MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) **15**, 893–902 doi: 10.1111/1755-0998.12365

Next-generation sequencing workflow for assembly of nonmodel mitogenomes exemplified with North Pacific albatrosses (*Phoebastria* spp.)

Z. T. LOUNSBERRY,* S. K. BROWN,* P. W. COLLINS,† R. W. HENRY,‡ S. D. NEWSOME§ and B. N. SACKS*¶

*Mammalian Ecology and Conservation Unit, Veterinary Genetics Laboratory, School of Veterinary Medicine, University of California, Davis, One Shields Avenue/Old Davis Rd., Davis, CA 95616, USA, †Santa Barbara Museum of Natural History, 2559 Puesta Del Sol, Santa Barbara, CA 93105, USA, ‡Institute for Marine Sciences, Center for Ocean Health, Long Marine Lab, University of California, Santa Cruz, 100 Shaffer Road, Santa Cruz, CA 95060, USA, §Department of Biology, University of New Mexico, 167 Castetter Hall, MSC03 2020, Albuquerque, NM 87131, USA, ¶Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis, One Shields Avenue/Old Davis Rd., Davis, CA 95616, USA

Abstract

Use of complete mitochondrial genomes (mitogenomes) can greatly increase the resolution achievable in phylogeographic and historical demographic studies. Using next-generation sequencing methods, it is now feasible to efficiently sequence mitogenomes of large numbers of individuals once a reference mitogenome is available. However, assembling the initial mitogenomes of nonmodel organisms can present challenges, for example, in birds, where mtDNA is often subject to gene rearrangements and duplications. We developed a workflow based on Illumina paired-end, whole-genome shotgun sequencing, which we used to generate complete 19-kilobase mitogenomes for each of three species of North Pacific albatross, a group of birds known to carry a tandem duplication. Although this duplication had been described previously, our procedure did not depend on this prior knowledge, nor did it require a closely related reference mitogenome (e.g. a mammalian mitogenome was sufficient). We employed an iterative process including de novo assembly, reference-guided assembly and gap closing, which enabled us to detect duplications, determine gene order and identify sequence for primer positioning to resolve any mitogenome ambiguity (via minimal targeted Sanger sequencing). We present full mtDNA annotations, including 22 tRNAs, 2 rRNAs, 13 protein-coding genes, a control region and a duplicated feature for all three species. Pairwise comparisons supported previous hypotheses regarding the phylogenetic relationships within this group and occurrence of a shared tandem duplication. The resulting mitogenome sequences will enable rapid, high-throughput NGS mitogenome sequencing of North Pacific albatrosses via direct reference-guided assembly. Moreover, our approach to assembling mitogenomes should be applicable to any taxon.

Keywords: black-footed albatross, de novo assembly, Laysan albatross, MiSeq, mtDNA, short-tailed albatross

Received 10 April 2014; revision received 12 December 2014; accepted 18 December 2014

Introduction

Mitochondrial DNA (mtDNA) is an essential molecular genetic tool for understanding phylogenetic history and population structure. Traditionally, studies used 200–1000 base pairs (bp) of sequence from the control region (CR) or coding regions such as the cytochrome *b* gene (Avise *et al.* 1987). Sequencing complete mitochondrial genomes (mitogenomes) has become increasingly feasible and can provide significantly more resolution for finer-scale spatial inferences over relatively recent time frames (Morin *et al.* 2010; Knaus *et al.* 2011). Use of complete mitogenomes as an alternative to smaller fragments also enables tests for selection on particular mtDNA genes and use of differential evolution among genes, gene regions and sites within codons for more accurate phylogenetic reconstruction (Ingman & Gyllensten 2007).

The enhanced resolution associated with whole mitogenomes can be especially useful in taxa where duplications affect the CR or adjacent regions, which can complicate traditional Sanger sequencing of those regions; such duplications occur in a broad range of metazoan taxa (Eda *et al.* 2010). For example, identical or

Correspondence: Benjamin N. Sacks, Fax: 530-752-3556; E-mail: bnsacks@ucdavis.edu

partially identical mtDNA sequence paralogs are pervasive in Procellariiformes, occurring in several species of albatrosses (Diomedeidae; Abbott *et al.* 2005; Eda *et al.* 2010). The maintenance of identical sequences among part, but not all, of two duplicated features (mosaic gene conversion) also has been reported in Diomedeidae CRs, including those of all three species of North Pacific albatrosses investigated in this study (Eda *et al.* 2010). These characteristics of albatross mitogenomes make targeted Sanger sequencing a multistep, time-consuming process (Eda *et al.* 2010; Kuro-O *et al.* 2010). Moreover, the presence of a polycytosine repeat motif directly upstream of the CR, as well as secondary structures, make amplifying and sequencing a single copy of the CR technically challenging (King *et al.* 2014). One way to circumvent this problem is to target unduplicated regions of the genome, but this requires longer sequences to provide the same information content as the hypervariable segments (HV1 and HV2) of the CR and also prevents the possibility of comparing to previous studies that have used single copies of duplicated regions to draw phylogenetic inferences (Eda *et al.* 2010; Kuro-O *et al.* 2010).

