Rapid mitochondrial genome sequencing based on Oxford Nanopore Sequencing and a proxy for Vertebrate Species Identification using MinION device

Short title: MinION sequencing for mammal species identification

Nicolás D. Franco-Sierra, Juan F. Díaz-Nieto*.

Grupo de investigación en Biodiversidad, Evolución y Conservación (BEC), Departamento de Ciencias Biológicas, Escuela de Ciencias, Universidad EAFIT. Carrera 49 N° 7 Sur-50, Medellín, Colombia.

*corresponding address: jdiazni@eafit.edu.co

Abstract

Molecular information is crucial for species identification when facing challenging morphology-based specimen identification. The use of DNA barcodes partially solves this problem, but in some cases when PCR is not an option (i.e. primers are not available, problems in reaction standardization) amplification-free approaches could lead to better and rapid results. Recent advances in DNA sequencing, like MinION device from Oxford Nanopore Technologies (ONT), allow to obtain genomic data with low laboratory and technical requirements, and at a relative low cost. In this study, we explore ONT sequencing for molecular species identification from a total DNA sample obtained from a neotropical rodent and also we test the technology for complete mitochondrial genome reconstruction via genome skimming. We were able to obtain "*de novo"* the complete mitogenome of *Melanomys caliginosus* (Cricetidae: Sigmodontinae) with an average depth coverage of 77.5X using ONT-only data and by combining multiple assembly routines to produce a coherent assembly (complete ORFs and RNAs). Our routine for an automated species identification was able to identify the sample with high certainty using unassembled sequence data in a reasonable computing time (1:56:31) and it was significatively reduced when *a priori* information related to the organism identity was known (00:04:57 and 00:01:47 when order and family were known, respectively). Our findings suggest ONT sequencing as a suitable candidate to solve species identification problems in metazoan non-model organisms and generate complete mtDNA datasets.

Keywords: Oxford Nanopore Sequencing, Vertebrate species identification, mitochondrial DNA sequencing, molecular species identification, genome skimming.

1. Introduction

Species are the fundamental study unit in biology, therefore accurate species identification is a vital process in the biological and medical sciences (Mayr, 1982; de Queiroz, 2005). DNA sequencing have been increasingly used in recent years for species identification given the challenges associated with morphological identification (e.g., need of adult individuals, specimens preserved in optimal conditions, inability for recognition of cryptic species, among many others) (Avise, 1994; Crawford *et al.*, 2013; Mendoza *et al.*, 2016). Yet, traditional DNA-based identification requires processing samples in a laboratory,using DNA sequencing techniques which traditionally involve robust equipment, methods that preclude obtaining sequences of non-model organisms (e.g., specific primers for PCR sequencing), and high-cost reagents that are not available at most laboratories (Wetterstrand, 2018; Erlich, 2015; Sboner *et al.*, 2011).

Recent developments in DNA sequencing techniques, such as single molecule sequencing detection like the nanopore-based method from Oxford Nanopore Technologies (ONT), are a breakthrough in molecular biology due to its multiple advantages such as long sequencing reads, portability (pocket-sized device called MinION), significantly reduced cost, and relative simplicity for its setup and operation compared with the traditional sequencing platforms (Jain *et al.*, 2016; Laver *et al.* 2015). Given those benefits, a new range of applications can be explored in fields like

microbiology, human genetics, basic genome research, microbiome studies, clinical and animal research (Schmidt *et al.*, 2017; Norris *et al.*, 2016). Caveats in this sequencing method are mostly associated to its high error rate (~10%) which is primarily attributed to the basecalling process (Magi *et al.*, 2016). Consequently, computational advances on basecalling from raw signal are still required to obtain higher sequence read quality compared to short read technologies, some misreadings can be captured from modified nucleotides, homopolymer sequences, and in consequence indels can be introduced (Magi *et al.*, 2016). Despite of this, use of the technology has been demonstrated with outstanding success in genomic surveillance of Ebola and Zika viruses outbreaks (Quick *et al.*, 2016; Faria *et al.*, 2016; Quick *et al.*, 2017), improved reconstruction in genomes of well-studied model organisms like Gram-negative bacterium *Escherichia coli* (Quick *et al.*, 2014), yeast *Saccharomyces cerevisiae* (Salazar *et al.*, 2017), nematode *Caenorhabditis elegans* (Tyson *et al.*, 2017) and even the resolution of complicated genomic regions (telomeres, centromeres, HLA locus, Y chromosome) of the human genome (Jain *et al.*, 2018a; Jain *et al.*, 2018b).

In the field of DNA-based species identification, ONT sequencing has been widely used in bacterial and viral applications, especially for clinical research, environmental, and microbiome studies (Greninger *et al.*, 2015; Benítez-Páez *et al.*, 2016). As a consequence, most computational pipelines and efforts have been developed for those taxa (model organisms), in tools like WIMP (What's in my pot?) (Juul *et al.*, 2015), and Kraken (Wood & Salzberg, 2014), among others.

Regarding molecular identification of metazoan organisms, advances have been done using barcodes in studies performed for identification of amphibians and reptiles in the Ecuadorian rainforest (Pomerantz *et al.*, 2018), and cost-effective method benchmarked with samples of hundreds of dipteran and hymenopteran insects (Srivathsan *et al.*, 2018). However, these approaches rely on the use of DNA barcoding and therefore they do require amplification by PCR, implying additional laboratory steps and reagents in order to obtain the desired sequence. In addition, when working with non-model organisms or with less-studied biotic groups at the molecular level, adequate primers are not available to achieve a successful amplification reaction (Schäffer et al., 2017).

DNA barcodes have been proved as a cost-effective, accessible and reliable solution to the species identification problem (Hebert *et al.*, 2003a). Since its proposal, many data of DNA barcodes have been obtained and released, especially for mitochondrial *COI* barcode (cytochrome oxidase subunit I) in metazoans resolving a wide range of questions (taxonomic, evolutionary, identification) in many insect, avian and other animal clades (Hebert *et al.*, 2003b; Stoeckle, 2003; Kerr *et al.* 2009; Blagoev *et al.*, 2016). Despite of that, some taxonomic groups would benefit from high quality reference sequence sets at genomic and mitogenomic scales (Lessa *et al.*, 2014), specially to solve discrepancies observed in single gene-based analysis (i.e. effectiveness of species trees over gene trees in molecular systematics,). Highly diverse taxa usually lack sufficient molecular information to properly address most of these questions. A particular example of such diverse group can be found in the sigmodontine rodents (e.g. cotton and rice rats, grass mice), a subfamily of the

Neotropical family Cricetidae consisting of 86 extant genera and about 400 species, of which 85 genera and 381 inhabit South America, making them an ubiquitous and widely distributed group (D'Elía & Pardiñas, 2016). To date, although a notable proportion of species members of the subfamily have available DNA sequences, these are mostly associated to the mitochondrial cytochrome oxidase b gene (*CYTB)* and only 3 species (less than 1% of the sigmodontine diversity) have a their complete mitochondrial genomes sequenced (3 records on GenBank accessed 1st October 2018), an dramatically low number for the genomic era (Coissac *et al.*, 2016). A more extensive taxon sampling is required to address phylogenomic approaches in this group (Lessa *et al.*, 2014).

The cost-benefit and accessibility of Oxford Nanopore Technology (ONT) allows to obtain genome-wide and organellar DNA sequences *in situ* without the need (and avoid the associated problems) of PCR-based amplification. Previous efforts have been made to obtain mitochondrial genomes using MinION-ONT sequencing data but to overcome the read noisiness and improve assembly quality, short-read data from Illumina platform were also implemented (Torres et al. 2018; Chandler et al. 2017). Herein, we propose that with enough mitochondrial DNA coverage, the mitogenome could be accurately reconstructed using ONT-only data combining several computational assembly routines to solve known issues and error-prone sites in the sequence. Although noisiness will be still present in sequence reads, this information is valuable as a rapid alternative for species identification purposes. We explore MinION-ONT sequencing use as a promising tool for mitogenomics and molecular

species identification of non-model organisms with a PCR-free approach. As a case of study, we present the first complete mitochondrial genome assembled using ONT-only data of the South American sigmodontine rodent *Melanomys caligno*sus and evaluate an assembly-free method for its molecular identification.

2. Materials and Methods

2.1 DNA extraction

The aim of this work is to demonstrate that ONT sequencing can be used as a tool for molecular species identification of non-model organisms. To achieve this, DNA sample was obtained from a wild-caught specimen of a sigmodontine rodent.

Fieldwork was conducted at a rural locality known as “Loma del Chocho” (6.163576 N, 75.548467 W; 2200m a.s.l.), municipality of Envigado, Antioquia, Colombia. A small rodent captured using Sherman traps was identified as *Melanomys caliginosus* according its morphological characters and it was deposited at the specimen collection of Universidad EAFIT (Medellín, Antioquia, Colombia.) coded as voucher JFD1322. The whole liver of the animal was extracted and a fraction was used fresh for DNA extraction. The remaining liver was stored frozen at -80°C.

Two different methods were used for whole DNA extraction: (1) GenElute™ Mammalian Genomic DNA Miniprep Kit Protocol (Sigma-Aldrich) following manufacturer’s instructions and (2) a DNA extraction protocol designed to obtain High Molecular Weight DNA (Quick, 2018). Both protocols used 25 mg of fresh liver as start material. For the High Molecular Weight DNA (HMW) protocol the following

procedure was used: the liver tissue (25 mg) was resuspended by gentle pipette mixing in 200 µl PBS (buffer phosphate saline 1X). 10 ml TLB was added (100 mM NaCl, 10 mM Tris-Cl pH 8.0, 25 mM EDTA pH 8.0, 0.5% (w/v) SDS, 20 µg/ml Sigma RNase A solution), vortexed at full speed for 5 s and incubated at 37 °C for 1 h. 50 µl Proteinase K (Sigma) was added and mixed by slow end-over-end rotation three times followed by 2 h at 50 °C with gentle mixing every 30 minutes. The lysate was phenol-purified by adding 10 ml TE saturated phenol on phase-lock gel falcon tubes (Dow Corning vacuum grease). Tubes were incubated at room temperature and rotation at 20 rpm for 10 minutes to achieve an homogeneous emulsion. The emulsion was centrifuged at 4500 rpm for 10 minutes in order to achieve complete phase separation. Aqueous phase was retained and an additional phenol:chloroform (1:1) purification step was performed under the same conditions. The DNA was precipitated by the addition of 4 ml 5 M ammonium acetate and 30 ml ice-cold absolute ethanol. DNA was pelleted by centrifugation at 4500 rpm for 10 minutes and recovered with a glass hook followed by washing twice in 70% ethanol. After spinning down at 10,000g, ethanol was removed followed by 15 minutes drying at room temperature. 100 µl EB (Elution Buffer: 10 mM Tris-Cl pH 8.0, 0.2% Triton X-100) was added to the DNA and left at 5 °C during 48 hours to fully resuspend. DNA integrity was assessed by agarose gel electrophoresis. DNA extracted from both methods was analyzed on 0.5% agarose gel in 1X TAE buffer (40 mM Tris-acetate, 1.0 mM EDTA, pH 8.3) for 14 hours at 30 V to ensure a good separation of HMW fragments if present.

