinervo, B., Venkatesh, B., Wiley, E.O., Allendorf, F.W., Amato, G., Baker, C.S., Bauer, A., Beja-Pereira, A., Bermingham, E., Bernardi, G., Bonvicino, C.R., Brenner, S., Burke, T., Cracraft, J., Diekhans, M., Edwards, S., Ericson, P.G.P., Estes, J., Fjelsda, J., Flesness, N., Gamble, T., Gaubert, P., Graphodatsky, A.S., Graves, J.A.M., Green, E.D., Green, R.E., Hackett, S., Hebert, P., Helgen, K.M., Joseph, L., Kessing, B., Kingsley, D.M., Lewin, H.A., Luikart, G., Martelli, P., Moreira, M.A.M., Nguyen, N., Orti, G., Pike, B.L., Rawson, D.M., Schuster, S.C., Seuanez, H.N., Shaffer, H.B., Springer, M.S., Stuart, J.M., Sumner, J., Teeling, E., Vrijenhoek, R.C., Ward, R.D., Warren, W.C., Wayne, R., Williams, T.M., Wolfe, N.D., Zhang, Y.P., Graph-Odatsky, A., Johnson, W.E., Felsenfeld, A., Turner, S., Genome, K.C.S., Mammals, G., Birds, G., Amphibians Reptiles, G., Fishes, G., General Policy, G., Anal, G., 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. J. Hered. 100, 659–674.

Hawkins, M.T.R., Leonard, J.A., Helgen, K.M., McDonough, M.M., Rockwood, L.L., Maldonado, J.E., 2016. Evolutionary history of endemic Sulawesi squirrels constructed from UCEs and mitogenomes sequenced from museum specimens. BMC Evol. Biol. 16, 80.

Hosner, P.A., Braun, E.L., Kimball, R.T., 2016a. Rapid and recent diversification of curassows, guans, and chachalacas (Galliformes: Cracidae) out of Mesoamerica: phylogeny inferred from mitochondrial, intron, and ultraconserved element sequences. Mol. Phylogenet. Evol. 102, 320–330.

Hosner, P.A., Faircloth, B.C., Glenn, T.C., Braun, E.L., Kimball, R.T., 2016b. Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes). Mol. Biol. Evol. 33, 1110–1125.

Hosner, P.A., Tobias, J.A., Braun, E.L., Kimball, R.T., 2017. How do seemingly non-vagile clades accomplish trans-marine dispersal? Trait and dispersal evolution in the landfowl (Aves: Galliformes). Proc. R. Soc. B: Biol. Sci. 284.

Hung, C.-M., Lin, R.-C., Chu, J.-H., Yeh, C.-F., Yao, C.-J., Li, S.-H., 2013. The De Novo assembly of mitochondrial genomes of the extinct passenger pigeon (*Ectopistes migratorius*) with next generation sequencing. PLoS ONE 8, e56301.

Hung, C.-M., Shaner, P.-J.L., Zink, R.M., Liu, W.-C., Chu, T.-C., Huang, W.-S., Li, S.-H.,

