DATA NOTE **Open Access**

Draft assembly and annotation of the Cuban crocodile (*Crocodylus rhombifer*) genome

Robert W. Meredith[1,2*], Yoamel Milián-García[3], John Gatesy[2*], Michael A. Russello[4*] and George Amato[2*]

Abstract

Objectives The new data provide an important genomic resource for the Critically Endangered Cuban crocodile (*Crocodylus rhombifer*). Cuban crocodiles are restricted to the Zapata Swamp in southern Matanzas Province, Cuba, and readily hybridize with the widespread American crocodile (*Crocodylus acutus*) in areas of sympatry. The reported de novo assembly will contribute to studies of crocodylian evolutionary history and provide a resource for informing Cuban crocodile conservation.

Data description The final 2.2 Gb draft genome for *C. rhombifer* consists of 41,387 scaffolds (contigs: N50 = 104.67 Kb; scaffold: N50-518.55 Kb). Benchmarking Universal Single-Copy Orthologs (BUSCO) identified 92.3% of the 3,354 genes in the vertebrata_odb10 database. Approximately 42% of the genome (960Mbp) comprises repeat elements. We predicted 30,138 unique protein-coding sequences (17,737 unique genes) in the genome assembly. Functional annotation found the top Gene Ontology annotations for Biological Processes, Molecular Function, and Cellular Component were regulation, protein, and intracellular, respectively. This assembly will support future macroevolutionary, conservation, and molecular studies of the Cuban crocodile.

Keywords Cuban crocodile, *Crocodylus rhombifer*, Critically Endangered, Conservation, Genome assembly, Genome annotation, Genomics

Objective

Crocodiles (Crocodylidae) are large semi-aquatic predators found throughout the tropics of Asia, Australia, Africa, and the Americas. Of the three extant genera (*Crocodylus*, *Osteolaemus*, and *Mecistops*) within Crocodylidae, *Crocodylus* is the largest, comprising 13 currently recognized species. The Cuban crocodile (*Crocodylus rhombifer*) is a Critically Endangered [1] island endemic, currently restricted to the smallest range of any extant member of the genus [2]. Fossil evidence suggests that it may be a Pleistocene relict formerly much more widespread in the Caribbean and Bahama islands [3, 4]. Now only found naturally in the Zapata Swamp in southern Matanzas Province, Cuba, *C. rhombifer* is restricted to the unique freshwater ecosystem characteristic of the Zapata peninsula. A long history of over-harvesting

*Correspondence:
Robert W. Meredith
meredithr@montclair.edu
John Gatesy
jgatesy@amnh.org
Michael A. Russello
michael.russello@ubc.ca
George Amato
gamato@amnh.org; georgeamato.ct@gmail.com
[1]Department of Biology, Montclair State University, Montclair, NJ, USA
[2]Institute for Comparative Genomics, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, USA
[3]Department of Integrative Biology, University of Guelph, Guelph, ON N1G 2W1, Canada
[4]Department of Biology, University of British Columbia, 3247 University Way, Kelowna, BC V1V 1V7, Canada

and land conversion continues to threaten this declining population. In addition, hybridization with the widespread American crocodile (*Crocodylus acutus*) in areas of sympatry may be an additional anthropogenic threat exacerbated by freshwater management and habitat modification activities [2].

A number of distinguishing morphological and behavioral traits have been described for this species [5, 6]. These include prominent cranial 'horns', heavy-scaled and colorful skin, robust skull structures, adaptations for a more terrestrial lifestyle, and aggressive, intelligent hunting strategies [5, 7]. Previous phylogenetic and phylogenomic studies are ambiguous about the exact phylogenetic placement of *C. rhombifer* within the monophyletic Neotropical *Crocodylus* radiation [2, 8–10]. Sequencing of whole genomes provides the best opportunity to test hypotheses concerning the biogeographic history and the evolution of novel morphological and behavioral traits. Such information may further offer insights into conservation threats and opportunities for this enigmatic species. Presented here is the first genome assembly for the Cuban crocodile.