2.2 Oxford Nanopore Sequencing run

MinION device was used for total DNA sequencing using Ligation Sequencing Kit 1D - SQK-LSK108. Two sequencing experiments were developed: (1) the sample processed using DNA extraction kit and (2) the one processed using HMW protocol. The sequencing library preparation procedure was the same for both experiments and was performed as follows: ~3 µg of total DNA (measured by fluorometry using Qubit 3.0 dsDNA BR assay) was adjusted to a volume of 45 µl with nuclease-free water (NFW). End repair and dA-tailing were performed on extracted DNA using NEBNext Ultra II End Repair/dA-Tailing Module (NEB E7546S) by adding 7 µl of reaction buffer, 3 µl of enzyme mix and 5 µl of NFW. This mixture was incubated at 20 °C for 5 min and 65 °C for 5 min. Then a 1X volume (60 µl) AMPure XP clean-up was performed and the DNA was eluted in 31 µl NFW. 1 µl aliquot was quantified by Qubit to ensure a retention of at least 700 ng DNA. Adapter ligation was performed, then by adding 20 µl Adaptor Mix (SQK-LSK108: AMX1D) and 50 µl NEB Blunt/TA Master Mix (NEB M0367S) to the 30 µl end-prepped DNA, mixing gently by flicking and incubating at room temperature for 30 min. The adaptor-ligated DNA was cleaned up by adding a 0.4X volume (40 µl) of AMPure XP beads, incubating for 5 min at room temperature and resuspending the pellet twice in 140 µl ABB buffer (SQK-LSK108: ABB). The purified DNA was resuspended by adding 15 µl of elution buffer (SQK-LSK108: ELB), resuspending the beads, incubating at room temperature for 10 min, pelleting the beads again, and transferring the supernatant

(pre-sequencing mix) to a new tube. 1 µl aliquot was quantified by Qubit to ensure a retention of at least 430 ng DNA.

MinION sequencing was performed using R9 flow cells: FLO-MIN106 (R9.4 chemistry) for the (1) sample treated using commercial GenElute DNA extraction kit, and FLO-MIN107 (R9.5 chemistry) for the (2) sample processed using HMW-DNA protocol. Flow cells were primed prior sequencing by loading 800 µl of priming mix (48% v/v Running Fuel Buffer (SQK-LSK108:RBF) in NFW) into the flow cell via the priming port, waiting 5 minutes, lifting the SpotON sample port cover and loading 200 µl of priming mix via the priming port, always avoiding the introduction of any air bubbles.

Sequencing library, of both runs, were prepared by adding 35 µl RBF, 25.5 Library Loading Beads (Library Loading Bead kit EXP-LLB001:LLB), 0.5 NFW to 14 µl of the pre-sequencing mix. The library was loaded dropwise onto the SpotON sample port using a P100 tip set to 75 µl and entered the flow cell by capillary action.

MinION sequencing run was controlled using MinKNOW software (version 1.4.2) setting a runtime of 48 hours for each experiment without live basecalling. Subsequently, the raw signal for each run (stored as FAST5 files) was basecalled using ONT Albacore Sequencing Pipeline Software (version 2.1.10) in order to obtain sequencing reads in FASTQ and FAST5 (including sequence data) formats. In order to assess performance of both sequencing runs, base and read counts were performed, and read length histograms for each dataset and for comparing runs were

plotted. Those tasks were performed using NanoStat 1.1.0, NanoComp 0.16.4, and NanoPlot 1.13.0, which are tools from NanoPack (De Coster *et al.*, 2018).

2.3 ONT sequencing data analysis and mitogenome reconstruction

Accurate species identification relies on available data deposited on public databases to perform comparisons. For vertebrates is common to find reported abundant mitochondrial DNA sequences, mostly from *COI* and *CYTB* genes since they are widely used as DNA barcodes (Hebert *et al.*, 2003a). For those reasons, mitochondrial DNA was selected as target locus for this study. Reads presumably assigned to mtDNA were selected by a BLAST-based strategy using mtBlaster (a script written by our group https://github.com/nidafra92/squirrel-project/blob/master/mtblaster.py). *Sigmodon hispidus* mitogenome (GenBank:KY707311.1) was used as reference sequence for this search, since it was the closest evolutionary related organism to the study specimen with an available complete mitogenome.

We observed that some genes were difficult to assemble as they showed premature stop codons, indicating misassembly. In an effort to resolve this issue we produced 4 different assemblies (using reads from the sequencing experiment of HMW-DNA sample) varying parameters aimed to correct Nanopore typical basecalling errors (i.e. homopolymers and modified nucleotides). Detailed workflow for each assembly was the following:

1) draft assembly of mtDNA reads was performed using MiniMap 2.1/Miniasm 1 (Li, 2016; Li, 2018). Draft assembly was polished applying 5 rounds of mapping reads against to the previous assembly using BWA 0.7.12-r1039 (Li & Durbin, 2010) and subsequently polishing with RACON 1.2.1 (Vaser *et al.*, 2017). A final polishing step was implemented with Nanopolish 0.8.5 (Loman *et al.*, 2015) on the 5-times RACON-polished assembly. For this final polishing the following parameters were modified in the “nanopolish variants” command line: ploidy status was set to “1” (--ploidy 1), the maximum haplotype combinations was restricted to 1 (-x 1) and the homopolymer caller was enabled (--fix-homopolymers).

2) draft assembly of mtDNA reads with Minimap2/Miniasm, polished 5 times by RACON and a final polish with Nanopolish enabling methylation aware feature (-q dcm) and disabling homopolymer calling correction.

3) draft assembly of mtDNA reads with Minimap2/Miniasm, polished 5 times by RACON and polished twice with Nanopolish enabling methylation aware feature (-q dcm) and using homopolymer calling correction.

4) A draft assembly was performed using mtDNA reads with Phred score higher than Q7 but called via Chiron 0.3 deep learning-based basecaller (Teng *et al.*, 2018). Those reads were assembled using Canu v1.7 (Koren *et al.*, 2018) (genomeSize=16.3k -nanopore-raw chiron_mtDNAreads.fasta correctedErrorRate=0.2 minOverlapLength=250 corMhapSensitivity=high corMinCoverage=0 contigFilter="2 0 1.0 0.5 0"), polished twice using Racon and 3 times by Nanopolish (--max-rounds 750) using both homopolymer caller and methylation aware (-q dcm) features.

The reconstructed mitogenome sequences were annotated using GeSeq (Annotation of Organellar Genomes) from CHLOROBOX webserver hosted at the Max Planck Institute of Molecular Plant Physiology (Tilich et al., 2017). For the accurate annotation of the given sequence these parameters were tuned: “circular sequence” option was checked, sequence source was set to “Mitochondrial”, tRNAscan-SE v2.0 in “Mammalia Mitochondrial tRNAs” mode was enabled, and Server References from NCBI was selected including all RefSeqs under Muroidea taxonomic rank (NCBI:txid337687).

Finally, the four annotated mitogenome assemblies were aligned in Geneious R11.1.4 (http://www.geneious.com, Kearse *et al.*, 2012) using Geneious aligner algorithm. The fully aligned region was inspected and using information derived from the sequences; a consensus was produced carefully reviewing polymorphic positions to ensure the integrity of the reading frames of the protein coding genes and the expected secondary structure of tRNAs and rRNAs. The complete procedure for the mitogenome reconstruction is schematized in Figure 1. Coverage depth of final assembly was assessed through mapping all the mtDNA reads to the final consensus sequence after all the polishing, correction, and curation steps using BWA-mem. Finally, Tablet 1.17 (Milne et al., 2013) was used for visualization and inspection of the generated alignment.

2.4 Confirming ONT sequence accuracy using Sanger sequencing

Error rates in ONT sequencing data are relatively high compared to the ones from short-read sequencing approaches and difficulties in assemblies have been observed in A-T biased sequences, homopolymer regions and modified nucleotides (e.g DNA methylation), all features present in mtDNA. To confirm the accuracy of our mitogenome assembly method, two regions from mtDNA were amplified by PCR an then sequenced by Sanger method to analyze identity or possible discrepancies. Primers for two markers corresponding to mtDNA sequences were available: 1) cytochrome b (*CYTB*) ~1150 bp amplicon as reported in Percequillo *et al.*, 2011 and 2) cytochrome oxidase subunit I (*COI*) ~650 bp amplicon as reported in Ivanova *et al.*, 2012. Those regions were amplified using polymerase chain reaction (PCR) method from the previously extracted DNA and then sequenced by Sanger method in order to confirm ONT assembly accuracy. The amplification programs were used with the following parameters: for *CYTB*; 2 min at 95ºC , 5 cycles x (30 sec 95°C, 40 sec 52°C, 1 min 72°C), 5 cycles x (30 sec 95°C, 40 sec 54°C, 1 min 72°C), 5 cycles x (30 sec 95°C, 40 sec 56°C, 1 min 72°C), 25 cycles x (30 sec 95 ° C, 40 sec 57°C, 1 min 72°C), 10 min at 72 ° C. For *COI* 2 min at 95°C, 35 cycles x (30 sec 95°C, 40 sec 52° C, 1 min 72°C), 10 min at 72°C. All reactions were performed in a C1000 Thermal Cycler (BioRad Technologies). The amplification products were visualized on 1% agarose gel and quantified in Nanodrop (ThermoFisher Nano2000) for subsequent shipment to the Sanger sequencing service at Molecular Cloning Laboratories (MCLAB, USA).

Sanger sequences for each marker were analysed in Geneious R11.1.4 (http://www.geneious.com, Kearse *et al.*, 2012) to trim poor quality bases (Phred

>Q20) and weak signal peaks. Reverse and forward sequence for both marker were assembled to form a consensus sequence for each marker. Sanger sequences were aligned in Geneious to the complete mitogenome assembly generated with the ONT data.