Massively parallel (next generation) sequencing (NGS) has been used to sequence full mitogenomes of birds containing duplicated mtDNA features (e.g. Cooke *et al.* 2012). However, these studies have utilized long-read technology (e.g. reads from a Roche 454 platform) which are typically more expensive to generate per base and exhibit slightly higher rates of sequencing error than short-read technologies (e.g. Illumina MiSeq or HiSeq; Luo *et al.* 2012). Whole-genome shotgun (WGS) short-read libraries potentially can provide novel solutions to problems associated with Sanger sequencing mtDNA at a reduced cost relative to long-read libraries. By nonselectively sequencing billions of bp of genomic DNA, among which mtDNA are present at 100- to 1000-fold frequency relative to nuclear DNA, it is possible to produce overlapping sequencing reads from independent molecules corresponding to every position of the mitogenome at several-hundred-fold depth, ensuring complete mtDNA representation (King *et al.* 2014). In the absence of long, complex duplications, such data sets can be readily assembled without prior knowledge of the gene order (i.e. de novo) into a single continuous and complete sequence (Cooke *et al.* 2012; Barker 2014). Mitogenomes with longer and more complex duplications, however, could require more data processing and minimal targeted Sanger sequencing to determine positioning of particular elements. In this study, we developed a workflow involving use of short-read WGS libraries to characterize the complete mitogenomes, including duplicated regions, of the three North Pacific albatrosses: black-footed albatross (*Phoebastria nigripes*), Laysan albatross (*P. immutabilis*) and short-tailed albatrosses (*P. albatrus*). The impetus for this study was to provide references needed for future high-throughput mitogenome sequencing in these particular species. However, the workflow we developed for this purpose also provides a resource that should be applicable to a broader range of nonmodel taxa with or without duplications in their organellar DNA.

Methods

DNA extraction, library preparation and Illumina sequencing

We extracted genomic DNA from pectoral muscle tissue sampled from by-caught albatrosses from Hawaii (n = 1 Laysan, 1 black-footed) and Alaska (n = 2 short tailed) using a DNeasy Blood and Tissue kit (Qiagen) following manufacturer protocols. Prior to fragmentation, we quantified extracted DNA concentrations using a Nanodrop (Thermo Scientific) and standardized concentrations to 10 ng/μL via dilution in sterile water. We fragmented genomic DNA to approximately 550 bp via sonication using a Bioruptor® ultrasonicator (Diagenode). To confirm that DNA was fragmented properly, we visualized prefragmented and post-fragmented samples on an agarose gel and additionally quantified the fragmented aliquots on a 2100 Bioanalyzer using a DNA High Sensitivity assay (Agilent). Then, we performed library preparations using half-volume reactions in a NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs), including size selection for 500-bp fragments using AMPure magnetic beads (Agencourt), following manufacturer instructions. During library preparation, we individually barcoded samples for multiplexing using the NEBNext Multiplex Oligos for Illumina (New England Biolabs). Lastly, we pooled our barcoded libraries and performed 250-bp paired-end sequencing on a MiSeq™ platform (Illumina) at the University of California, Davis Genome Center Core Facility.

Read preprocessing

We demultiplexed samples and trimmed adapter sequence using MISEQ REPORTER version 2.3.32 (Illumina). We then paired forward and reverse mated reads with a minimum overlap of 20 bp using FLASH version 1.2.7 (Magoc & Salzberg 2011). Finally, we trimmed sequencing artefacts from the 5′ ends, trimmed poor-quality 3′ sequence and filtered the paired reads for overall quality (remaining reads with >50% bases having quality score <2 were discarded) using NGSSHORT (Chen *et al.* 2013). We used these quality-filtered, mated reads in all subsequent analyses.

Identifying duplication and sequence assembly

We approached mitogenome assembly for each individual ($n = 4$) of the three species using a workflow we designed to handle genome duplications in previously unsequenced mitogenomes (code available in Appendix S1, Supporting information). To ensure that our assembly was independent of prior knowledge of the duplication in the *Phoebastria* mitogenome (Eda *et al.* 2010; Kuro-O *et al.* 2010), we relied entirely on the reads generated in this study and guidance from orthologous mitogenomes. Initially, we used for this purpose the mitogenome of a confamilial albatross, *Thalassarche melanophris* (AY158677.2; Slack *et al.* 2006), but also explored whether a more distantly related mitogenome could suffice (described below). Although the *T. melanophris* mitogenome contained a duplication as well (Abbott *et al.* 2005), it differed from that of the North Pacific albatrosses in that the *T. melanophris* cytochrome *b* pseudogene at the start site of the duplication was substantially shorter than that of *Phoebastria* spp.

We used an iterative de novo approach to assemble sequences (Fig. 1; Appendix S1, Supporting information). For the initial (tentative) assembly, we extracted all reads that aligned to the *T. melanophris* mitogenome by combining the '–very-sensitive' and '—no-unal' parameterization in BOWTIE2 (Langmead *et al.* 2009), but reserved nonaligning reads for subsequent steps. We converted the resulting sequence alignment/map (SAM) files to FASTA format using a combination of SAMTOOLS (Li *et al.* 2009) and command line functions in AWK. These sequence files were significantly reduced in size relative to the entire read data set, enabling rapid de novo assembly of the mtDNA into contigs (i.e. consensus sequences derived from overlapping reads) using a greedy assembler in CAP3 (Huang & Madan 1999). We chose CAP3 for the assembly because it does not rely on specifying a k-mer length and can be run as part of a loop among several samples more easily than an assembler that requires an optimized k-mer length. After confirming orthology of the de novo assembled contigs to avian mtDNA using a Basic Local Alignment Search Tool (BLASTN), we aligned reads back to them and visually checked for accuracy and even read depth in TABLET version 1.13.12.17 (Milne *et al.* 2013). During this step, we removed poorly assembled contigs (<10 reads aligning or a nucleotide mismatch frequency >0.6%) and manually trimmed 3′ or 5′ ends covered by a single read.

