Short Reads, Circular Genome: Skimming SOLiD Sequence to Construct the Bighorn Sheep Mitochondrial Genome

JOSHUA M. MILLER, RENÉ M. MALENFANT, STEPHEN S. MOORE, AND DAVID W. COLTMAN

From the Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada (Miller, Malenfant, Moore, and Coltman); and the Department of Agricultural, Food, and Nutritional Science, University of Alberta, Edmonton, Alberta T6G 2P5, Canada (Moore).

Address correspondence to Joshua M. Miller at the address above, or e-mail: jmm1@ualberta.ca.

As sequencing technology improves, an increasing number of projects aim to generate full genome sequence, even for nonmodel taxa. These projects may be feasibly conducted at lower read depths if the alignment can be aided by previously developed genomic resources from a closely related species. We investigated the feasibility of constructing a complete mitochondrial (mt) genome without preamplification or other targeting of the sequence. Here we present a full mt genome sequence (16,463 nucleotides) for the bighorn sheep (Ovis canadensis) generated though alignment of SOLiD short-read sequences to a reference genome. Average read depth was 1240, and each base was covered by at least 36 reads. We then conducted a phylogenomic analysis with 27 other bovid mitogenomes, which placed bighorn sheep firmly in the Ovis clade. These results show that it is possible to generate a complete mitogenome by skimming a low-coverage genomic sequencing library. This technique will become increasingly applicable as the number of taxa with some level of genome sequence rises.

Key words: ABI SOLiD, Caprini, mtDNA, phylogenomics, short reads, Ovis

Whole-genome sequencing is becoming routine with recent advances in sequencing technology. Human genome resequencing is now common and brings the promise of personalized medicine (Chee-Seng et al. 2001), and several model and domestic organisms have full genome sequences as well (Adams et al. 2000; Dalloul et al. 2010). This expansion has also opened new avenues of research for nonmodel organisms (Ouborg et al. 2010; Tautz et al. 2010; Ekblom and Galindo 2011), encouraging full-genome sequencing of wild taxa (Jackson et al. 2006; Haussler et al. 2009). These efforts are aided by the fact that many species are considered “genome enabled” (Kohn et al. 2006) due to availability of genomic resources in a closely related species. Such resources allow researchers to sequence new taxa at shallower read depths and then align reads to a reference, rather than perform de novo assembly.

The benefits of various sequencing technologies for genome sequencing have been reviewed extensively elsewhere (Morozova et al. 2009; Ekblom and Galindo 2011). For most researchers, the choice comes to balancing the number of base pairs (bp) sequenced in a single read (read length), the number of times an individual base is sequenced (read depth), and accuracy of the sequence generated (Ekblom and Galindo 2011). The ABI SOLiD platform uses short reads (35–75 bp) and is useful for resequencing studies because once reads are aligned to a reference genome, sequencing errors can be distinguished from true nucleotide variations by the nature of its sequence-by-ligation strategy (Ondov et al. 2008; McKernan et al. 2009).

Mitochondrial (mt) DNA has a long history of use for both phylogenetics and population genetic analyses (Avise et al. 1987; Moritz et al. 1987; Moritz 1994). mtDNA is an ideal marker due to its fast mutation rate relative to nuclear genome, ease of sequencing from conserved primers, and lack of recombination (Kocher et al. 1989; Moore 1995; Birky 2001; though see White et al. 2008; Galtier et al. 2009 for some caveats to these generalities). These features have led mtDNA, specifically the cytochrome c oxidase I (COI) gene, to be proposed for use as the “barcode of life” (Hebert et al. 2003).