2014. Drastic population fluctuations explain the rapid extinction of the passenger pigeon. Proc. Natl. Acad. Sci. 111, 10636–10641.
Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., Ho, S.Y.W., Faircloth, B.C., Nabholz, B., Howard, J.T., Suh, A., Weber, C.C., da Fonseca, R.R., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M.S., Zavidovych, V., Subramanian, S., Gabaldón, T., Capella-Gutiérrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W.C., Ray, D., Green, R.E., Bruford, M.W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E.P., Bertelsen, M.F., Sheldon, F.H., Brumfield, R.T., Mello, C.V., Lovell, P.V., Wirthlin, M., Schneider, M.P.C., Prosdocimi, F., Samaniego, J.A., Velazquez, A.M.V., Alfaro-Núñez, A., Campos, P.F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D.M., Zhou, Q., Perelman, P., Driskell, A.C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F.E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F.K., Jønsson, K.A., Johnson, W., Koepfli, K.-P., O'Brien, S., Haussler, D., Ryder, O.A., Rahbek, C., Willerslev, E., Graves, G.R., Glenn, T.C., McCormack, J., Burt, D., Ellegren, H., Alström, P., Edwards, S.V., Stamatakis, A., Mindell, D.P., Cracraft, J., Braun, E.L., Warnow, T., Jun, W., Gilbert, M.T.P., Zhang, G., 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346, 1320–1331.
Jetz, W., Thomas, G.H., Joy, J.B., Hartmann, K., Mooers, A.O., 2012. The global diversity of birds in space and time. Nature 491, 444.
Kanda, K., Pflug, J.M., Sproul, J.S., Dasenko, M.A., Maddison, D.R., 2016. Successful recovery of nuclear protein-coding genes from small insects in museums using illumina sequencing. PLoS ONE 10, e0143929.
Kimball, R.T., Braun, E.L., 2008. A multigene phylogeny of Galliformes supports a single origin of erectile ability in non-feathered facial traits. J. Avian Biol. 39, 438–445.
Kimball, R.T., Braun, E.L., 2014. Does more sequence data improve estimates of galliform phylogeny? Analyses of a rapid radiation using a complete data matrix. PeerJ 2, e361.
Knapp, M., Hofreiter, M., 2010. Next generation sequencing of ancient DNA: requirements, strategies and perspectives. Genes 1, 227–243.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.
Lemmon, A.R., Emme, S.A., Lemmon, E.M., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61, 727–744.
Long, C., Kubatko, L., 2017. Identifiability and Reconstructibility of Species Phylogenies under a Modified Coalescent. arXiv preprint arXiv:1701.06871.
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.
McCormack, J.E., Tsai, W.L.E., Faircloth, B.C., 2016. Sequence capture of ultraconserved elements from bird museum specimens. Mol. Ecol. Resour. 16, 1189–1203.
Meiklejohn, K.A., Danielson, M.J., Faircloth, B.C., Glenn, T.C., Braun, E.L., Kimball, R.T., 2014. Incongruence among different mitochondrial regions: a case study using complete mitogenomes. Mol. Phylogenet. Evol. 78, 314–323.
Meiklejohn, K.A., Faircloth, B.C., Glenn, T.C., Kimball, R.T., Braun, E.L., 2016. Analysis of a rapid evolutionary radiation using ultraconserved elements: evidence for a bias in some multispecies coalescent methods. Syst. Biol. 65, 612–627.
Murray, G.G.R., Soares, A.E.R., Novak, B.J., Schaefer, N.K., Cahill, J.A., Baker, A.J., Demboski, J.R., Doll, A., Da Fonseca, R.R., Fulton, T.L., Gilbert, M.T.P., Heintzman, P.D., Letts, B., McIntosh, G., O'Connell, B.L., Peck, M., Pipes, M.-L., Rice, E.S., Santos, K.M., Sohrweide, A.G., Vohr, S.H., Corbett-Detig, R.B., Green, R.E., Shapiro, B., 2017. Natural selection shaped the rise and fall of passenger pigeon genomic diversity. Science 358, 951–954.
Persons, N.W., Hosner, P.A., Meiklejohn, K.A., Braun, E.L., Kimball, R.T., 2016. Sorting out relationships among the grouse and ptarmigan using intron, mitochondrial, and ultra-conserved element sequences. Mol. Phylogenet. Evol. 98, 123–132.
Prum, R.O., Berv, J.S., Dornburg, A., Field, D.J., Townsend, J.P., Lemmon, E.M., Lemmon, A.R., 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526, 569–573.
Pyron, R.A., Burbrink, F.T., Wiens, J.J., 2013. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. BMC Evol. Biol. 13, 93.