Data description

For a detailed description of all methods see Table 1, Data file 1. High molecular weight DNA was extracted from a non-hybrid Cuban crocodile ([2]; Table 1, Data file 2) using the QIAGEN® MagAttract HMW DNA Kit. 10X Genomics Chromium Genome library preparation and sequencing was performed at the New York Genome Center. The libraries were 150 bp paired-end sequenced on an Illumina HiSeqX machine (1,717.59 million reads at ~65X coverage; mean read length of 138.5 bp; Table 1, Data file 3).

Two assemblies were performed. First, the linked reads were assembled into 41,387 scaffolds (contigs: N50=104.67 Kb; scaffolds: N50=518.55 Kb) using the Supernova assembler (v 2.1.1; [11]). The estimated genome size was 2.61 GB, and the assembly size was 2.20 Gb. The Supernova scaffolds were screened for contaminants via the NCBI Foreign Contamination Screen (https://github.com/ncbi/fcs), resulting in 39,474 scaffolds. For the second build, the Supernova assembly was run through RagTag [12] with the *Crocodylus porosus* genome (Cpor 3.0; [13]) as a reference. The

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	Table 1, *Data file 1 Detailed description of the methodology*	*Microsoft Word (.docx)*	Figshare: https://doi.org/10.6084/m9.figshare.25388386 [32]
Data file 2	Table 1, *Data file 2 Photos of Cuban crocodiles*	*Microsoft Word (.docx)*	Figshare: https://doi.org/10.6084/m9.figshare.25388386 [32]
Data file 3	Table 1, *Data file 3_C.rhombifer10X_Assembly_statistics*	*Spreadsheet (.xlsx)*	Figshare: https://doi.org/10.6084/m9.figshare.25388386 [32]
Data file 4	Table 1, *Data file 4 BUSCO Comparisons*	*Microsoft Word (.docx)*	Figshare: https://doi.org/10.6084/m9.figshare.25388386 [32]
Data file 5	Table 1, *Data file 5 Interspersed Repeat landscape*	*Microsoft Word (.docx)*	Figshare: https://doi.org/10.6084/m9.figshare.25388386 [32]
Data file 6	Table 1, *Data file 6 Percentages of repeat elements*	*Microsoft Word (.docx)*	Figshare: https://doi.org/10.6084/m9.figshare.25388386 [32]
Data file 7	Table 1, *Data file 7 C.rhombifer10X_Pannzer*	*Portable document format (.pdf)*	Figshare: https://doi.org/10.6084/m9.figshare.25388386 [32]
Data file 8	Table 1, *Data file 8 Venn diagram*	*Microsoft Word (.docx)*	Figshare: https://doi.org/10.6084/m9.figshare.25388386 [32]
Data file 9	Table 1, *Data file 9 GO_counts_Table*	*Spreadsheet (.xlsx)*	Figshare: https://doi.org/10.6084/m9.figshare.25388386 [32]
Data file 10	Table 1, *Data file 10_Orthofinder_Results_Crocs_Only*	*Spreadsheet (.xlsx)*	Figshare: https://doi.org/10.6084/m9.figshare.25388386 [32]
Data file 11	Table 1, *Data file 11_Statistics_PerSpecies*	*Spreadsheet (.xlsx)*	Figshare: https://doi.org/10.6084/m9.figshare.25388386 [32]
Data file 12	Table 1, *Data file 12 Concordance factor statistics*	*Spreadsheet (.xlsx)*	Figshare: https://doi.org/10.6084/m9.figshare.25388386 [32]
Data file 13	Table 1, *Data file 13 Phylogeny*	*Microsoft Word (.docx)*	Figshare: https://doi.org/10.6084/m9.figshare.25388386 [32]
Data set 1	*Sequencing reads of C. rhombifer genomic DNA*	*Fastq files (.fq.gz)*	NCBI SRA Database: SAMN36978604 https://identifiers.org/ncbi/bioproject:PRJNA1005273 [33]
Data set 2	*Genomic Assembly of C. rhombifer*	*Fasta file (.fa)*	NCBI GenBank Database: JAVSML000000000 https://identifiers.org/nucleotide:JAVSML000000000 [34]

RagTag assembly placed 19,264 contigs (25,753 scaffolds; N50=6,528.07 Kb: Table 1, Data file 3).