As sequencing technology improves, the use of entire mt genomes rather than specific genes becomes increasingly feasible. Full mt genomes provide more complete and robust datasets compared with studies that use only 1 or 2 regions of the genome (DeFilippis and Moore 2000; Ingman et al. 2000; Inoue et al. 2001; Hassanin et al. 2009; Meadows et al. 2011). Full mt genomes have often been obtained
through targeted amplification of mtDNA followed by sequencing with next-generation long-read sequencing technologies, such as 454 Life Sciences (Roche). This approach has been applied to a variety of taxa including beavers (Castor spp.; Horn et al. 2011), killer whales (Orcinus Orca; Morin et al. 2010), the yellow croaker (Larimichthys crocea, Cui et al. 2009), several parasitic worm species (Jex et al. 2010), as well as ancient DNA samples (Briggs et al. 2009; Lari et al. 2011).

However, with the proliferation of genomic sequence data from related species, we wanted to explore the possibility of constructing a complete mt genome for an uncharacterized wildlife taxon without preamplification or other targeted enrichment. A similar effort was undertaken by Nabholz et al. (2010) who mined a transcriptomic data set for mt sequence in order to gain insight into the phylogeny of passerine birds. Although this method was successful in terms of generating a robust data set, it is limited to only yielding sequence for the protein-coding genes in an mt genome.

Here for the first time, we present a complete mt genome generated from skimming a short-read genomic sequencing library. As part of a larger project, a genomic sequencing library for the bighorn sheep (Ovis canadensis) was generated using an ABI SOLiD 4 platform. From this, we constructed an mt genome utilizing resources for the closely related domestic sheep (Ovis aries, ca. 3 million years divergence, Bunch et al. 2006) as a reference for alignment and annotation. Lastly, we conducted a phylogenomic analysis, placing the sequence in a larger phylogeny of 27 other bovid taxa.

**Methods**

**DNA Extraction and Next-Generation Sequencing**

Total genomic DNA was extracted from tissue of a single bighorn sheep from Ram Mountain (Alberta, Canada), using standard phenol–chloroform extraction protocols (Sambrook and Russell 2001). Mate-paired libraries were then prepared using the reagents and protocols provided by Applied Biosystems (Life Technologies Corporation, Carlsbad, CA). Specifically, 50 micrograms of DNA was sheared into 1.5-kb fragments using a HydroShear (DigiLab Genomic Solutions Inc., Holliston, MA). The fragmented DNA was end repaired and ligated to LMP CAP adapters, size selected by electrophoresis on a 1% agarose-LE gel, and then purified using the SOLiD Library quick gel extraction kit. Next, the sheared DNA was circularized with a biotinylated internal adaptor and then purified using the SOLiD Library micro column purification kit. Purified circularized DNA was nick translated, digested with T7 exonuclease and S1 nuclease, and end repaired before binding the library molecules to streptavidin beads. The double-stranded P1 and P2 sequencing adapters were ligated to the end-repaired DNA prior to amplification of the library. The amplified library was purified using the SOLiD Library micro column purification kit.

Average molecule sizes of the library were confirmed by analysis with a DNA 1000 chip using the Bioanalyzer (Agilent, Mississauga, ON, Canada), and the final concentration of the library was quantitated by the SOLiD Library TaqMan Quantitation Kit using the StepOne Plus real-time PCR (Life Technologies Corporation). Emulsion PCR was performed using the SOLiD EZ bead system (Life Technologies Corporation). Both forward and reverse tags were sequenced to 50 bases using an Applied Biosystems SOLiD System (V4 chemistry) that involves ligation-based sequencing and a 2-base encoding method (McKernan et al. 2009) using a massively parallel sequencing platform SOLiD 4 sequencer (Life Technologies Corporation).

**Sequence Alignment and Phylogenetic Analysis**

Raw reads and associated quality scores were imported into CLC Genomics Workbench 4.5.1 (CLC bio, Cambridge, MA, USA). Direct alignment to domestic sheep mt genome (GenBank accession NC_001941) was conducted using default parameters for local alignment of color-space sequences. Specifically, we set a mismatch cost of 2, an insertion cost of 3, and a deletion cost of 3. Length fraction was set at 0.5 while the minimum identity was 0.8, meaning that at least 50% of a sequence read had to have at least 80% similarity to the reference sequence to be mapped. We included the mate-paired information by specifying an expected distance range of 100–200 bp between sequence pairs.