Species identification using mitogenomic ONT nucleotides

To perform a molecular identification procedure from ONT raw data the following considerations were taken. Most molecular species identification are based on DNA barcode methods (Hebert *et al.*, 2003a). For almost all animal groups, a segment of the mitochondrial cytochrome c oxidase 1 gene (known as "*COX1*" or "*COI*") is used. In addition to this marker, other gene regions (depending on the taxonomic group) from mitochondrial genes are used, such as *CYTB*, *COX2*, *NAD1*, *NAD2* or mitochondrial coded rRNAs. Consequently, public data from mitochondrial loci is available for a wide range of animal groups. Given its high mutation rates, uniparental heritability, and relatively fast coalescent times (Hudson & Turelli, 2003), mtDNA sequences have been widely used for a molecular identification of metazoan organisms (Hebert *et al.*, 2003b; Stoeckle, 2003; Kerr *et al.* 2009; Blagoev *et al.*, 2016). Therefore, our molecular identification strategy from unassembled ONT reads is mtDNA-based.

Basecalled reads from HMW-DNA run were initially pre-filtered to obtain only putative metazoan mtDNA reads. In order to do that, a metazoan reference database was constructed using the whole mitogenome sequence of the well-known model organisms: nematode *Caenorhabditis elegans* (GenBank:NC_001328.1), fruit fly

Drosophila melanogaster (GenBank:NC_024511.2), mosquito *Anopheles gambiae* (GenBank:NC_002084.1), chicken *Gallus gallus* (GenBank:NC_001323.1), tunicate *Ciona intestinalis* (GenBank:NC_017929.1), lab mouse *Mus musculus* (GenBank:NC_005089.1), lab rat *Rattus norvegicus* (GenBank:AY172581.1), human *Homo sapiens* (GenBank:NC_012920.1), zebrafish *Danio rerio* (GenBank:NC_002333.2), pufferfish *Takifugu rubripes* (GenBank:NC_004299.1) and African clawed frog *Xenopus laevis* (GenBank:NC_001573.1).

An initial nucleotide BLAST (blastn) run was performed. BLAST+ 2.8 (Camacho *et al.*, 2009) was used with 60% identity cut-off value against metazoan model organism mtDNA database. Low identity cut-off was established taking into account average ONT error rate of 10% and 75% identity threshold to presume homology at nucleotide level. Positives hits from this search were classified as presumably metazoan mtDNA sequences. Recovered reads were blasted against a custom metazoan mtDNA database containing 4,686,865 metazoan mtDNA sequences from GenBank (accessed on 11 August 2018). For this search, BLAST parameters were set as follows: max_target_seqs=1, word_size=11, gapopen=2, gapextend=2, penalty=-3, reward=2, max_hsps=1, perc_identity=85, task=blastn. Stringency of BLAST parameters was lowered to mitigate the effect of mismatch and indels in positive reads due to read noisiness inherent to the sequencing technology. Based on BLAST results, an unique taxonomic assignation (species-level) was given to every single read with positives against the metazoan mtDNA sequence database. This information was used to compute a weighted frequency score for each read based on blast % identity. For instance, read #1 matched to a DNA sequence of species A with

95% identity and read #2 matched to species B with 90% identity, at this point, candidate species are species A (0.95 score) and species B (0.90), preferring classification as species A. Those scores were updated as more reads were identified using the database and this criteria. This information was summarized and represented in pie charts to suggest the possible identification of the DNA used in the sequencing run.

The identification process (above described) was performed with database delimitation at three different taxonomic levels to evaluate performance and runtimes of the routine when *a priori* information of the sample identity was available: (1) when no information is available (the complete metazoa mtDNA dataset), (2) when sample identity is known at order-level, and (3) when sample identity is known at family-level. All BLAST searches were run in a distributed manner in a computer cluster at the Apolo Scientific Computing Center, Universidad EAFIT. The distributed processing was achieved with the aid of DC-BLAST (Divide and Conquer BLAST for HPC) utility (Yim & Cushman, 2017). The setup used for executions consisted of two HPC nodes with the following specifications: 32 cores per node, 64 GB RAM per node, Intel ® Xeon ® CPU E5-2683 v4 @ 2.10GHz processor.

Phylogenetic reconstruction.

Molecular identification methods, as the one proposed here, can be used as a proxy to species identification of an unknown individual when its morphological recognition is difficult due to specimen integrity or in groups where cryptic diversity is observed. Eventhough, to achieve an accurate species identification a combination of

approaches are required. Phylogenetic reconstruction (using molecular data, morphological characters, or both) can be an important method for species identification when a dense taxon sampling is available (Weksler, 2006; Davalos & Jansa, 2004; Jansa & Weksler, 2004; Díaz-Nieto et al., 2016a; Díaz-Nieto et al., 2016b). With the intention of corroborate our findings using the automated species identification described in the previous section (Species identification using mitogenomic ONT nucleotides), we performed a phylogenetic reconstruction using the *CYTB* gene, an appropriate marker for our purposes because of it has the denser available taxon and geographic sampling within mammals (including sigmodontine rodents). Our ingroup taxon sampling included all sequences available in Genbank from the genus *Melanomys,* our JFD1322 *CYTB* sequence, and 5 additional *CYTB* sequences from *Melanomys* specimens collected at several localities near to JFD1322 (see supplemental table 1). Because the genus *Melanomys* has been previously recognized as paraphyletic by several authors (Weksler, 2006; Pine et al., 2012) our ingroup also included Genbank sequences from the genus *Sigmodontomys*. Two *Oryzomys palustris CYTB* sequences were used as outgroup for the analysis. Our complete taxon sampling can be found in supplemental table 1.

DNA sequences were aligned using MAFFT 7.271 (Katoh & Standley, 2013) in --auto mode. The best-fitting nucleotide substitution model was determined by the BIC in jModelTest 2.1.10 (Darriba *et al.*, 2012). Our CYTB matrix was analyzed using Bayesian Inference (BI) and Maximum Likelihood (ML) searches. Bayesian Inference was implemented in MrBayes v3.2 (Ronquist *et al.*, 2012) by running 4 independent Markov Chain Monte Carlo (MCMC) analyses including 1 cold chain and 3 heated

chains. The length of the chain was 10,000,000 generations, sampling every 1000 generations and implementing the optimal substitution model from jModelTest. A majority rule consensus tree was generated from the sampled trees after discarding a relative burn-in fraction of 35%. Maximum likelihood analysis was implemented in RAxML ver. 8.0 (Stamatakis, 2014) based on 10 independent searches, using the GTRGAMMA as substitution model, and evaluating nodal support by running 5,000 bootstrap pseudoreplicated datasets and also using GTRGAMMA as substitution model. Bipartitions of the bootstrap searches were summarized on the best ML topology using SumTrees from Dendropy package (Sukumaran & Holder, 2010). Finally, we calculated uncorrected *p*-distances between and within clades based on our *CYTB* aligned matrix using MEGA X (Kumar *et al.*, 2018). Calculations were made estimating variance via bootstrap with 1000 replicates, considering all substitutions (transitions+transversions), uniform rates among sites, and pairwise deletion treatment for gap/missing data.

3. Results

3.1 Performance of ONT sequencing runs

Two successfully sequencing runs were obtained from the DNA obtained using both extraction protocols. A higher DNA yield was obtained from HMW-DNA extraction protocol (65.4 ug) compared to that obtained using the GenElute extraction kit (4 ug). DNA integrity assessment on agarose gel showed little fragmentation for the sample treated with the HMW-DNA extraction protocol, whereas a highly smeared band was observed for the sample treated with the GenElute DNA extraction kit, suggesting

significant fragmentation of the DNA (supplemental figure 1). Accordingly, higher sequencing yield was observed for the sample obtained from the HMW-DNA extraction protocol (526.18 Mbp) compared to the yield obtained from the kit extraction (354.37 Mbp). Fewer reads were generated from the HMW library (139,431) compared to the kit library (352,088) as average read length was higher for the former (3,773 bp) and lower for the latter (1,006.5 bp). DNA fragmentation was considerably higher in the kit library as observed both in sequencing summaries and in agarose gel electrophoresis. Details on DNA characteristics and sequencing statistics for both sequencing libraries can be observed in table 1. In figure 1, read length distributions of both libraries can be observed showing longer fragments sequenced in HMW-DNA run compared to the GenElute kit run.

After the filter to obtain the highest possible amount of mitochondrial DNA reads (mtDNA reads) from our target sample (JFD1322), low mtDNA yields were obtained from both libraries: 0.36% of mtDNA for the GenElute kit library and 0.53% of mtDNA for the HMW library. Subsequently, we decided to apply a second filter based of read length by only keeping those reads longer than 6 kbp in order to decrease the possibility of capturing NUMTs (nuclear mitochondrial segments. After this second filter we retained 24 reads (avg. length of 8,862 bp) from the kit library and 127 reads (avg. length of 11,026 bp) from the HMW library. Assuming a 16 kpb long mitochondrial genome, the theoretical mtDNA sequencing depth was 13.25X for the kit library and 87.5X for the HMW library. Based on this data, just the HMW reads

could be successfully assembled (depth > 30X) into a mitochondrial genome. Full statistics on mtDNA reads can be seen in Table 2.

3.2 Complete mitochondrial genome of *Melanomys caliginosus*

A mtDNA molecule was successfully reconstructed from the four assembly strategies and prioritized consensus derived from HMW library reads with an average coverage depth of 77.5X (max depth of 84X). As a result our DNA sequenced ONT-only data, we obtained the resulting mitogenome assembly of *Melanomys caliginosus* JFD01322 as a closed circular 16,309-bp molecule (GenBank accession number MH939287) which contains the typical set of 37 mitochondrial genes (13 protein coding genes, 22 tRNAs and two rRNAs) (Fig. 1). A total of 28 genes were transcribed on the heavy-coding strand, while the rest (9) were transcribed on the light-coding strand. The nucleotide composition of the entire mitogenome (A: 35.0%, T:28.5%, C: 24.9% and G: 11.6%) is A+T-biased (63.5%) and exhibits positive AT-skew (0.102) and negative GC-skew (−0.362) values. Coding sequences occupy 91.7% of the total genome length. The protein-coding genes encompassed 11,387 bp of the entire assembled sequence (69.8%). Accuracy of the obtained assembly from ONT data was assessed through confirmation with Sanger sequencing of partial sequences of *COI* (662 bp) and *CYTB* (1187 bp) gene. The segments of *COI* (GenBank:MH939280) and *CYTB* (GenBank:MH939281) genes amplified by PCR

and sequenced by Sanger method were identical (100% identity) to the final curated consensus mitogenome obtained from ONT data.