To obtain a tentative gene order, we aligned contigs by hand using BLASTN results and a standard avian gene order (Abbott *et al.* 2005). Because the regions between contigs were not known (i.e. were not assembled de

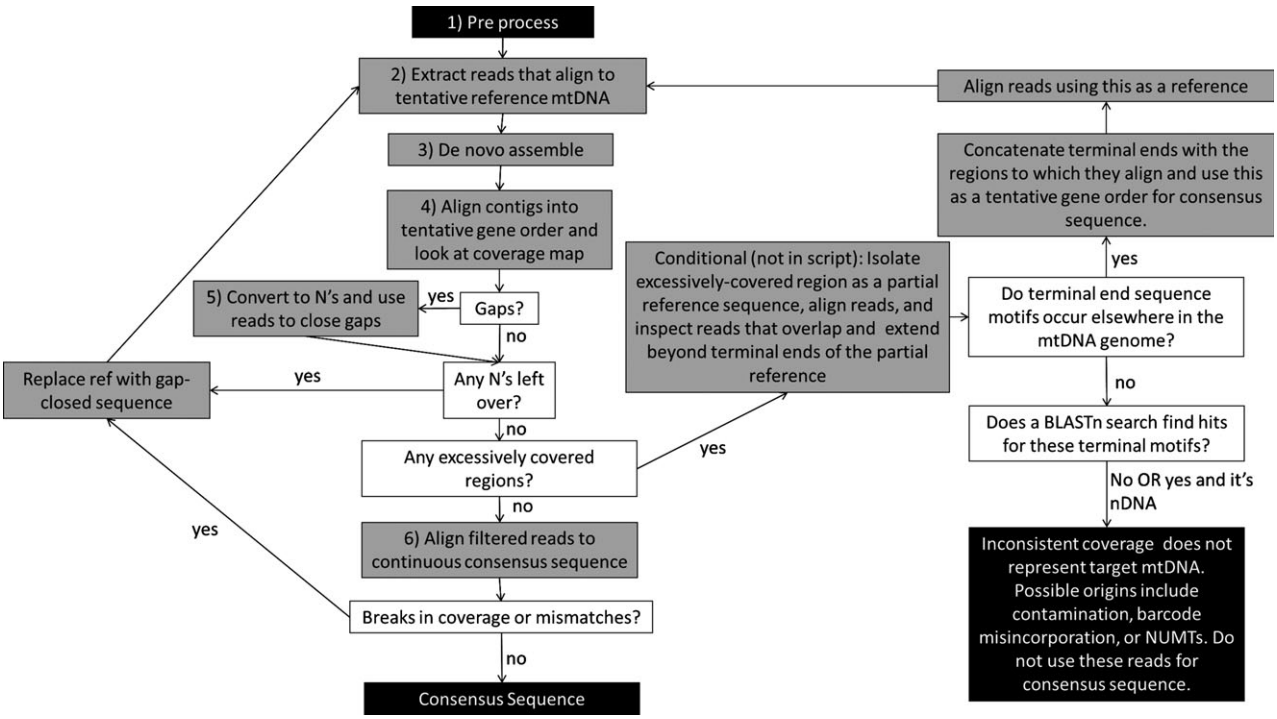

Fig. 1 Flowchart of assembly workflow. Grey boxes correspond to specific steps (with numbers corresponding to their respective command line codes, where applicable; Appendix S1, Supporting information); white boxes correspond to questions asked before and after steps; and black boxes correspond to the start and end points of the workflow.

novo), we left gaps between contigs and attempted to fill them using that individual's full filtered read set, including those reads not initially aligning to the *T. melanophris* reference. Specifically, we manually created a single scaffold sequence in BIOEDIT version 7.1.3.0 (Hall 1999) for each individual based on this tentative gene order, filled gaps between contigs with N's, and entered each individual's scaffold and corresponding reads into the gap closing algorithm within SOAPDENOVO (GapCloser; Li *et al.* 2010). We aligned reads back to the tentative consensus sequence output from GAPCLOSER using BOWTIE2 and inspected the alignment for remaining gaps and evenness of read depth in TABLET.

To discover duplications, we inspected excessively covered regions, which often indicate spurious alignment of paralogous loci (Medvedev *et al.* 2009). Because paralogous features start and end at different parts of the genome and thus each copy should have different flanking sequences, we focused on reads aligning at the 5′ and 3′ ends of overrepresented regions. Specifically, we constructed reference sequences from contigs abutting flanking regions (i.e. contigs representing the 5′ and 3′ terminating ends of excessively covered regions) and inserted a string of 'N's extending into the flanking regions. This allowed reads with flanking sequence not initially aligning to the *T. melanophris* sequence to align to the 'N's and enabled a BLASTN search to guide positioning of paralogs. Once boundaries were identified, we inferred the correct placement of the duplication and used this information to construct a consensus sequence manually with a newly inferred gene order (i.e. including the tentative duplication). We then used it as a reference in place of the previous sequence (which was constructed based on a standard avian mtDNA gene order), in another round of alignment and assembly. We accepted final consensus sequences when all mtDNA reads aligned to the consensus sequences with no mismatching motifs or gaps. Any remaining ambiguities in duplicated regions could then be addressed through Sanger sequencing.

To investigate the applicability of this workflow to nonmodel organisms with no available closely related reference mitogenome, we performed the same analysis on each sample, but replacing the *T. melanophris* reference with a mouse (*Mus musculus*) mitogenome (GenBank Accession no. AP013031). Because of the evolutionary distance between *Mus* and *Phoebastria*, we ran an iterative preliminary step to extract as many mtDNA reads as possible for de novo assembly (Appendix S1, Supporting information). To this end, we relaxed alignment parameters by adding the '-L 10' and '–very-sensitive-local' parameters in BOWTIE2. We then de novo assembled the reads we extracted and used them as seeds for 5′- and 3′-end extension using the targeted assembly software MAPSEMBLER2 and each individual's read data set (Peterlongo & Chikhi 2012). From this point on, the procedure was the same as described above and resulted in the same contiguous consensus sequences.

Sanger sequencing

We applied Sanger sequencing to resolve the positioning ambiguity between paralogs. We conducted polymerase chain reaction (PCR) using the following primer pairs: Lcyt246.dio (5′-CCTCCACGCAAACGGAG-3′; Eda *et al.* 2006) and AlbCblock2.R (5′-GTTGCTGATTTCTCGTG AG-3′, designed in this study). The PCR mixtures contained 1 μL genomic DNA extract, 1× PCR Buffer (Life Technologies), 1.5 mM $MgCl_2$, 2 mM each dNTPs, 1mg/ml bovine serum albumen (New England Biolabs), 0.075μM forward and reverse primers and 0.1 U of Amplitaq (Life Technologies). We amplified DNA using a thermal profile consisting of one 10 min cycle at 94 °C followed by 30 cycles combining a 30 s denaturation step at 94 °C, a 30 s annealing step at 60 °C and a 90 s extension at 74 °C followed by a final extension step for 10 min at 74 °C.