In members of the genus Ovis, there is approximately 75-bp tandem repeat at the 5’ end of the control region (Hiendleder et al. 2002), where bighorn sheep carry 2 repeat units whereas domestic sheep most often carry 4. The 2 excess repeats were removed from the aligned sequence after comparison to a reference for the control region from a Canadian Rocky Mountain bighorn sheep (AY091486, Hiendleder et al. 2002). Annotations were added to the bighorn sheep mt genome, using the annotation tool in CLC Genomics Workbench, based on similarity with the domestic sheep reference after protein-coding genes were checked for premature stop codons.

Once aligned, variations between the bighorn and domestic sheep mt genomes were assessed with the sequence analysis tools in CLC Genomics Workbench. For SNP detection, we set a window length of 11 bp, the maximum number of gaps and mismatches at 2, and the minimum average quality of the surrounding bases at 15. These parameters set the thresholds for the Neighborhood Quality Standard algorithm (Altschuler et al. 2000) used by CLC. Putative SNPs were ignored if coverage was less than 4 reads or if the minor allele frequency of the variant was less than 35%. Similarly, insertion/deletion polymorphisms were called only if read depth was greater than 4 and over 35% of the reads showed the variant. After removal of the 5’ tandem repeats, the sequences for the bighorn sheep and domestic mitogenomes were realigned in CLC, and the number of gaps and nucleotide variants was reassessed.

For phylogenetic analysis, 26 additional bovid mt genomes were accessed from GenBank (Table 1). Sequences were aligned with MAFFT 6.8.5.0 (Katoh and Toh 2008) online (default algorithm, offset = 0), as well as MUSCLE 3.8.31 (Edgar 2004a, 2004b). Alignments were
then compared using AltAVisT 1.0 (Morgenstern et al. 2003) to determine regions that aligned ambiguously. This highlighted the 5' tandem repeats in the control region, which were then excluded from all subsequent analyses.

jModelTest 0.1.1 (Posada 2008) was used to identify the nucleotide substitution model that best fit the data set (GTR + I + G). This model was applied in GARLI 1.0 (Zwickl 2006) where 2 replicates were run with 100 bootstraps per replicate. Bootstrap results were mapped onto the original GARLI replicate tree using SumTrees 3.1.0, part of the DendroPy library version 3.7.1 (Sukumaran and Holder 2010). In addition, the data were examined in a Bayesian framework using MrBayes 3.2.1 (Huelsenbeck et al. 2001, 2002). This analysis was run with default priors for 1,000,000 generations using 8 chains. A 25% burn-in was applied after assessment of average standard deviation of split frequencies and log likelihood values. Consensus trees were visualized and edited in Mesquite 2.73 (Maddison WP and Maddison DR 2010). Clades were considered strongly supported if they had bootstrap values greater than 70, or posterior probabilities greater than 0.95.

**Results**

Genomic sequencing resulted in 312 million short reads. Of these, approximately 470,000 aligned to the reference mt genome, with an average read depth of 1240 (range 36–8630). Areas of high and low coverage are dispersed throughout the genome and appear to be associated with GC content such that average coverage drops when GC content rises above 40% (Supplementary Figure S1). The average distance between paired reads was 1055 nucleotides (nt)(Supplementary Figure S2). After removal of the excess tandem repeats, the final length of the bighorn sheep mt genome was 16,463 nt (GenBank accession number JN181255). Only 37 heterozygous sites were found after all reads had been aligned to the reference genome. These ambiguities were not retained in the final sequence, but rather replaced with the most common allele.

Comparison to the domestic sheep mt genome revealed 793 nt differences (Supplementary Table 1). In addition, we found 175 gaps between the 2 sequences. Relative to the domestic sheep mitogenome, there were 164 deletions (all but 4 of which are in the control region, and 150 of which were associated with the 5' tandem repeat) and 150 of which were in the control region). The order of features on the bighorn sheep mt genome was the same as in domestic sheep, and no premature stop codons were found after annotation of protein-coding genes. To get a sense of the distribution of these substitutions, we looked at the number of substitutions relative to the number of nucleotides for the components of the mt genome (Table 2). This analysis showed that the control region had the highest proportion of substitution, approximately 23%.