3.3. Species-level identification

Taxonomic identity of our sample (JFD01322) could be obtained using the whole dataset from HMW library using the previous exposed BLAST-based methodology. When the search was performed without any prior information on the sample identification, against the whole mtDNA database of Metazoa, the analysis took 1:56:31 hours to be performed in our computer cluster;when performed against our order-level (Rodentia) dataset, the analysis took 00:04:57 minutes to be completed; and when the analysis was performed using the family-level (Cricetidae) dataset it took 00:01:47 minutes to be completed. Results of the three analyses agreed on the identification of the sample, showing assignation of ~80% of the valid reads (mtDNA reads) with the *Melanomys* genus. And the most likely identification for the library, at species-level, was *Melanomys caliginosus*, corresponding with the morphological identification of the collected voucher.

When the search was performed with the whole metazoan dataset few hits (5.39%) from different animal groups (other than rodents) could be observed (Architaenioglossa, Hemiptera, Siluriformes). Such event could be explained as possible contamination of our specimen since is was collected from a wild habitat, next to a stream, with the reported presence of individuals belonging to those animal groups: snails of the clade Architaenioglossa, catfish from the order Siluriformes and psyllid from the order Hemiptera (Stevenson et al., 2006). This fact shows the

sensitivity of the sequencing method and also proves this information useful to get information on the ecological context of an unknown metazoan DNA sample using ONT sequencing. A graphical summary of this results can be seen in figure 4.

3.4. Phylogenetic reconstruction

In addition to the taxonomic identification based on ONT molecular dataset, and the initial identification of the specimen based on morphological data, a *CYTB* phylogenetic reconstruction was performed to evaluate the identification of our sample JFD01322 in a taxonomic context, and not only relying on GenBank identifications. .

The phylogenetic tree (figure 3) shows haplotypes of the genus *Melanomys* as a strongly supported monophyletic clade (based on PP). It shows four well-supported major clades within *Melanomys*, with uncorrected p-distances varying from 0.0629 to 0.0701, but no clear relationships among them are found based on their low support values. The sequence derived from our specimen (voucher JFD1322) is found in a strongly supported clade of *Melanomys* haplotypes of other material collected from Valle de Aburra (same geographic location as JFD01322). Haplotypes of genus *Sigmodontomys* and *Tanyuromys*. are each part of a monophyletic group independent of *Melanomys* haplotypes. Following the identification of CYTB haplotypes from the GenBank, the species *M. caliginosus* would be paraphyletic, as no monophyletic clade is recovered from haplotypes with this identification. However,

clade C includes topotypical material (D'Elía & Pardiñas, 2016) and consequently would represent the nominal form *caliginosus*; clade A would correspond to *M. chrysomelas* (following Hanson and Bradley, 2008); and clade D would include the names *indoneus* and *columbianus* (following Hanson and Bradley, 2008). The appropriate name applied to clade B is uncertain based on the lack of resolution of this clade with any of the other clades for which a nominal form is available.

4. Discussion

Species identification efforts have improved using molecular data or 'holistic datasets' that complement taxonomic information obtained only from morphology of a given individual. Availability of molecular datasets (partial genes or genomic) is still restricted to certain animal groups; many groups in which high diversity occurs still lack proper molecular information which difficult addressing evolutionary questions, biogeographic hypothesis, and accurate species recognition and delimitation. Traditional barcoding approaches have been proved useful but they are not suitable for certain taxa lacking references genomes or specific primers for PCR amplification when universal primers do not work as expected. The use of Oxford Nanopore sequencing stands as an alternative to facilitate the obtention of molecular datasets from organellar genomes (via genome skimming) for metazoan research where mitochondrial genomes could be easily generated PCR-free. Given accessibility and cost-benefit of the technology this could be an useful tool to reduce the gap in mitochondrial DNA sequence knowledge for non-model organism research and biodiversity studies.

Herein, we used MinION device from ONT to explore genome skimming strategy to generate mitochondrial genomes using ONT-only data. Previous studies have shown complete mtDNA assemblies combining ONT datasets and high throughput short-read datasets from Illumina platform to improve overall consensus quality (Torres et al. 2018; Chandler et al. 2017). The complete mitogenome assembly of a clinical-interest nematode, *Nippostrogylus brasiliensis*, was improved using dual datasets (Chandler et al. 2017). When mapping raw reads to the assembly discrepancies with consensus could be observed possibly due to DNA methylation and epigenetic modifications. To solve these issues, authors used Illumina data for error correction.

We observed same complications when reconstructing the mitochondrial assembly using a simple *de novo* assembly pipeline from mtDNA reads (Minimap/Miniasm-RACON). During the annotation process we observed plenty of truncated ORFs for several genes due to premature stop codons. Those misassemblies occurred in polynucleotide regions mostly, but they were also observed in other sequences contexts. This could be due to the presence modified nucleotides which affects basecalling, hence misasigning bases. To tackle this problem we used the correction features included in Nanopolish (methylation aware correction and homopolymer correction) but on its own, those approaches were not enough to obtain an assembly free of error, although they fixed some frameshifts and substitutions. Combining different assemblies routines in basecalling and correction steps (number of runs of RACON or Nanopolish) allowed to obtain the likely basecall

that does not interrupt the functionality in at least one of the assemblies. Our four different procedures produced an scenario to generate a prioritized consensus sequences of the mitochondrial genome; when polymorphisms were present, we chose the one which preserved the reading frame of the protein coding sequence or the RNA stability, depending the region where it occurred. Despite of basecalling issues to be solved in highly repetitive regions or in DNA contexts when nucleotide modification is common, it is possible to obtain *de novo* a complete mitochondrial genome (without sequencing errors) using ONT-only sequencing data with enough coverage (above 50X). In addition, partial sequences of CYTB and *COI* generated from Sanger sequencing showed 100% identity with our ONT final assembly.

We also explore the use of sequencing reads prior any assembly strategy to apply a molecular species identification method for metazoan based on mitochondrial information. The approach used here could successfully identify the organism of study due to the use of mitochondrial data even in its uncorrected form (i.e. raw reads without any polishing, error rate correction or any assembly). The selection of mtDNA for this task was crucial since its variation rate is sufficient to discriminate taxa at species level (Hudson & Turelli, 2006). Use of nuclear information was avoided for being uninformative. Mammals, specially Neotropical and non-model ones, lack genomic information in DNA sequence databases (Lessa *et al.*, 2014). Nuclear sequences possess lower mutation rates compared to mitochondrial loci, specially in highly conserved genomic regions of vertebrates (UCEs and structural regions) (Hudson & Turelli, 2006). Prefiltering sequencing reads using low stringency BLAST

parameters allowed to recover reads from a variety of metazoan groups. This is advantageous for several reasons: the identification is not researcher-biased to give results for certain taxonomic groups (i.e. cricetids only, in this case). This gives improved sensitivity as observed in the complete analysis (fig. 4A). Hits from different orders were recovered with considerably high identities (above 90%). This allow to get ecological information of the sample (its context) and suggest possible organisms located nearby. For the molecular species identification could be observed that any *a priori* information to delimitate the database reduces significantly the search time, allowing to obtain an identification in a matter of minutes. This strategy could be used in a future to perform the identification in a real time wise manner while the sequencer is running.

Acknowledgements

This research was funded by Universidad EAFIT with internal research projects support in 2017 and 2018, MSc scholarship (Universidad EAFIT) to NDF, and a research fellowship for young researchers (COLCIENCIAS program “Jóvenes Investigadores por la paz 2017 - convocatoria 775-2017”) awarded to NDF under Javier Correa Alvarez’s mentorship. The authors acknowledge supercomputing resources made available by the Centro de Computación Científica Apolo at Universidad EAFIT (http://www.eafit.edu.co/apolo). We would like to thank Juan M. Martínez, Valentina Grisales, Mauricio Serna, and Manuela Londoño at Universidad EAFIT for their help in fieldwork and laboratory experiments. We also thank Beatriz Aristizábal at HPTU Hospital and Uriel Hurtado at CIB for their kind assistance

providing lab equipment and molecular biology reagents to complete specific experiments when needed.

References

Avise, J. C. (1994). *Molecular Markers, Natural History and Evolution*. Boston, MA: Springer US. http://doi.org/10.1007/978-1-4615-2381-9

Benítez-Páez, A., Portune, K. J., & Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. *GigaScience*, *5*(1), 1–9. http://doi.org/10.1186/s13742-016-0111-z

Blagoev, G. A., deWaard, J. R., Ratnasingham, S., deWaard, S. L., Lu, L., Robertson, J., … Hebert, P. D. N. (2016). Untangling taxonomy: A DNA barcode reference library for Canadian spiders. *Molecular Ecology Resources*, *16*(1), 325–341. http://doi.org/10.1111/1755-0998.12444

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 421. http://doi.org/10.1186/1471-2105-10-421

Chandler, J., Camberis, M., Bouchery, T., Blaxter, M., Le Gros, G., & Eccles, D. A. (2017). Annotated mitochondrial genome with Nanopore R9 signal for Nippostrongylus brasiliensis. *F1000Research*, *6*(0), 56. http://doi.org/10.12688/f1000research.10545.1

Coissac, E., Hollingsworth, P. M., Lavergne, S., & Taberlet, P. (2016). From barcodes to genomes: Extending the concept of DNA barcoding. Molecular Ecology, 25(7), 1423–1428. http://doi.org/10.1111/mec.13549

Crawford, A. J., Cruz, C., Griffith, E., Ross, H., Ibáñez, R., Lips, K. R., … Crump, P. (2012). DNA barcoding applied to ex situ tropical amphibian conservation programme reveals cryptic diversity in captive populations. *Molecular Ecology Resources*, *13*(6), n/a-n/a. http://doi.org/10.1111/1755-0998.12054

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). JModelTest 2: More models, new heuristics and parallel computing. *Nature Methods*. http://doi.org/10.1038/nmeth.2109

Davalos, L. M., & Jansa, S. A. (2004). Phylogeny of the Lonchophyllini (Chiroptera: Phyllostomidae). J Mammalogy. http://doi.org/10.1644/1545-1542(2004)085<0404:POTLCP>2.0.CO;2

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, *34*(March), 2666–2669. http://doi.org/10.1093/bioinformatics/bty149

D'Elía, G., & Pardiñas, U. F. J. (2015). Mammals of South America. In J. L. Patton, U. F. J. Pardiñas, & G. D'Elía (Eds.), *Mammals of South America, Volume 2: Rodents*. Chicago, IL: The University of Chicago Press.

de Queiroz, K. (2005). Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences*. http://doi.org/10.1073/pnas.0502030102