We attempted to sequence amplicons in both directions using an internal primer modified from Kuro-O *et al.* (2010; AlbCR1.F: 5′-AGACTTGCGGCCTGAAAA AC-3′) and the same reverse primer used in the PCR (see Results for further explanation). We sequenced PCR products using a BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) electrophoresed on an ABI 3730 DNA sequencer and visualized chromatograms using Bioedit. (Note: although we made use of some previously published primers for Sanger sequencing, we could have just as easily designed all primers directly from conserved segments of the NGS assemblies.)

Annotation and mitogenome comparisons

We annotated our consensus mitogenomes using the Dual Organellar Genome Annotator (DOGMA; Wyman *et al.* 2004) and compared the annotation to the *T. melanophris* mitogenome for additional reference. After annotating and trimming the small, tandem repeat region after the 5′ end of CR2 (because of low confidence in assembly accuracy associated with small repeat regions), we aligned the sequences to one another using the multiple sequence comparison algorithm in MEGA6 (Tamura *et al.* 2013). We then divided consensus sequences into their respective annotated mtDNA features and calculated base pair composition for each feature in MEGA6. For direct pairwise comparisons, we calculated the number of transitions and transversions among species for each feature in the 'PopGenome' package in R (R Core

Team 2012; Pfeifer *et al*. 2014) and the synonymous–non-synonymous substitution (K_a/K_s) ratio for all protein-coding genes in DNAsp (Librado & Rozas 2009). We then calculated approximate pairwise divergences based on sequence differences (number of mutations/sequence length). To compare CR directly within and among species, we estimated the best fit nucleotide substitution model in MEGA6 and, using this model, computed a maximum-likelihood tree (with 500 bootstrap replicates) of a relevant 223 bp segment of HV1.

Results

Mitogenome assemblies

After adapter trimming, mate pairing and filtering for quality, we obtained 7.71 million paired reads averaging 362 bp (range: 233–452 bp) in length for analysis. On average among individuals, 1.81% (range: 1.28–2.27%) of reads aligned to the *T. melanophris* reference sequence, resulting in an average per site, per individual read depth of 694X (range: 127–1297X; Fig. 2a). During this step, we identified a region in these initial assemblies with approximately twice the read depth as other regions, presumably reflecting a duplication (Fig. 2a). A BLASTN search revealed that these regions corresponded to the mtDNA CR as well as to portions of cytochrome *b*, NADH dehydrogenase 6 (ND6) and associated tRNAs. Closer inspection of the reads aligned to the double-covered locations further revealed a mismatching motif between approximately half of the reads corresponding to HV1 of the CR (Fig. 2b). We hypothesized that the initial de novo assembly failed to distinguish two duplicated paralogs, which were identical in sequence for most of the duplication (i.e. except a segment towards the 5′ terminating end of the CR [HV1]), and spuriously collapsed them into a single contig.

Assuming the double-covered region corresponded to a duplication, our alignment should have systematically excluded most of the reads containing the boundaries of the duplication due to partial misalignment. To test this, we determined the start site of the putative duplication by examining the 5′ terminating end of the region where read depth doubled, which was approximately 600 bp into the cytochrome *b* gene (on the *T. melanophris* reference), continuing downstream through the CR. We created a FASTA file from the ~2-kbp conserved, presumably duplicated sequence from cytochrome *b* to the CR, added N's to 5′ and 3′ ends, and used this sequence as a reference to extract additional reads overlapping putative boundaries (Fig. 2c). As expected, there were two distinct sequence motifs at the 5′ terminating end of the conserved duplicated reference. Based on a BLASTN search of reads containing each motif, one of these motifs aligned to an upstream portion of the cytochrome *b* gene and the other one aligned to the 3′ end of the CR. Reads with motif 1 mapped seamlessly to a complete and functional cytochrome *b* gene and reads with motif 2 revealed a partial cytochrome *b* gene (i.e. beginning ~600 bp into the gene sequence) directly following the CR. Whereas Fig. 2c illustrates only the reads at the 5′ end of the duplication, we used the same procedure to determine

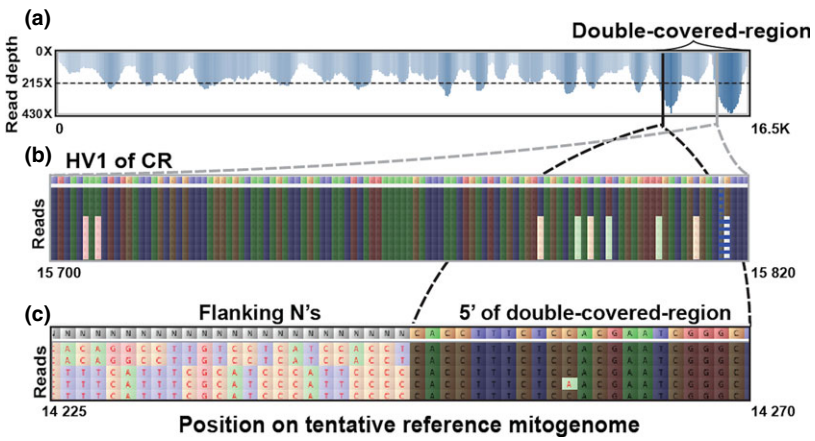

Fig. 2 Visualization of duplication discovery steps as illustrated with sample BFAL1 (which had the fewest reads), including (a) coverage map, illustrating read depth corresponding to sites along the initial tentative consensus sequence after the first iteration of assembly and gap closure, with suspected region of gene duplication/paralogs indicated; (b) alignment of reads to the HV1 portion of the CR within the tentative consensus sequence, illustrating two distinct sets of sequence motifs (mismatches to reference CR contig shown as highlighted nucleotides and sorted for ease of reading) within this double-covered segment; (c) alignment of reads mapping to the start site of the duplication, illustrating the 5′ end of this segment with two distinct motifs (highlighted) and the 3′ end (~600 bp into the cytochrome *b* gene) with identical read sequences (not highlighted).