Phylogenomic analyses of the 28 mt genomes were based on 16,128 characters, after 883 nt were excluded due to ambiguous alignment. Of these, 4527 were phylogenetically informative.

---

**Table 1** Species name and accession numbers for mitogenomes used in this study

<table>
<thead>
<tr>
<th>Species</th>
<th>Common name</th>
<th>Accession no.</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aoudad</td>
<td>Aoudad</td>
<td>FJ207522</td>
<td>Hassanin et al. (2009)</td>
</tr>
<tr>
<td>Arabian tahr</td>
<td>Arabian tahr</td>
<td>FJ207523</td>
<td>Hassanin et al. (2009)</td>
</tr>
<tr>
<td>Cattle</td>
<td>Cattle</td>
<td>AV526085</td>
<td>HY Chung and JM Ha, unpublished data</td>
</tr>
<tr>
<td>Swamp buffalo</td>
<td>Swamp buffalo</td>
<td>NC_006295</td>
<td>JX Qiang et al., unpublished data</td>
</tr>
<tr>
<td>Takin</td>
<td>Takin</td>
<td>FJ207524</td>
<td>Hassanin et al. (2009)</td>
</tr>
<tr>
<td>Markhor</td>
<td>Markhor</td>
<td>FJ207525</td>
<td>Hassanin et al. (2009)</td>
</tr>
<tr>
<td>Domestic goat</td>
<td>Domestic goat</td>
<td>NC_005044</td>
<td>Hassanin et al. (2010)</td>
</tr>
<tr>
<td>Alpine ibex</td>
<td>Alpine ibex</td>
<td>FJ207526</td>
<td>Hassanin et al. (2009)</td>
</tr>
<tr>
<td>Nubian ibex</td>
<td>Nubian ibex</td>
<td>FJ207527</td>
<td>Hassanin et al. (2009)</td>
</tr>
<tr>
<td>Spanish ibex</td>
<td>Spanish ibex</td>
<td>FJ207528</td>
<td>Hassanin et al. (2009)</td>
</tr>
<tr>
<td>Siberian ibex</td>
<td>Siberian ibex</td>
<td>FJ207529</td>
<td>Hassanin et al. (2009)</td>
</tr>
<tr>
<td>Sumatran serow</td>
<td>Sumatran serow</td>
<td>FJ207534</td>
<td>Hassanin et al. (2009)</td>
</tr>
<tr>
<td>Biesbok</td>
<td>Biesbok</td>
<td>FJ207530</td>
<td>Hassanin et al. (2009)</td>
</tr>
<tr>
<td>Himalayan tahr</td>
<td>Himalayan tahr</td>
<td>FJ207531</td>
<td>Hassanin et al. (2009)</td>
</tr>
<tr>
<td>Chinese muntjac</td>
<td>Chinese muntjac</td>
<td>NC_008491</td>
<td>HW Wang et al., unpublished data</td>
</tr>
<tr>
<td>Chinese goral</td>
<td>Chinese goral</td>
<td>FJ207532</td>
<td>Hassanin et al. (2009)</td>
</tr>
<tr>
<td>Mountain goat</td>
<td>Mountain goat</td>
<td>FJ207535</td>
<td>Hassanin et al. (2009)</td>
</tr>
<tr>
<td>Muskox</td>
<td>Muskox</td>
<td>FJ207536</td>
<td>Hassanin et al. (2009)</td>
</tr>
<tr>
<td>Domestic sheep</td>
<td>Domestic sheep</td>
<td>NC_001941</td>
<td>Hiendleder et al. (1998)</td>
</tr>
<tr>
<td>Bighorn sheep</td>
<td>Bighorn sheep</td>
<td>JN181255</td>
<td>This study</td>
</tr>
<tr>
<td>Mouflon</td>
<td>Mouflon</td>
<td>HM236185</td>
<td>Meadows et al. (2011)</td>
</tr>
<tr>
<td>Urial</td>
<td>Urial</td>
<td>HM236187</td>
<td>Meadows et al. (2011)</td>
</tr>
<tr>
<td>Chiru</td>
<td>Chiru</td>
<td>NC_007441</td>
<td>Xu et al. (2005)</td>
</tr>
<tr>
<td>Isard</td>
<td>Isard</td>
<td>FJ207537</td>
<td>Hassanin et al. (2009)</td>
</tr>
<tr>
<td>Chamois</td>
<td>Chamois</td>
<td>FJ207539</td>
<td>Hassanin et al. (2009)</td>
</tr>
</tbody>
</table>
Table 2  Number of nucleotide differences between components of the bighorn and the domestic sheep mt genomes

