Yes, we can use it: A formal test on the accuracy of low-pass nanopore long-read sequencing for mitophylogenomics + barcoding research using the Caribbean spiny lobster Panulirus argus

J. Antonio Baeza
Department of Biological Sciences, Clemson University, Clemson, SC 29634, USA. Email: jbaezan@clemson.edu.

ABSTRACT
Whole mitochondrial DNA (mtDNA) genomes are the subject of choice for re-sequencing efforts and among species genomics. Protocols for sequencing and assembling mitogenomes of long reads have been growing in number and techniques over the last decade. Despite the diversity of protocols, most studies have compared mitogenome assemblies generated de novo to those assembled with reference-based assembly pipelines. This comparison is critical as one can evaluate the accuracy of the mitogenome assemblies. Furthermore, few studies have explored how long read assembly pipelines generate accurate mitogenomes. Here, we present a proof-of-concept study using long reads to sequence the mitochondrial genome of Panulirus argus, a Caribbean spiny lobster of economic and biological significance. We explored the accuracy of long-read mitogenome assemblies using short & long reads across a single species mitogenome. The results demonstrated that long-read mitogenome assemblies are highly accurate, even with some additional steps to correct errors in the alignment of the long-read assemblies. This study suggests that long reads can be used for mitogenome sequencing and assembly, especially when there is limited material.

INTRODUCTION
The aims of this study were threefold. First, I tested if a mitochondrial genome can be sequenced and assembled from long read nanopore sequencing data exclusively using both de novo and reference-based assembly pipelines. Second, I explored the quality (i.e., accuracy) of the long-read-based assembled genome by comparing them to a gold-standard mitochondrial genome retrieved from the same individual but generated using short read Illumina sequencing data. Sequence accuracy was explored for different long read assembly pipelines with multiple metrics including completeness, identity, and coverage. Furthermore, a detailed quantification of analysis error type for long read assemblies was conducted. Third and lastly, I tested if the de novo and reference-based long read assemblies are useful for mitophylogenomic and barcoding research. This study was conducted using short & long reads from a single Caribbean spiny lobster species, Panulirus argus. The number of single nucleotide homopolymer deletions was by far the most common error in the alignment of the long-read assemblies to the reference assembly that was mostly due to indels at the flanks of homopolymer regions comprising all four nucleotide types.

MATERIAL & METHODS
Field collections were approved by FWCC (SAL-11-1319-DR). A single adult specimen of P. argus was collected from Alligator Reef, FL, USA. A small piece of muscle was dissected from a dismembered pereopod and immediately preserved in 95% ethanol. Next, total genomic DNA from ~15-25 mg muscle tissue was extracted using standard protocols. Approximately one microgram of gDNA extracted from each sample was used for Illumina and MinION ONT library construction following standard protocols.

The pipeline Novoplasty was used for mitogenome assembly with short reads. For long reads, the mitochondrial genome of P. argus was assembled using the following workflow:

Accuracy of long read mitogenome assemblies
Alignment of the different long read assemblies to the reference genome revealed discordance between each of the long read assemblies and the reference assembly that was mostly due to indels at the flanks of homopolymer regions comprising all four nucleotide types.

The number of single nucleotide homopolymer deletions was by far the most common error detected in all long read assemblies followed by single nucleotide homopolymer insertions. The main effect of polishing with Medaka, across de novo- and reference-based mitochondrial genomes, was a decrease in the number of homopolymer deletions (Fig. 2).

CONCLUSIONS
In conclusion, using nanopore long read sequencing and various bioinformatics pipelines, this study assembled for the first time a complete and highly accurate mitochondrial genome for the Caribbean spiny lobster Panulirus argus. The assembled genomes were ‘imperfect’ but permitted to identify reliably the sequenced specimen as belonging to P. argus and differentiate the specimen from other closely and distantly related species in the same genus, family, and superorder. This study will facilitate the transferring of genomic technologies to low-income countries in the greater Caribbean region with which to monitor mislabeling of this resource.

ACKNOWLEDGEMENTS
Many thanks to Dr. Tom Matthews (FWCC, Florida). Many thanks to Dr. Vinca P. Richards (Clemson University) for bioinformatics support.

RESULTS
Mitogenome assembly using short & long reads
Novoplasty assembled and circularized the mitochondrial chromosome of P. argus (15,728 bp) with an average coverage of 710x (MH068821). The mtDNA of P. argus comprised 13 protein-coding genes (PCGs), 2 ribosomal RNA genes (rRNAs), and 22 transfer RNA (tRNA) genes. All long read pipelines, with and without ‘extra’ polishing with Medaka, also assembled and circularized the mitochondrial genome of P. argus. All long read assemblies were very similar (a close match) to the reference genome (patristic distance range: 6.3621x10^-5 - 6.3755x10^-4) (Fig. 1).

Annotation of long read mitogenome assemblies
Annotation of long read assembled mitochondrial genomes, either de novo or reference-based with or without extra polishing with Medaka, indicated that gene number and synteny were identical to that of the reference genome of P. argus. Importantly, all but 1-2 of the genes did have at least one and usually more than one internal stop codon that interrupted the open reading frame. Although highly accurate, the errors in each long read assembled mitochondrial genome precluded generating a reliable in silico annotation (Fig. 3).

Mitophylogenomics using long read assemblies
In a maximum likelihood phylogenetic analysis using a short 500 bp fragment of the cox1 gene (1899 terminals retrieved from GenBank), long and short read assembled mitogenomes clustered together into a single well supported monophyletic clade (bootstrap value [bv] = 100). Importantly, in the same analysis, this clade comprising long read assembled mitogenomes plus the short read reference assembly and a total of 340 other sequences belonging to P. argus obtained from GenBank clustered together into a strongly supported monophyletic clade [bv = 98] (Fig. 4).

Figure 1. Panulirus argus. Circular representation of the mitochondrial genome. The mtDNA comprises 13 protein-coding genes (PCGs), 2 ribosomal RNA genes (rRNAs), and 22 transfer RNA (tRNAs) genes.

Figure 2. Sequence errors per de novo (Flye and Unicycler) and reference-based assemblies (Rebaker) with and without ‘extra’ polishing with the software Medaka for the Caribbean spiny lobster Panulirus argus mitochondrial genome.

Figure 3. Annotation of all novel (Flye & Unicycler) and reference-based (Rebaker) assemblies with and without ‘extra’ polishing with the software Medaka for the Caribbean spiny lobster Panulirus argus mitochondrial genome.

Figure 4. Barcoding analysis of the order Achelata using a 500 bp fragment of the cox1 gene, including cox1/S1 gene fragments retrieved from mitochondrial genome of the Caribbean spiny lobster Panulirus argus assembled with long reads alone and short reads + 1 884 other species belonging to the order Achelata retrieved from GenBank (the barcode comprising long and short reads mitogenomes obtained during this study).

Figure 5. Annotation of all novel (Flye & Unicycler) and reference-based (Rebaker) assemblies with and without ‘extra’ polishing with the software Medaka for the Caribbean spiny lobster Panulirus argus mitochondrial genome.

ACKNOWLEDGEMENTS
Many thanks to Dr. Tom Matthews (FWCC, Florida). Many thanks to Dr. Vinca P. Richards (Clemson University) for bioinformatics support.

Figure 6. Annotation of all novel (Flye & Unicycler) and reference-based (Rebaker) assemblies with and without ‘extra’ polishing with the software Medaka for the Caribbean spiny lobster Panulirus argus mitochondrial genome.