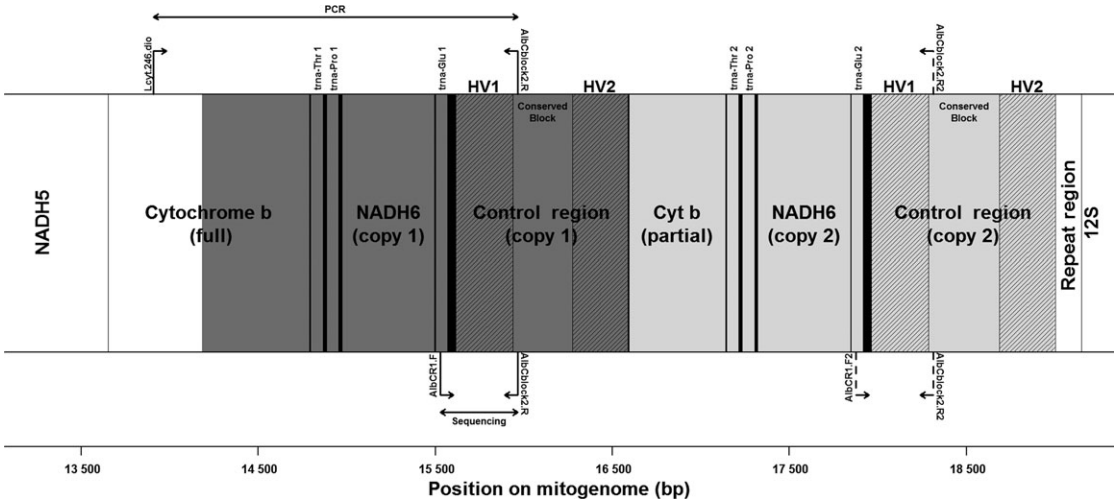

Fig. 3 Gene map illustrating the gene order of the duplicated mtDNA features. Unduplicated regions are shown in white; duplications are shown in dark and light grey (with HV1 and HV2 copies cross-hatched); and nonannotated regions are shown in black. Copies of tRNAs are indicated by their respective amino acids. PCR priming sites are labelled with arrows and numbers (1 = Lcyt246.dio, 2 = AlbCblock2.R) above the gene map and internal sequencing primers are labelled with arrows and numbers (2 = AlbCblock2.R, 3 = AlbCR1.F) below the gene map, including both targeted (solid-lined arrows) and incidental (dash-lined arrows) priming sites. PCR and sequencing regions are highlighted with labelled arrows, and the scale bar indicates position along the final consensus mitogenome.

that the 3′ end of the duplicated region (associated with the CR) abutted against two different motifs: one aligning to part of the cytochrome *b* gene and one containing the repeat region between the CR and 12s rRNA. This pattern indicated that the duplicated region spanned ~2 kbp beginning with a downstream portion of the cytochrome *b* gene (pseudogene), through complete copies of ND6 and all associated tRNAs, ending with a complete copy of the CR (Fig. 3).

The only ambiguity that remained unresolved was the order of two HV1 paralogs within the duplicate CRs; this was because the length of identical sequence surrounding these variants was greater than our maximum paired read length. We therefore attempted to Sanger sequence this region by first amplifying the larger region containing only one variant (HV1 from CR1) using primers Lcyt246.dio and AlbCblock2.R, and then sequencing the portion of this amplicon corresponding to the HV1 variant using primers AlbCR1.F and AlbCblock2.R (Fig. 3). Despite numerous attempts for each individual under a range of PCR conditions (including multiple replications under identical conditions), we succeeded in producing usable sequences only for the black-footed albatross sample (BFAL1, used for NGS) and one additional unrelated black-footed albatross (BFAL2, GenBank Accession nos. KM878669–KM878670), but not in the other two species. Our read depth from the WGS data representing each individual was also significantly lower in this region (minimum read depth per individual spanning this region ranged from 15X in black-footed to 71X in Laysan), presumably due to similar biochemical issues (see Discussion). Nevertheless, we obtained two sequences via Sanger methods enabling us to determine the correct positioning of HV1 in the black-footed albatross CRs. We relied on a 3-bp motif common to one copy of the HV1 paralogs of all three species (AAA in HV1-CR1 and GAG in HV1-CR2 in all three species) to infer the position in the other two species for which we were unable to Sanger sequence the region directly.

To finalize mitogenome assembly under the model proposed and supported above, we returned to the full set of reads to recover any that were left out due to partial mismatches with the *T. melanophris* mitogenome for reads flanking the duplication. To do this, we attempted to manually construct a final consensus sequence containing both duplications. Because assembly algorithms could not distinguish between paralogs, we aligned these contigs by hand in Bioedit into a new gene order that included both copies of the duplicated region as inferred above (Fig. 3), yielding our hypothetical consensus sequence. Lastly, we aligned reads against the final hypothetical consensus sequence and observed no unevenness or gaps in coverage and no mismatching bases other than sequencing errors and a heteroplasmic site in the CR for the entire ~19-kbp mitogenome (Fig. 4).