Díaz-Nieto, J. F., Jansa, S. A., & Voss, R. S. (2016a). Phylogenetic relationships of Chacodelphys (Marsupialia: Didelphidae: Didelphinae) based on "ancient" DNA sequences. Journal of Mammalogy. http://doi.org/10.1093/jmammal/gyv197

Díaz-Nieto, J. F., Jansa, S. A., & Voss, R. S. (2016b). DNA sequencing reveals unexpected Recent diversity and an ancient dichotomy in the American marsupial genus Marmosops (Didelphidae: Thylamyini). Zoological Journal of the Linnean Society. http://doi.org/10.1111/zoj.12343

Erlich, Y. (2015). A vision for ubiquitous sequencing. *Genome Research*, *25*(10), 1411–6. http://doi.org/10.1101/gr.191692.115

Faria, N. R., Sabino, E. C., Nunes, M. R. T., Alcantara, L. C. J., Loman, N. J., & Pybus, O. G. (2016). Mobile real-time surveillance of Zika virus in Brazil. *Genome Medicine*, *8*(1), 97. http://doi.org/10.1186/s13073-016-0356-2

Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., … Chiu, C. Y. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Medicine*, *7*(1), 99. http://doi.org/10.1186/s13073-015-0220-9

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003a). Biological identifications through DNA barcodes. Proceedings of the Royal Society B: Biological Sciences, 270(1512), 313–321. http://doi.org/10.1098/rspb.2002.2218

Hebert, P. D. N., Ratnasingham, S., & de Waard, J. R. (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B: Biological Sciences*, *270*(Suppl_1), S96–S99. http://doi.org/10.1098/rsbl.2003.0025

Hudson, R. R., & Turelli, M. (2003). STOCHASTICITY OVERRULES THE "THREE-TIMES RULE": GENETIC DRIFT, GENETIC DRAFT, AND COALESCENCE TIMES FOR NUCLEAR LOCI VERSUS MITOCHONDRIAL DNA. Evolution, 57(1), 182–190. http://doi.org/10.1111/j.0014-3820.2003.tb00229.x

Ivanova, N. V, Clare, E. L., & Borisenko, A. V. (2012). *DNA Barcodes in Mammals*. (W. J. Kress & D. L. Erickson, Eds.) (Vol. 858). Totowa, NJ: Humana Press. http://doi.org/10.1007/978-1-61779-591-6

Jansa, S. A., & Weksler, M. (2004). Phylogeny of muroid rodents: Relationships within and among major lineages as determined by IRBP gene sequences. Molecular Phylogenetics and Evolution. http://doi.org/10.1016/j.ympev.2003.07.002

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 256. http://doi.org/10.1186/s13059-016-1122-x

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., … Loose, M. (2018a). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*(4), 338–345. http://doi.org/10.1038/nbt.4060

Jain, M., Olsen, H. E., Turner, D. J., Stoddart, D., Bulazel, K. V, Paten, B., … Miga, K. H. (2018b). Linear assembly of a human centromere on the Y chromosome. Nature Biotechnology, 36(4), 321–323. http://doi.org/10.1038/nbt.4109

Juul, S., Izquierdo, F., Hurst, A., Dai, X., Wright, A., Kulesha, E., … Turner, D. J. (2015). What's in my pot? Real-time species identification on the MinION. *BioRxiv*, 030742. http://doi.org/10.1101/030742

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*. http://doi.org/10.1093/molbev/mst010

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., … Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. http://doi.org/10.1093/bioinformatics/bts199

Kerr, K. C. R., Lijtmaer, D. A., Barreira, A. S., Hebert, P. D. N., & Tubaro, P. L. (2009). Probing evolutionary patterns in neotropical birds through DNA barcodes. *PLoS ONE*, *4*(2). http://doi.org/10.1371/journal.pone.0004379

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. http://doi.org/10.1101/gr.215087.116

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.

Molecular Biology and Evolution, *35*(6), 1547–1549. http://doi.org/10.1093/molbev/msy096

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, *3*, 1–8. http://doi.org/10.1016/j.bdq.2015.02.001

Lessa, E. P., Cook, J. a., D 'Elía, G. & Opazo, J. C. (2014). Rodent diversity in South America: transitioning into the genomics era. *Frontiers in Ecology and Evolution*, *2*(July), 1–7. http://doi.org/10.3389/fevo.2014.00039

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589–595. http://doi.org/10.1093/bioinformatics/btp698

Li, H. (2016). Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, *32*(14), 2103–2110. http://doi.org/10.1093/bioinformatics/btw152

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. http://doi.org/10.1093/bioinformatics/bty191

Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, *12*(8), 733–735. http://doi.org/10.1038/nmeth.3444

Magi, A., Giusti, B., & Tattini, L. (2016). Characterization of MinION nanopore data for resequencing analyses. Briefings in Bioinformatics, 18(6), bbw077. http://doi.org/10.1093/bib/bbw077

Mayr, E. (1982). The Growth of Biological Thought: Diversity, Evolution, and Inheritance. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. http://doi.org/10.1016/0162-3095(84)90038-4

Mendoza, Á. M., Torres, M. F., Paz, A., Trujillo-Arias, N., López-Alvarez, D., Sierra, S., … Gonzalez, M. A. (2016). Cryptic diversity revealed by DNA barcoding in Colombian illegally traded bird species. *Molecular Ecology Resources*, *16*(4), 862–873. http://doi.org/10.1111/1755-0998.12515

Milne, I., Stephen, G., Bayer, M., Cock, P. J. A., Pritchard, L., Cardle, L., … Marshall, D. (2013). Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*, *14*(2), 193–202. http://doi.org/10.1093/bib/bbs012

Norris, A. L., Workman, R. E., Fan, Y., Eshleman, J. R., & Timp, W. (2016). Nanopore sequencing detects structural variants in cancer. *Cancer Biology*

and Therapy, *17*(3), 246–253. http://doi.org/10.1080/15384047.2016.1139236

Pine, R. H., Timm, R. M., & Weksler, M. (2012). A newly recognized clade of trans-Andean Oryzomyini (Rodentia: Cricetidae), with description of a new genus. Journal of Mammalogy, 93(3), 851–870. http://doi.org/10.1644/11-MAMM-A-012.1

Pomerantz, A., Peñafiel, N., Arteaga, A., Bustamante, L., Pichardo, F., Coloma, L. A., … Prost, S. (2018). Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *GigaScience*, *7*(4), 1–14. http://doi.org/10.1093/gigascience/giy033

Prain, T. 1956. On a collection of *Pomacea* from Colombia, with description of a new subspecies.Journal of Conchology 24: 73-79.

Quick, J., Quinlan, A. R., & Loman, N. J. (2014). A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *GigaScience*, *3*(1), 22. http://doi.org/10.1186/2047-217X-3-22

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., … Carroll, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, *530*(7589), 228–232. http://doi.org/10.1038/nature16996

Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., … Loman, N. J. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature Protocols*, *12*(6), 1261–1276. http://doi.org/10.1038/nprot.2017.066

Quick, J. (2018). Ultra-long read sequencing protocol for RAD004. protocols.io. http://doi.org/10.17504/protocols.io.mrxc57n

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., … Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, *61*(3), 539–42. http://doi.org/10.1093/sysbio/sys029

Salazar, A. N., Gorter de Vries, A. R., van den Broek, M., Wijsman, M., de la Torre Cortés, P., Brickwedde, A., … Abeel, T. (2017). Nanopore sequencing enables near-complete de novo assembly of *Saccharomyces cerevisiae* reference strain CEN.PK113-7D. *FEMS Yeast Research*, *17*(7). http://doi.org/10.1093/femsyr/fox074

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., & Gerstein, M. B. (2011). The real cost of sequencing: higher than you think! *Genome Biology*, *12*(8), 125. http://doi.org/10.1186/gb-2011-12-8-125

Schäffer, S., Zachos, F. E., & Koblmüller, S. (2017). Opening the treasure chest: A DNA-barcoding primer set for most higher taxa of Central European birds and mammals from museum collections. *PLOS ONE*, *12*(3), e0174449. http://doi.org/10.1371/journal.pone.0174449

Schmidt, K., Mwaigwisya, S., Crossman, L. C., Doumith, M., Munroe, D., Pires, C., … Livermore, D. M. (2017). Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *Journal of Antimicrobial Chemotherapy*, *72*(1), 104–114. http://doi.org/10.1093/jac/dkw397

Srivathsan, A., Baloğlu, B., Wang, W., Tan, W. X., Bertrand, D., Ng, A. H. Q., … Meier, R. (2018). A MinION™-based pipeline for fast and cost-effective DNA barcoding. *Molecular Ecology Resources*, 0–2. http://doi.org/10.1111/1755-0998.12890

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. http://doi.org/10.1093/bioinformatics/btu033

Stevenson, P., Pérez-Torres, J., & Muñoz-Saba, Y. (2006). Informe Nacional sobre el Avance en el Conocimiento y la Información de la Biodiversidad 1998 – 2004. In Informe Nacional sobre el Avance en el Conocimiento y la Información de la Biodiversidad 1998 – 2004. http://doi.org/10.1016/0167-5087(83)90973-0

Stoeckle, M. (2003). Taxonomy, DNA, and the Bar Code of Life. *BioScience*, *53*(9), 796. http://doi.org/10.1641/0006-3568(2003)053[0796:TDATBC]2.0.CO;2

Sukumaran, J., & Holder, M. T. (2010). DendroPy: A Python library for phylogenetic computing. *Bioinformatics*, *26*(12), 1569–1571. http://doi.org/10.1093/bioinformatics/btq228

Teng, H., Cao, M. D., Hall, M. B., Duarte, T., Wang, S., & Coin, L. J. M. (2018). Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *GigaScience*, *7*(5), 1–9. http://doi.org/10.1093/gigascience/giy037

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*, *45*(W1), W6–W11. http://doi.org/10.1093/nar/gkx391

Torres, L., Welch, A. J., Zanchetta, C., Chesser, R. T., Manno, M., Donnadieu, C., … Pante, E. (2018). Evidence for a duplicated mitochondrial region in Audubon's shearwater based on MinION sequencing. *Mitochondrial DNA Part A*, *0*(0), 1–8. http://doi.org/10.1080/24701394.2018.1484116

Tyson, J. R., O'Neil, N. J., Jain, M., Olsen, H. E., Hieter, P., & Snutch, T. P. (2018). MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. *Genome Research*, *28*(2), 266–274. http://doi.org/10.1101/gr.221184.117

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. http://doi.org/10.1101/gr.214270.116

Weksler, M. (2006). Phylogenetic Relationships of Oryzomine Rodents (Muroidea: Sigmodontinae): Separate and Combined Analyses of Morphological and Molecular Data. Bulletin of the American Museum of Natural History, 296, 1–149. http://doi.org/10.1206/0003-0090(2006)296[0001:PROORM]2.0.CO;2

Wetterstrand, K. A. (2018). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Retrieved October 10, 2018, from https://www.genome.gov/sequencingcostsdata

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. http://doi.org/10.1186/gb-2014-15-3-r46

Yim, W. C., & Cushman, J. C. (2017). Divide and Conquer (DC) BLAST: fast and easy BLAST execution within HPC environments. *PeerJ*, *5*, e3486. http://doi.org/10.7717/peerj.3486

Data Accessibility

NCBI BioProject ID: PRJNA492505; NCBI BioSample accession: SAMN10103913; NCBI SRA accession: PRJNA492505.