<table>
<thead>
<tr>
<th>Component</th>
<th>Number of substitutions</th>
<th>Total length (nt)</th>
<th>Proportion of substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP6</td>
<td>42</td>
<td>679</td>
<td>0.062</td>
</tr>
<tr>
<td>ATP8</td>
<td>9</td>
<td>201</td>
<td>0.045</td>
</tr>
<tr>
<td>CO1</td>
<td>79</td>
<td>1545</td>
<td>0.051</td>
</tr>
<tr>
<td>CO2</td>
<td>32</td>
<td>684</td>
<td>0.047</td>
</tr>
<tr>
<td>CO3</td>
<td>34</td>
<td>766</td>
<td>0.044</td>
</tr>
<tr>
<td>Control region</td>
<td>269</td>
<td>1194</td>
<td>0.225</td>
</tr>
<tr>
<td>cytB</td>
<td>66</td>
<td>1140</td>
<td>0.058</td>
</tr>
<tr>
<td>ND1</td>
<td>43</td>
<td>955</td>
<td>0.045</td>
</tr>
<tr>
<td>ND2</td>
<td>56</td>
<td>1042</td>
<td>0.054</td>
</tr>
<tr>
<td>ND3</td>
<td>21</td>
<td>346</td>
<td>0.061</td>
</tr>
<tr>
<td>ND4</td>
<td>92</td>
<td>1478</td>
<td>0.062</td>
</tr>
<tr>
<td>ND4L</td>
<td>7</td>
<td>197</td>
<td>0.036</td>
</tr>
<tr>
<td>ND5</td>
<td>117</td>
<td>1821</td>
<td>0.064</td>
</tr>
<tr>
<td>ND6</td>
<td>27</td>
<td>528</td>
<td>0.051</td>
</tr>
<tr>
<td>rRNA</td>
<td>57</td>
<td>2530</td>
<td>0.023</td>
</tr>
<tr>
<td>tRNA</td>
<td>21</td>
<td>1510</td>
<td>0.014</td>
</tr>
</tbody>
</table>

Bighorn sheep was found to be firmly placed in a monophyletic group of all *Ovis* species (Figure 1). The majority of nodes were supported both by high posterior probability values and high bootstrap support (Figure 1). An exception to this trend was the sister relationship between the *Ammotragus–Rupicapra* clade and the *Ovis–Oreamnos* clade that received 47% bootstrap support but 100% support in the Bayesian analysis. Similarly, the placement of *Oreamnos* received 67% bootstrap support, but a 100% posterior probability.

Discussion

In this study, we presented the mt genome for bighorn sheep constructed from a short-read genomic sequencing library without preamplification or other targeting of the region. Alignment using the domestic sheep mt genome as a reference resulted in a sequence that had complete coverage with deep read depth. Phylogenomic analyses placed the bighorn sheep within the *Ovis* clade but separate from Eurasian sheep. This topology had high support in both Bayesian and a maximum likelihood analyses and matches with previous phylogenies of the genus based on both partial mt and nuclear DNA (Bunch et al. 2006; Rezaei et al. 2009).