We repeated the procedure above substituting the mouse mitogenome for the *T. melanophris* mitogenome as the initial reference, which resulted in 0.25% (range: 0.18–0.28%) of total albatross reads aligning directly. Although this represented approximately 7-fold fewer

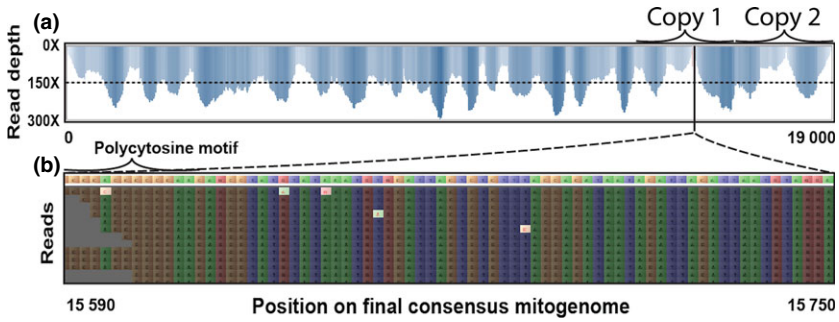

Fig. 4 Alignment to BFAL1 reads to its final consensus sequence showing (a) a coverage map with both copies of the duplication indicated; and (b) magnification of the reads aligning to 5′ end of the first CR, illustrating no systematically mismatching bases (i.e. other than the expected occasional sequencing errors), which resolved misalignment issues associated with paralogous sequence. Areas of low coverage (~10×) represent the 5′ terminating ends of the CR, which contains a polycytosine repeat motif (indicated above read set) presumably interfering with the sequencing technology.

reads than aligned to the more closely related reference, the resulting contigs provided sufficient seeds for the workflow to ultimately produce identical alignments after four or fewer iterations.

Having completed assembly of these mitogenomes, including correct HVl paralog order (GenBank Accession nos.: KJ735512–KJ735514), we compared the total number of reads aligning back to them to those aligning to the reference mitogenomes initially used in our workflow (Fig. 4). The average proportion of reads that aligned to the *T. melanophris* reference sequence (1.81%) represented, on average, 75% of the total mtDNA reads (2.41% of reads, range: 1.69–3.03%) ultimately mapping to the final, individual-specific consensus sequences. In comparison, the 0.25% of reads mapping to the mouse mitogenome represented a much smaller proportion (10.4%) of the total number of reads ultimately mapping to the final consensus sequences. The ratio of read depths between the initial reference and final consensus mitogenome sequences showed a similar pattern. For example, the average read depth aligning to the *T. melanophris* reference was 694X (range: 127–1297X), which was approximately 77% of the 893X (range: 162–1685X) average read depth aligning to the final consensus mitogenomes.

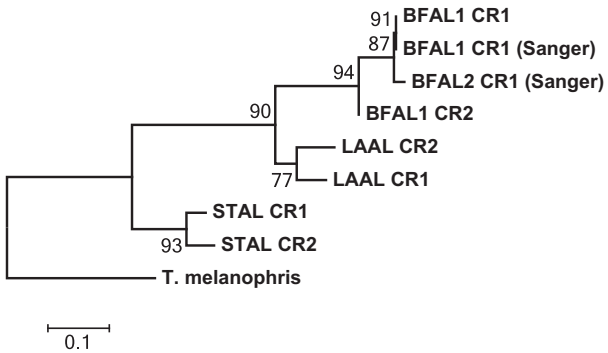

Fig. 5 Maximum-likelihood tree estimated under the HKY+G substitution model (determined to have the best fit), illustrating divergence in a 223-bp segment of HV1 between the two paralogous control regions within and among individuals. Branch lengths indicate sequence divergence and numerals represent percentage of 500 bootstrapped trees supporting nodes. Sequence data derived from non-MiSeq methods are included in parentheses. The *T. melanophris* reference refers to the GenBank sequence referenced in the text. The Laysan albatross (LAAL), short-tailed albatross (STAL) and black-footed albatross (BFAL) sample, BFAL1, were sequenced on the MiSeq lane in this study. BFAL2 was an additional individual that we Sanger sequenced.

Annotation and description

The two short-tailed albatross samples had identical mitogenomes, providing no information on intraspecific variability (but generally validating the accuracy of base calls). Among species, the same features/orthologs occurred in the same arrangement in all three *Phoebastria* spp. and, except for the duplication, those of confamilial albatrosses (partial *Diomedea chrysostoma*, GenBank Accession no. AP009193.1, Watanabe *et al.* 2006; and *T. melanophyris*). These features include 22 tRNAs, two rRNAs, 13 protein-coding genes and a CR (Table S1, Supporting information). The *Phoebastria* duplication was composed of a partial cytochrome *b* pseudogene, a full copy of NADH dehydrogenase 6, a full copy of the CR and three full copies of tRNAs. The partial cytochrome *b* pseudogene, duplicate NADH dehydrogenase 6 and all three duplicate tRNAs were identical to the corresponding sequence in the original features. The CRs within *Phoebastria* spp. differed partially between HV1 copies and were mostly identical at HV2 except at the 3′ end. Nevertheless, HV1 was more similar between CR1 and CR2 within species (i.e. individual representatives) than among species (Fig. 5). The NADH dehydrogenase

3 gene contained a noncoding frame-shift mutation at position 164 in all three species similar to previous observations in some other birds and in reptiles (Mindell *et al.* 1998). There was also evidence for two cytosine–thymine heteroplasmic sites in the black-footed (sites 15 739 and 18 086) and Laysan albatrosses (15 829 and 18 176) in the CR.

The full *Phoebastria* spp. mitogenomes contained 700 mutational sites among the three species, including 689 substitutions and 11 indels. Base pair composition (T: 24.4%, C: 31.0%, A: 30.6%, G: 14.0%) was nearly identical across species (Table S2, Supporting information). The number of transitions and transversions, as well as K_a/K_s ratio, varied depending on the mtDNA feature (Table S2, Supporting information). Pairwise comparisons of nucleotide substitutions revealed that black-footed and Laysan were approximately 1.8% divergent, black-footed and short-tailed were 2.8% divergent, and Laysan and short tailed were also 2.8% divergent.