Reddy, S., Kimball, R.T., Pandey, A., Hosner, P.A., Braun, M.J., Hackett, S.J., Han, K.-L., Harshman, J., Huddleston, C.J., Kingston, S., 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. Syst. Biol. syx041.
Rocha, L.A., Aleixo, A., Allen, G., Almeda, F., Baldwin, C.C., Barclay, M.V.L., Bates, J.M., Bauer, A.M., Benzoni, F., Berns, C.M., Berumen, M.L., Blackburn, D.C., Blum, S., Bolaños, F., Bowie, R.C.K., Britz, R., Brown, R.M., Cadena, C.D., Carpenter, K., Ceríaco, L.M., Chakrabarty, P., Chaves, G., Choat, J.H., Clements, K.D., Collette, B.B., Collins, A., Coyne, J., Cracraft, J., Daniel, T., de Carvalho, M.R., de Queiroz, K., Di Dario, F., Drewes, R., Dumbacher, J.P., Engilis, A., Erdmann, M.V., Eschmeyer, W., Feldman, C.R., Fisher, B.L., Fjeldså, J., Fritsch, P.W., Fuchs, J., Getahun, A., Gill, A., Gomon, M., Gosliner, T., Graves, G.R., Griswold, C.E., Guralnick, R., Hartel, K., Helgen, K.M., Ho, H., Iskandar, D.T., Iwamoto, T., Jaafar, Z., James, H.F., Johnson, D., Kavanaugh, D., Knowlton, N., Lacey, E., Larson, H.K., Last, P., Leis, J.M., Lessios, H., Liebherr, J., Lowman, M., Mahler, D.L., Mamonekene, V., Matsuura, K., Mayer, G.C., Mays, H., McCosker, J., McDiarmid, R.W., McGuire, J., Miller, M.J., Mooi, R., Mooi, R.D., Moritz, C., Myers, P., Nachman, M.W., Nussbaum, R.A., Foighil, D.Ó., Parenti, L.R., Parham, J.F., Paul, E., Paulay, G., Pérez-Emán, J., Pérez-Matus, A., Poe, S., Pogonoski, J., Rabosky, D.L., Randall, J.E., Reimer, J.D., Robertson, D.R., Rödel, M.-O., Rodrigues, M.T., Roopnarine, P., Rüber, L., Ryan, M.J., Sheldon, F., Shinohara, G., Short, A., Simison, W.B., Smith-Vaniz, W.F., Springer, V.G., Stiassny, M., Tello, J.G., Thompson, C.W., Trnski, T., Tucker, P., Valqui, T., Vecchione, M., Verheyen, E., Wainwright, P.C., Wheeler, T.A., White, W.T., Will, K., Williams, J.T., Williams, G., Wilson, E.O., Winker, K., Winterbottom, R., Witt, C.C., 2014. Specimen collection: an essential tool. Science 344, 814–815.
Rowe, K.C., Singhal, S., Macmanes, M.D., Ayroles, J.F., Morelli, T.L., Rubidge, E.M., Bi, K.E., Moritz, C.C., 2011. Museum genomics: low-cost and high-accuracy genetic data from historical specimens. Mol. Ecol. Resour. 11, 1082–1092.
Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. Genome Res.
Sproul, J.S., Maddison, D.R., 2017. Sequencing historical specimens: successful preparation of small specimens with low amounts of degraded DNA. Mol. Ecol. Resour. 17, 1183–1201.
Staats, M., Erkens, R.H.J., van de Vossenberg, B., Wieringa, J.J., Kraaijeveld, K., Stielow, B., Geml, J., Richardson, J.E., Bakker, F.T., 2013. Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. PLoS ONE 8, e69189.
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.
Straub, S.C.K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R.C., Liston, A., 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. Am. J. Bot. 99, 349–364.
Swofford, D.L., 2003. PAUP*: Phylogenetic Analysis using Parsimony, Version 4.0 b10. Sinauer Associates, Inc, Sunderland (MA).
Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., Earl, A.M., 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9, e112963.
Wang, N., Kimball, R.T., Braun, E.L., Liang, B., Zhang, Z., 2013. Assessing phylogenetic relationships among galliformes: a multigene phylogeny with expanded taxon sampling in Phasianidae. PLoS ONE 8, e64312.
Wang, N., Hosner, P.A., Liang, B., Braun, E.L., Kimball, R.T., 2017. Historical relationships of three enigmatic phasianid genera (Aves: Galliformes) inferred using phylogenomic and mitogenomic data. Mol. Phylogenet. Evol. 109, 217–225.
Wood, H.M., González, V.L., Lloyd, M., Coddington, J., Scharff, N., 2018. Next-generation museum genomics: Phylogenetic relationships among palpimanoid spiders using sequence capture techniques (Araneae: Palpimanoidea). Phylogenet. Evol. Mol.
Zhang, G., 2015. Genomics: Bird sequencing project takes off. Nature 522 34–34.
Zimmer, E.A., Wen, J., 2015. Using nuclear gene data for plant phylogenetics: progress and prospects II. Next-gen approaches. J. Systemat. Evol. 53, 371–379.