**Figure 1.** Phylogenetic tree constructed from 28 full mt genomes representing the tribe Caprini. Support values are shown for each node (posterior probability/bootstrap value). Branch lengths are proportional to number of substitutions per site.
Though the phylogenetic placement of bighorn sheep within Ovis has not been controversial, other relationships within this genus (Hiendaelder et al. 2002; Bunch et al. 2006; Meadows et al. 2007; Rezaei et al. 2009) as well as those in the larger Caprini tribe (Ropiquet and Hassanin 2005; Hassanin et al. 2009; Shafer and Hall 2010) have been subject to much debate. In this analysis, we found evidence for the sister relationship between Ovis ammon and Ovis vignei, a topology that had not been previously suggested. However, both the bootstrap support and posterior probability for this relationship are lower than most of the other relationships seen in the tree. With respect to the tribe Caprini, inclusion of the bighorn sheep mt genome resulted in a sister relationship between the Ammotragus–Rupicapra clade and the Ovis–Oreamnos clade (Figure 1). Previous analyses bases on mt genomes found the Ammotragus–Rupicapra clade as sister to the Capra clade, but this had low posterior probability support (Hassanin et al. 2009). In addition, Oreamnos has been found to be a “rogue taxon” making its phylogenetic placement extremely difficult (Shafer and Hall 2010). Whereas our topology does have high Bayesian support, bootstrap support is low. Expanded taxon sampling may help to firmly place some of these controversial nodes (Zwickl and Hillis 2002; Heath et al. 2008). For instance, mt genomes for thinhorn sheep (Ovis dalli), Asian mouflon (Ovis orientalis), and snow sheep (Ovis nivicola) would likely help resolve the controversies within Ovis, whereas inclusion of sequence for the Nilgiri tahr (Nilgiritragus hylocrius) could help elucidate relationships in the tribe Caprini.

A concern when conducting any analysis based on mt DNA is the coamplification of nuclear inserts of mt genes (numts) that can lead to incorrect phylogenetic estimates (Sorenson and Quinn 1998; Richly and Leister 2004). It may seem that this risk is compounded when the mtDNA sequence is skinned directly from a genomic library. However, although amplification of numts likely occurred during preparation of the genomic sequencing library, we do not feel that it had an influence on the assembly. Numts would be at a significantly lower copy number than true mtDNA, so incorporation into the alignment was likely rare. However, in cases where alignment did occur and there was variation between the mtDNA and numt, such variants would likely not pass the stringency criteria for SNP calling.

Here we have shown that skimming lower coverage genomic sequencing libraries is a valid option to generate complete mt genomes. This eliminates the need to preamplify the sequence using multiple overlapping primer sets or long-range PCR. Similar efforts have been successfully conducted using transcriptomic libraries (Naeholz et al. 2010). However, starting from genomic sequence allows inclusion of more elements of the mitochondria, including the control region, which in some cases is the most information rich (Meadows et al. 2011). We expect this technique to become increasingly relevant, especially as the cost of next-generation sequencing continues to fall and more taxa are genome-sequenced at some level. As we move to total genomic sequences and phyllogenomic analyses, it is important not to forget what has been one of the gold standards of molecular biology.

**Supplementary Material**

Supplementary material can be found at http://www.jhered.oxfordjournals.org/.

**Funding**

Natural Sciences and Engineering Research Council of Canada grant (312207) to DWC; Graduate Research Scholarship from Alberta Innovates—Technology Futures and the University of Alberta to JMM.

**Acknowledgments**

The library preparation and sequencing were performed in the Agricultural Genomics and Proteomics Unit at the University of Alberta. We would like to thank the University of Alberta’s Biology 521 class for instruction on phylogenetic analyses and Aaron B. A. Shafer as well as 3 anonymous reviewers for helpful comments on earlier drafts.

**References**


Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reed CA, Saunders NC. 1987. Intraspecific phylogeography: the mitochondrial DNA


Zwickl DJ. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Austin (TX): The University of Texas.


Received April 27, 2011; Revised August 31, 2011; Accepted August 31, 2011

Corresponding Editor: Warren Johnson