Discussion

In the present study, we used a WGS approach to visualize and fully annotate mitogenomes for North Pacific albatrosses. Although minimal Sanger sequencing was required to determine the order of HV1 paralogs, our approach eliminated the need for multiple PCR steps (e.g. semi-nested PCR; Kuro-O *et al.* 2010) to isolate paralogous copies of the CR, and also produced sequences for the entire mitogenome, providing a framework for higher resolution comparative studies. Because an objective of this study was to assess the general utility of WGS sequencing for mitogenome assembly in nonmodel taxa, including those with duplications, we presented some steps (e.g. duplication discovery and targeted Sanger sequencing) that would have been unnecessary in the absence of duplications. Additionally, in the case of small duplication events or duplications for which paralogs do not evolve in concert, a simple de novo assembly approach could be sufficient (Cooke *et al.* 2012). Moreover, our application of the workflow using a reference mitogenome from an entirely different class of vertebrate (i.e. a mammal) demonstrated its broad utility to nonmodel taxa. Although we utilized knowledge of the class-wide (avian) gene order to assist in assembly, such information should be readily available for many taxa even where congeneric or confamilial references are lacking. Thus, our approach provided a framework for the assembly of simple mitogenomes as well as for discovery of putative duplication or rearrangement events in taxa that have previously presented challenges to full-mitogenome assembly (e.g. Barker 2014).

Our results also provided independent support for the conclusions of previous studies that used Sanger sequencing to reveal mosaic gene conversion in *Phoebastria* spp. mitogenomes (Eda *et al.* 2010). In the albatrosses sequenced in the present study, CR duplicates were more similar, but not identical, within species than among species, which is the characteristic of a gene family subject to mosaic gene conversion (Shao *et al.* 2005; Eda *et al.* 2010). Taking this characteristic of the *Phoebastria* spp. mitogenome (as well as those of seabirds with similar duplication events; Morris-Pocock *et al.* 2010) into consideration is critical when studying the CR in a phylogeographic context. Using NGS to sequence entire mitogenomes eliminates the need to focus on one of the two copies of the CR, allowing more robust phylogeographic inferences. Moreover, once a reference mitogenome sequence is available for a focal study species, the duplication-specific steps are no longer necessary and it becomes feasible to sequence many samples in a single lane and to use open-source, reference-guided consensus sequence generating software (e.g. SAMTOOLS' mpileup; Li *et al.* 2009) to rapidly assemble individual-specific mitogenomes.

Conversely, the technical difficulties in Sanger sequencing HV1 paralogs that we encountered make large-scale application of Sanger methods to multiple individuals at the very least a time- and resource-consuming task (Eda *et al.* 2010; Kuro-O *et al.* 2010). Based on PCR amplicon size (visualized via gel electrophoresis), we determined that we were eventually successful in isolating our target paralog within the duplicated elements (approximately 2 kbp spanning cytochrome *b* to the conserved block of CR1; Fig. 3). However, after many attempts under identical reaction conditions, we successfully Sanger sequenced portions of these amplicons in only two samples. We attribute the difficulty in sequencing this region to the presence of a polycytosine homopolymer and secondary structures directly adjacent to the target sequence, which often results in poor-quality data from Sanger (Butler 2005) and other sequencing approaches, including NGS (Luo *et al.* 2012; Knief 2014). Even using our short-read WGS approach, we observed exceptionally low read depth in this region consistent with patterns reported in other vertebrates sequenced using the same Illumina short-read platform (MiSeq; King *et al.* 2014). Ultimately, however, the Illumina-based approach applied in our study (and others; King *et al.* 2014) yielded sufficient read depths spanning the polycytosine homopolymer and containing all HV1 sequence data.

Due to excessive mitogenome coverage using a whole-genome shotgun approach, it is possible to get appropriate read depth for a moderate number of barcoded individual DNAs pooled in a single lane. Target-capture (e.g. with RNA baits) can further enrich libraries, enabling even larger numbers of barcoded individual

DNAs to be pooled and sequenced in a single lane (Hancock-Hanser *et al.* 2013). The mitogenome sequences presented here provide a template to design *Phoebastria* (or, more broadly, Diomedeidae or Procellariiformes)-specific bait probes, which could be used in large-scale, cross-species mtDNA sequencing efforts (Hancock-Hanser *et al.* 2013; Li *et al.* 2013). Mitogenome target enrichment also is used in studies that utilize ancient (archaeological) and historic (museum) specimens, which would make it possible to study temporal changes in mtDNA diversity directly over shallow evolutionary time scales (Enk *et al.* 2014; Mitchell *et al.* 2014).

Acknowledgements

We thank Shannon Fitzgerald (Alaska Fisheries Science Center – NOAA) and Jamie Marchetti (NOAA Fisheries –PIRO-HI) for collection of by-catch samples, Hannah Nevins (Wildlife Health Center, University of California, Davis; Oikonos) and Jessie Beck (Oikonos) for postmortem sampling, Mark Statham for assistance with laboratory procedures and the UCD Bioinformatics Core for assistance with data processing. Laysan and black-footed albatross were collected under MBTA #MB-052060-0. Short-tailed albatross were provided through a MOU from UWFWS (Ellen Lance, Anchorage-AK). We are grateful to Anna Santure and 3 anonymous reviewers for helpful comments on earlier drafts of this manuscript. This work was funded by University of New Mexico and the Mammalian Ecology and Conservation Unit (VGL) at UC Davis.

References

Abbott CL, Double MC, Trueman JWH, Robinson A, Cockburn A (2005) An unusual source of apparent mitochondrial heteroplasmy: duplicate mitochondrial control regions in *Thalassarche* albatrosses. *Molecular Ecology*, **14**, 3605–3613.

Avise JC, Arnold J, Ball RM *et al.* (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, **18**, 489–522.

Barker FK (2014) Mitogenomic data resolve basal relationships among passeriform and passeridan birds. *Molecular Phylogenetics and Evolution*, **79**, 313–324.