Final mitochondrial DNA assembly: GenBank accession MH939287.

Partial mitochondrial gene sequences by Sanger method: GenBank accessions MH939280-MH939286.

Author Contributions

NDF and JFD conceived and designed the study; JFD performed fieldwork, specimen sampling and morphological identification; NDF carried out molecular experiments, ONT sequencing runs and computational data analysis; NDF wrote the early manuscript draft, which was reviewed and improved by JFD.

Tables and Figures (with captions)

FIGURE 1 Computational procedure followed for the mitochondrial genome reconstruction based on only ONT sequencing data.

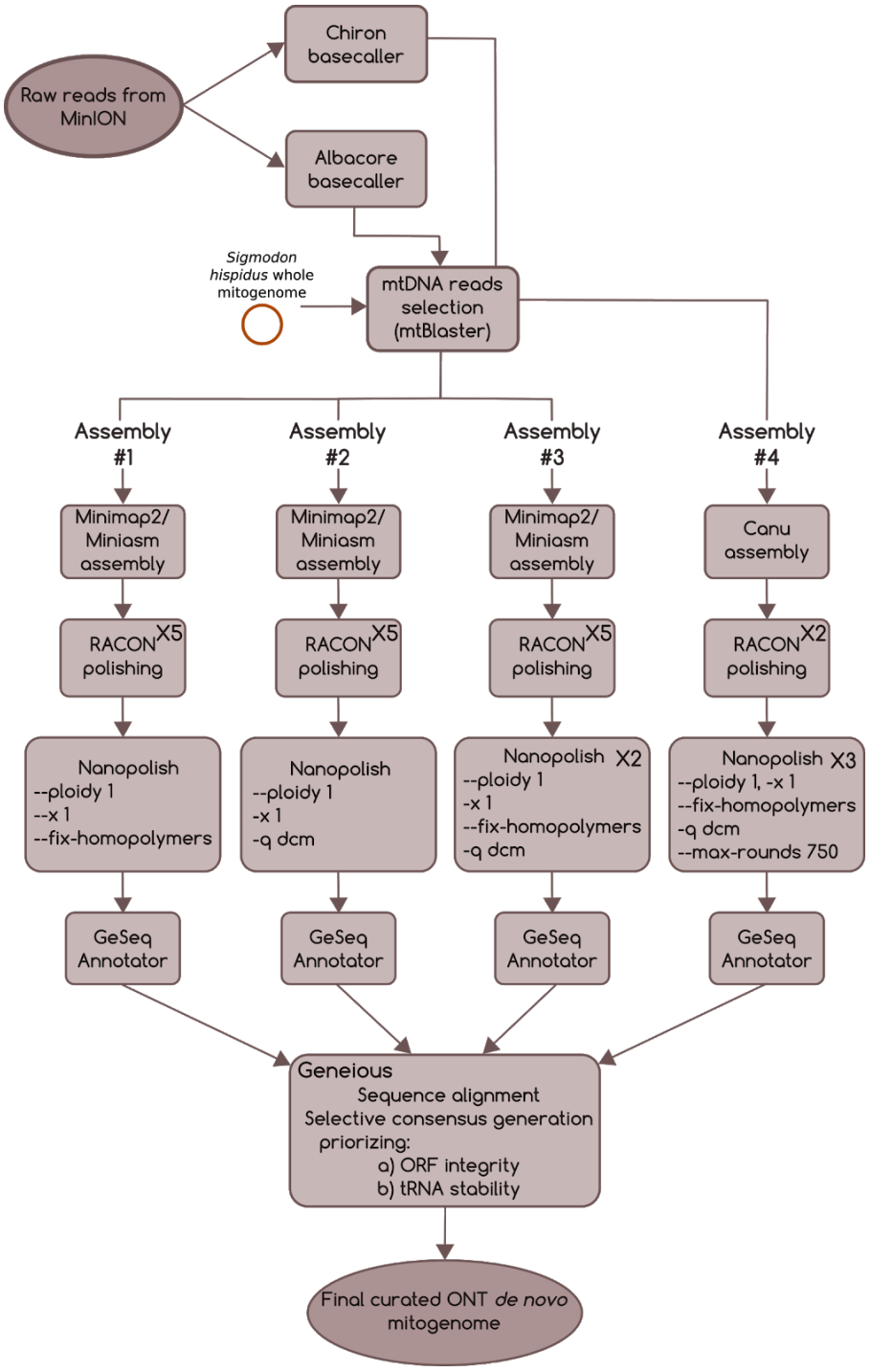


FIGURE 2 Distribution of sequencing read lengths generated from both libraries (GenElute Extraction Kit, HMW DNA). **A.** Distribution of read length after log transformation for both sequencing libraries **B.** Weighted histogram of read length after log transformation for both sequencing libraries **C.** Box plot comparing both sequencing runs based in their log-transformed read lengths

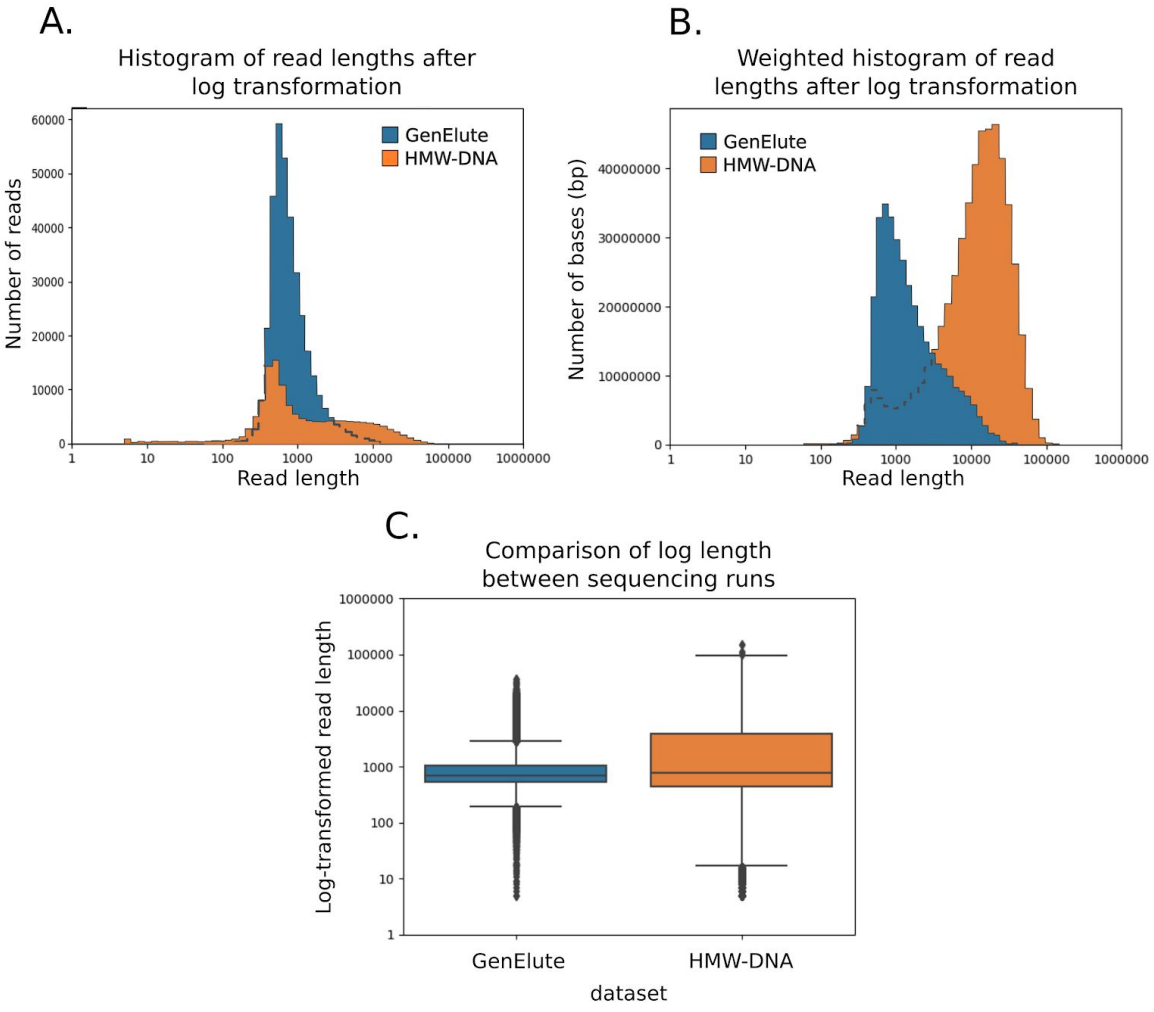


FIGURE 3 Circular map of the mitochondrial genome of *Melanomys caliginosus* (specimen JFD01322) obtained from ONT sequencing. Color-legend indicating gene type is shown in bottom left. Inner circle shows %GC content and light-gray arrows indicate direction of transcription.