Butler JM (2005) *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*, 2nd edn. Elsevier, London.

Chen C, Khaleel SS, Huang H, Wu CH (2013) ngsShoRT: A Software for Pre-processing Illumina Short Read Sequences for De Novo Genome Assembly. *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, ACM.

Cooke GM, King AG, Johnson RN, Boles WE, Major RE (2012) Rapid characterization of mitochondrial genome rearrangements in Australian songbirds using next-generation sequencing technology. *Journal of Heredity*, **103**, 882–886.

Eda M, Baba Y, Koike H, Higuchi H (2006) Do temporal size differences influence species identification of archaeological albatross remains when using modern reference samples?. *Journal of Archaeological Science*, **33**, 349–359.

Eda M, Kuro-O M, Higuchi H, Hasegawa H, Koike H (2010) Mosaic gene conversion after a tandem duplication of mtDNA sequence in Diomedeidae (albatrosses). *Genes Genetic Systems*, **85**, 129–139.

Enk JM, Devault AM, Kuch M, Murgha YE, Rouillard J, Poinar HN (2014) Ancient whole genome enrichment using baits built from modern DNA. *Molecular Biology and Evolution*, **31**, 1292–1294.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, **41**, 95–98.

Hancock-Hanser BL, Frey A, Leslie MS, Dutton PH, Archer FI, Morin PA (2013) Targeted multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Molecular Ecology Resources*, **13**, 254–268.

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. *Genome Research*, **9**, 868–877.

Ingman M, Gyllensten U (2007) Rate variation between mitochondrial domains and adaptive evolution in humans. *Human Molecular Genetics*, **16**, 2281–2287.

King JL, LaRue BL, Novroski NM *et al.* (2014) High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. *Forensic Science International: Genetics*, **12**, 128–135.

Knaus BJ, Cronn R, Liston A, Pilgrim K, Schwartz MK (2011) Mitochondrial genome sequences illuminate maternal lineages of conservation concern in a rare carnivore. *BMC Ecology*, **11**, 10.

Knief C (2014) Analysis of plant microbe interactions in the era of next generation sequencing technologies. *Frontiers in Plant Science*, **5**, 216.

Kuro-O M, Yonekawa H, Saito S *et al.* (2010) Unexpectedly high genetic diversity of mtDNA control region through severe bottleneck in vulnerable albatross *Phoebastria albatrus*. *Conservation Genetics*, **11**, 127–137.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li R, Zhu H, Ruan J *et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, **20**, 265–272.

Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJP (2013) Capturing protein-coding genes across highly divergent species. *BioTechniques*, **54**, 321–326.

Librado P, Rozas J (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS ONE*, **7**, e30087.

Magoc T, Salzberg S (2011) FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–2963.

Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, **6**, S13–S20.

Milne I, Stephen G, Bayer M *et al.* (2013) Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*, **14**, 193–202.

Mindell DP, Sorenson MD, Dimcheff DE (1998) An extra nucleotide is not translated in mitochondrial ND3 of some birds and turtles. *Molecular Biology and Evolution*, **15**, 1568–1571.

Mitchell KJ, Wood JR, Scofield RP, Llamas B, Cooper A (2014) Ancient mitochondrial genome reveals unsuspected taxonomic affinity of the extinct Chatham duck (*Pachyanas chathamica*) and resolves divergence times for New Zealand and sub-Antarctic brown teals. *Molecular Phylogenetics and Evolution*, **70**, 420–428.

Morin PA, Archer FI, Foote AD *et al.* (2010) Complete mitochondrial genome phylogeographic analysis of killer whales (*Orcinus orca*) indicates multiple species. *Genome Research*, **20**, 908–916.

Morris-Pocock JA, Taylor SA, Bir TP, Friesen VL (2010) Concerted evolution of duplicated mitochondrial control regions in three related seabird species. *BMC Evolutionary Biology*, **10**, 14.

Peterlongo P, Chikhi R (2012) Mapsembler, targeted and micro assembly of large NGS datasets on a desktop computer. *BMC Bioinformatics*, **13**, 48.

Pfeifer B, Wittelsbürger U, Onsins SER, Lercher MJ (2014) PopGenome: an efficient swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, **31**, 1929–1936.

R Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shao R, Barker SC, Mitani H, Aoki Y, Fukunaga M (2005) Evolution of duplicate control regions in the mitochondrial genomes of metazoa: a case study with Australasian Ixodes ticks. *Molecular Biology and Evolution*, **22.3**, 620–629.

Slack KE, Jones CM, Ando T *et al.* (2006) Early penguin fossils, plus mitochondrial genomes, calibrate avian evolution. *Molecular Biology and Evolution*, **23**, 1144–1155.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution*, **30**, 2725–2729.

Watanabe M, Nikaido M, Tsuda TT *et al.* (2006) New candidate species most closely related to penguins. *Gene*, **378**, 65–73.

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, **20**, 3252–3255.

Z.T.L. performed all data analyses including sequence data processing, bioinformatic analyses and mtDNA comparative analyses. B.N.S. assisted in study design, assembly workflow, data interpretation and writing. Z.T.L., B.N.S. and S.K.B. performed the laboratory procedures including DNA extraction and library preparation. P.C., R.H. and S.N. contributed to fieldwork including sample collection and processing. All authors contributed to writing the manuscript.

Data Accessibility

Annotated mitochondrial DNA sequences have been submitted to GenBank (Accession nos. KJ735512–KJ735514, KM878669–KM878670). Assembly Workflow Scripts are available as Online Supporting information, Appendix S1.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Annotated command line code for de novo assembly workflow.

Table S1 Full annotation of the mitochondrial genome of all three North Pacific albatross species: black-footed (BFAL), Laysan (LAAL), and short-tailed (STAL) albatrosses.

Table S2 Pairwise comparisons of *Phoebastria* mitogenomes.