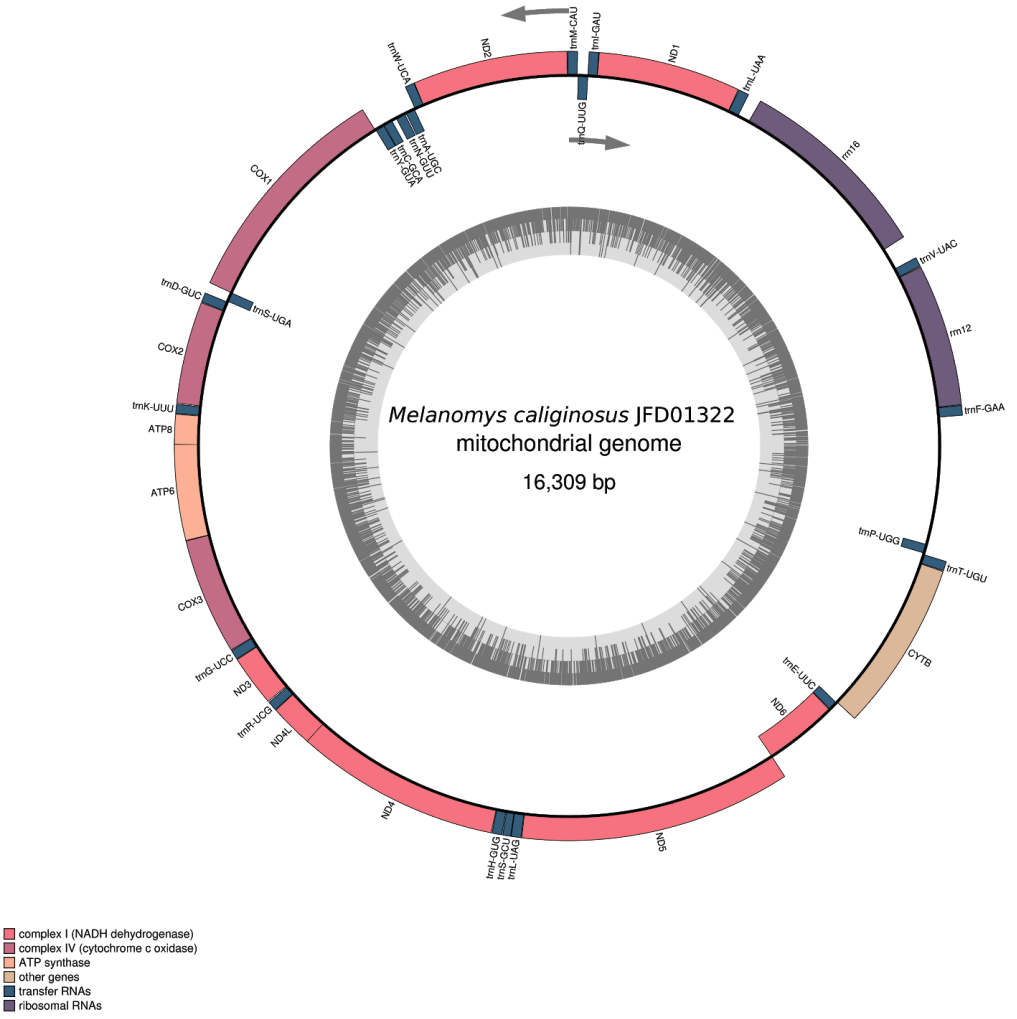


FIGURE 4 Species identification using ONT nucleotide data. **A.** results when previous information of the sample identity is limited to Metazoa. **B.** results when previous information of sample identity is available at order-level. **C.** results when previous information of sample identity is available at family-level. ***** all analyses were run in a computer cluster (OS: Linux Rocks 6.2) using two nodes, with 64 GB RAM and 32 cores per node. Processor type of nodes used is Intel ® Xeon ® CPU E5-2683 v4 @ 2.10GHz.

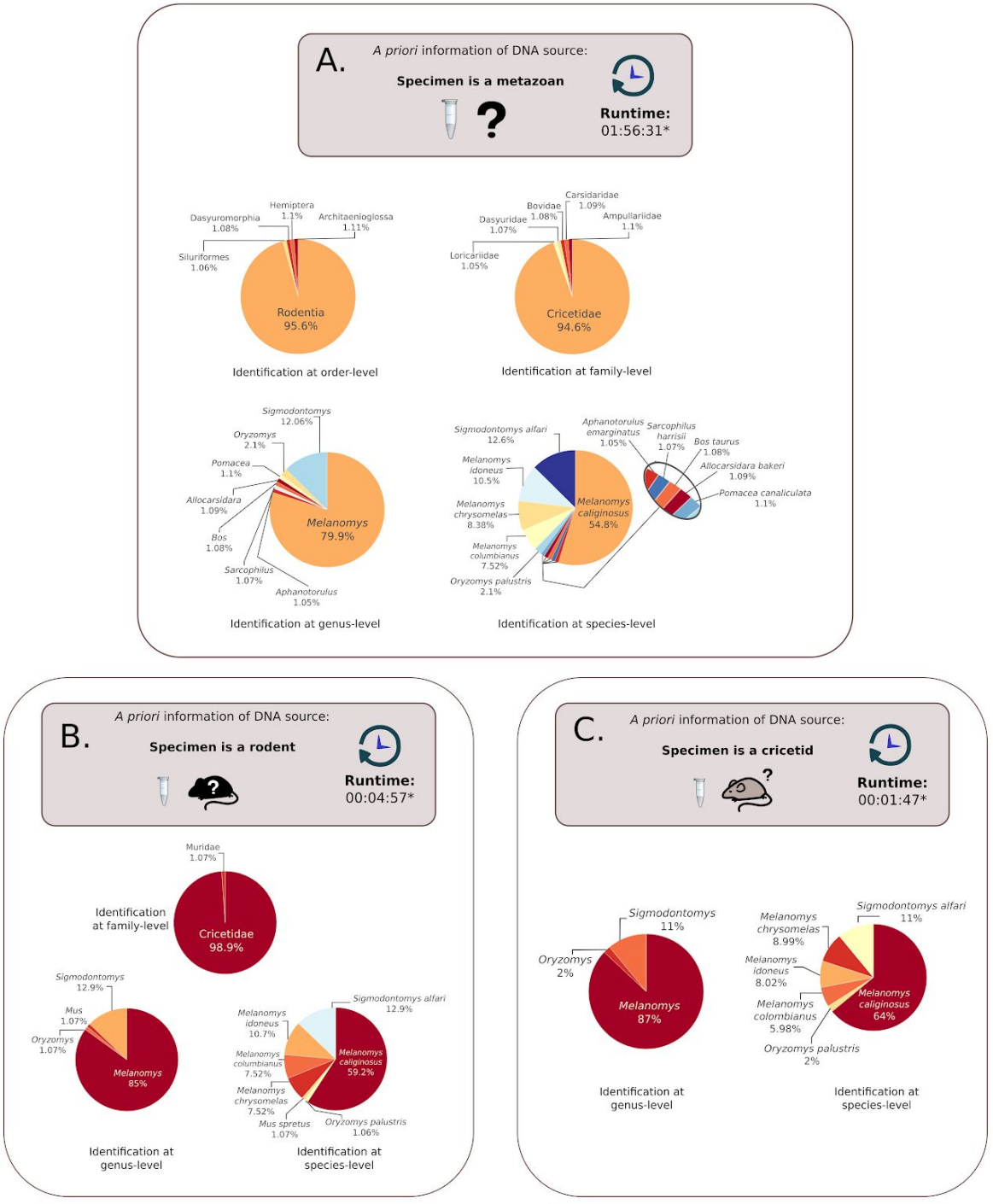


FIGURE 5 Phylogenetic reconstruction of *Melanomys* and *Sigmodontomys CYTB* sequences. **A.** Geographical distribution of haplotypes used in the phylogenetic analysis. **B.** Phylogenetic tree reconstructed by Bayesian Inference. Terminals include GenBank accession numbers (bolded letters) followed by the identification obtained from GenBank, collection localities are found in parenthesis, and numbers are mapped in figure A. Red terminal corresponds to sample JFD1322. Support values are shown on nodes (Posterior Probabilities/Bootstrap Support values). Color boxes represent our main ingroup mitochondrial haplogroups recovered by our analyses.

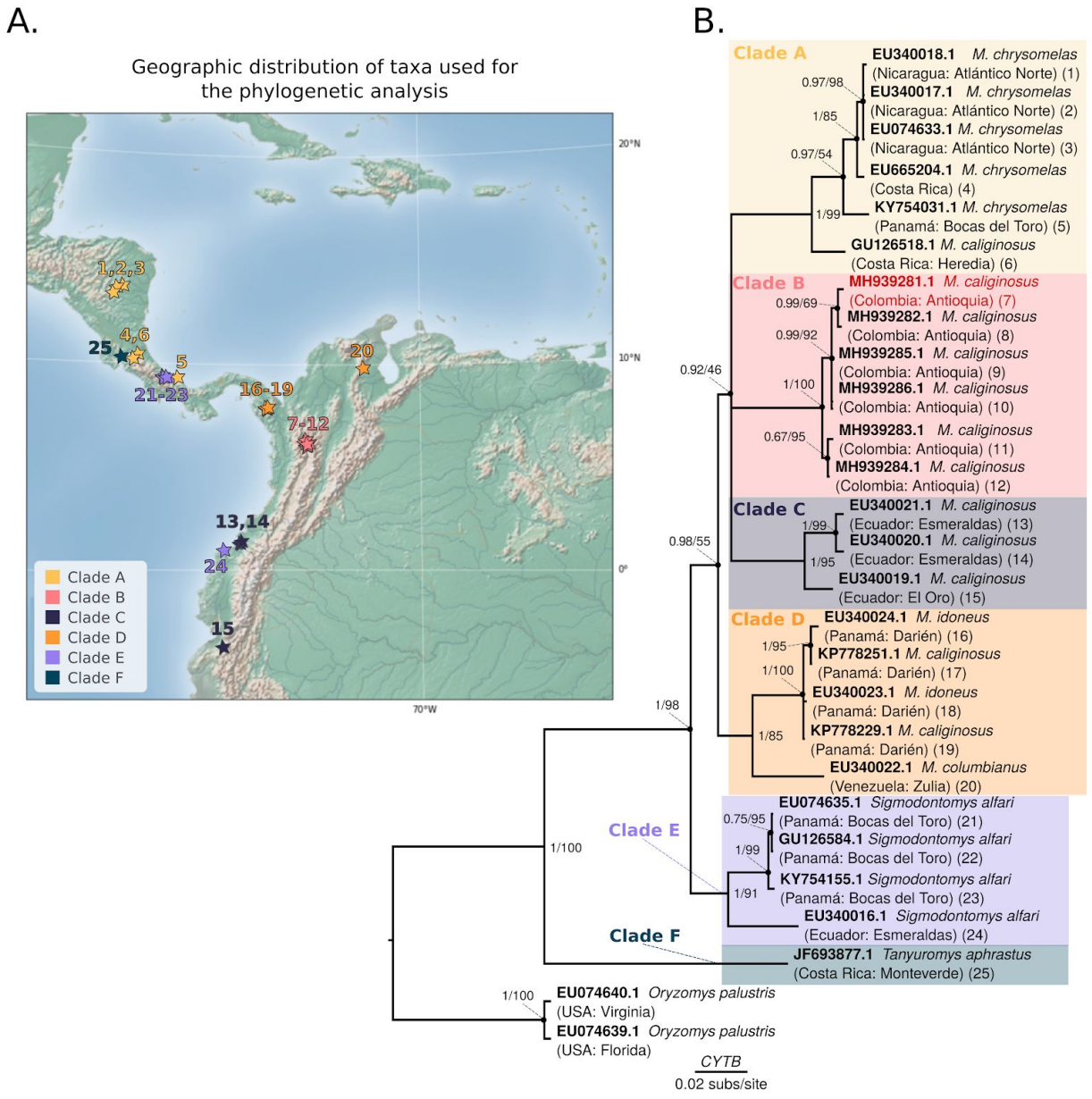


TABLE 1 Sequencing performance of two MinION runs using total DNA extracted from liver of specimen JFD1322 of the species *Melanomys caliginosus*.

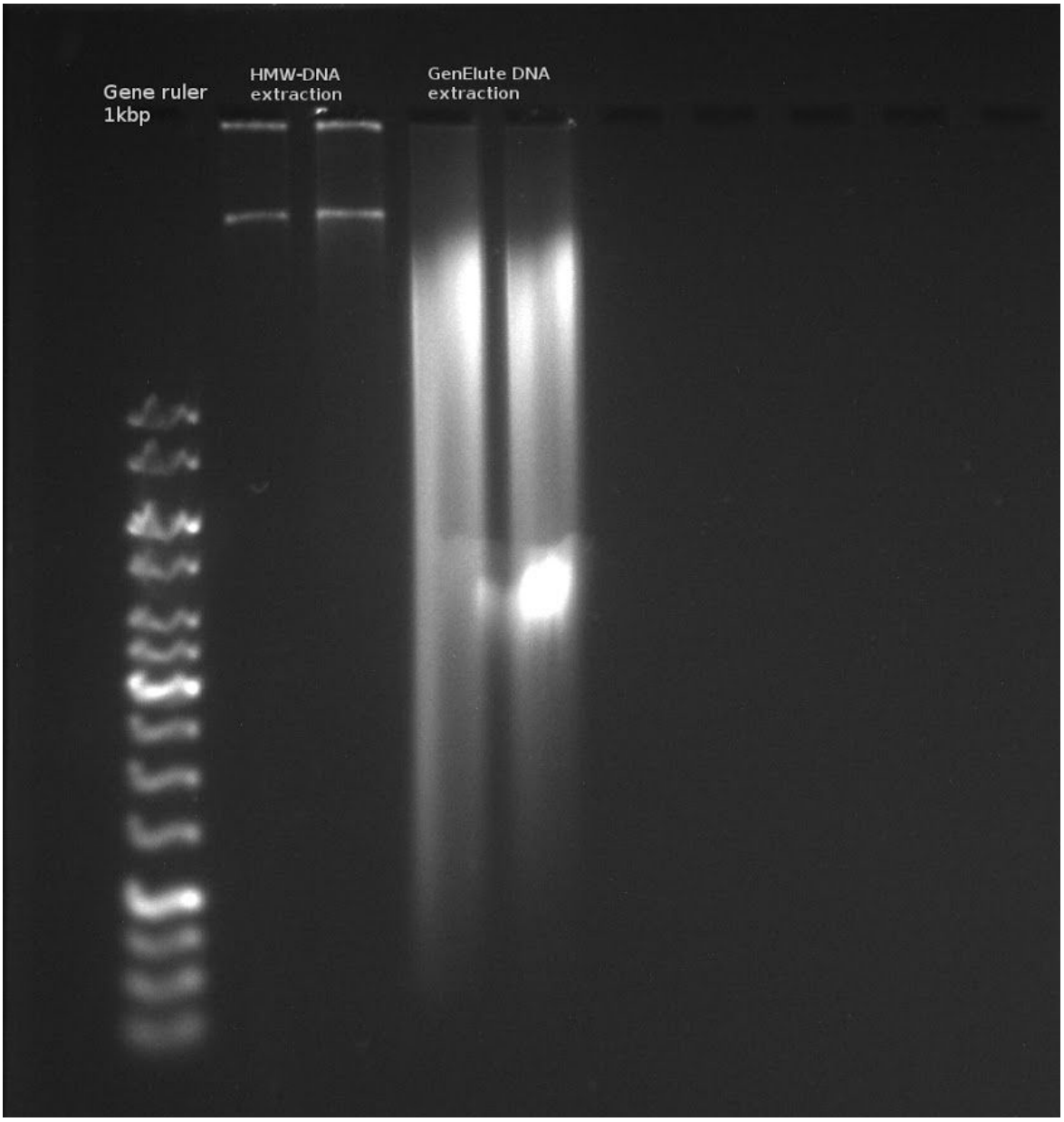
	DNA extraction method	
	GenElute	**HMW DNA protocol**
DNA yield (ng DNA/25mg tissue)	4,077	65,400
Flow cell version	FLO-MIN106 (R9.4)	FLO-MIN107 (R9.5)
Sequencing run yield (Mbp)	354.37	526.18
Number of reads	352,088	139,431
Average read length (bp)	1006.5	3773
Average base quality (Phred score)	8.9	6.4
Longest read (bp)	36,744	149,424

TABLE 2 Sequencing yield of mitochondrial DNA for both runs inferred from the total DNA sequencing data using mtBlaster.

	DNA extraction method	
	GenElute	**HMW DNA protocol**
mtDNA sequencing yield (Mbp)	1.28 (0.36%)	2.8 (0.53%)
Number of mtDNA reads	1,045	1,609
Average mtDNA read length (bp)	1,227	1,756
Average base quality (Phred score)	9.4	8.1
Median mt DNA read length (bp)	744	919
mtDNA in reads >6 kbp (Mbp)	0.212 (0.06%)	1.4 (0.27%)
Number of mtDNA reads > 6kbp	24	127
Average mtDNA reads >6 kbp read length (bp)	8,862	11,026
Theoretical mtDNA sequencing depth (for a 16 kbp mitogenome)	13.25X	87.5X

Supporting information

Supplemental figure 1. low-agarose gel electrophoresis (0.5%). GeneRuler 1 kb was used as DNA ladder (lane 1). DNA obtained from the HMW protocol shows little smearing pattern and high molecular weights, over 10 kbp (lanes 2 and 3). DNA obtained from extraction using commercial GenElute kit shows a strong smearing pattern (lanes 4 and 5).



Supplemental table 1. Taxa used for the phylogenetic reconstruction

Phylogenetic group	Taxa	Accession number	Location
Clade A	*Melanomys chrysomelas* voucher TK121431	EU074633.1	Nicaragua: RAAN, Rosa Grande
	Melanomys chrysomelas voucher TK121427	EU340018.1	Nicaragua: Atlántico Norte, Rosa Grande, Siuna
	Melanomys chrysomelas voucher TK121417	EU340017.1	Nicaragua: Atlántico Norte, Rosa Grande, Siuna
	Melanomys chrysomelas voucher RMT4658	EU665204.1	Costa Rica
	Melanomys chrysomelas	KY754031.1	Panamá: Bocas Del Toro, Isla Bastimentos, Old Point
	Melanomys caliginosus	GU126518.1	Costa Rica: Prov. Heredia, La Flaminia
Clade B	*Melanomys caliginosus* voucher JFD01322	MH939281.1	Colombia: Antioquia, Envigado, 'Loma del Chocho'
	Melanomys caliginosus voucher DAG00150	MH939282.1	Colombia: Antioquia, Caldas, Reserva del Alto de San Miguel
	Melanomys caliginosus voucher JFD00237	MH939283.1	Colombia: Antioquia, Sabaneta, Vereda la Doctora
	Melanomys caliginosus voucher JFD00248	MH939284.1	Colombia: Antioquia, Bello, Vereda Sabana Larga
	Melanomys caliginosus voucher YXR00014	MH939285.1	Colombia: Antioquia, Sabaneta, La Romera
	Melanomys caliginosus voucher YXR00016	MH939286.1	Colombia: Antioquia, Sabaneta, La Romera
Clade C	*Melanomys caliginosus* voucher TK135895	EU340021.1	Ecuador: Esmeraldas, Comuna San Fransisco de Bogotá
	Melanomys caliginosus voucher TK135894	EU340020.1	Ecuador: Esmeraldas, Comuna San Fransisco de Bogotá
	Melanomys caliginosus	EU340019.1	Ecuador: El Oro, Zaruma,

	voucher TK135789		Cerro Urcu
Clade D	*Melanomys idoneus* voucher TK22586	EU340024.1	Panamá: Darién, Cana
	Melanomys idoneus voucher ROM116303	EU340023.1	Panamá: Darién, Cana
	Melanomys caliginosus voucher LSUMZ:M-579	KP778251.1	Panamá: Cana
	Melanomys caliginosus voucher LSUMZ:M-568	KP778229.1	Panamá: Cana
	Melanomys columbianus voucher MHNLS7698	EU340022.1	Venezuela: Zulia, Misión Tukuko
Clade E	*Sigmodontomys alfari*	KY754155.1	Panamá: Bocas Del Toro, Nuri
	Sigmodontomys alfari voucher USNM449895	EU074635.1	Panamá: Boca del Toro, Isla San Cristóbal
	Sigmodontomys alfari	GU126548.1	Panamá: Bocas del Toro, Isla San Cristóbal
	Sigmodontomys alfari voucher TK135621	EU340016.1	Ecuador: Esmeraldas, Estación Experimental 'La Chiquita'
Clade F	*Tanyuromys aphrastus* voucher KU161003	JF693877.1	Costa Rica: Monteverde Cloud Forest Reserve
Outgroup	*Oryzomys palustris* voucher SCVA15	EU074640.1	USA: Virginia, Norfolk Co
	Oryzomys palustris voucher EVGL06	EU074639.1	USA: Florida, Miami-Dade Co, Everglades National Park

Supplemental table 2. Uncorrected *p*-distance matrix calculated from *CYTB* sequence alignment of our ingroup (including species of *Melanomys* and *Sigmodontomys*) (see text). Between group mean distances are shown in lower left matrix, within group mean distances are shown along the diagonal.

	Clade A	**Clade B**	**Clade C**	**Clade D**	**Clade E**	**Clade F**
Clade A	0.0149					
Clade B	0.0685	0.0040				
Clade C	0.0701	0.0660	0.0181			
Clade D	0.0651	0.0651	0.0629	0.0193		
Clade E	0.0697	0.0645	0.0660	0.0622	0.0207	
Clade F	0.1175	0.1227	0.1195	0.1191	0.1179	-