Completeness and quality of the two *C. rhombifer* genomic builds were assessed by Benchmarking Universal Single-Copy Orthologs (BUSCO v5.1.2; [14]) using the vertebrata_odb10 database (3,354 markers) and compared to published Crocodylia genomes (Table 1, Data file 4). The Supernova build had 91.3% of the BUSCO genes complete (single and duplicate), 5.1% fragmented (171 genes), and 2.6% missing (85 genes). The RagTag build had 95% of the BUSCO genes complete (single and duplicate), 3.0% fragmented (102 genes), and 2.0% missing (67 genes) (Table 1, Data file 4).

RepeatModeler and RepeatMasker [15] and Earl Gray [16, 17] identified ~ 1000Mbp of the builds as interspersed repeat elements. Retroelements (17–18%) and Unclassified (16–18%) were the most common (Table 1, Data file 5, 6). Protein sequences were predicted using two ab initio methodologies BRAKER2 [18–23] and MetaEuk [24]. This resulted in 30,138 unique protein-coding sequences (17,737 unique genes) (Table 1, Data file 7). PANNZER2 [25] was used for functional annotation. The top gene ontology annotations for biological processes, molecular function, and cellular component were regulation, protein, and intracellular, respectively (Table 1, Data files 8, 9). Orthofinder [26, 27] was used to perform comparative genomic analyses between all published crocodylian genomes. A total of 175,928 genes were compared among the five species. Of these, 93.5% were placed into 26,551 orthogroups, with 0.6% of genes in species-specific orthogroups (Table 1, Data files 10, 11).

BUSCO Phylogenomics [28] identified and aligned 1,912 single-copy BUSCO genes present in 12 taxa (five Crocodylia; seven outgroups). IQ-TREE inferred the maximum-likelihood concatenated protein tree with bootstrap support [29–31]. All recovered nodes had 100% bootstrap support (Table 1, Data file 12, 13).

Limitations

The draft genome was generated using short-read shotgun sequencing via 10X genomics for a scale sample. As a result, the assembly is somewhat fragmented and smaller than the genome size estimate. The Cuban crocodile is naturally restricted to a developing country (Cuba) with limited research resources and access to sequencing technology. Consequently, obtaining genomic data from a non-hybrid wild caught specimen was limited to the most accessible sequencing technology available at the time of collection. If and when more funds become available, the completeness and accuracy of the genome will be built upon using long-read sequencing technologies.

Abbreviations

Kb	kilobases
Gb	gigabases
Mbp	million base pairs
bp	basepair
BUSCO	Benchmarking Universal Single-Copy Orthologs
IUCN	International Union for the Conservation of Nature

Acknowledgements

We dedicate this work to the memory of Vicente Berovides Álvarez (El Bero), whose passion for Cuban fauna conservation knew no bounds, committing his professional life to this noble task with exceptional support to Cuban crocodiles' protection. Bero was a referent in the field and supervisor of all generations of Cuban crocodile specialists. His legacy will always remain among us. We thank Roberto Ramos Targarona's (Toby) team, particularly Gustavo Sosa Rodríguez and Etiam Pérez Fleitas, for arduous field expeditions to Zapata Swamp that permitted collecting the sample used in this study and for setting the basis for genomics studies of Cuban crocodiles. Toby's dedication and passion for Cuban crocodiles' conservation will remain a permanent source of motivation for present and future Cuban crocodile specialists. We are deeply grateful to Georgina Espinosa López for acquiring sample permits and establishing collaborations between the University of Havana and the University of British Columbia Okanagan (UBCO). Lastly, we thank all members of the Ecological & Conservation Genomics Laboratory at UBCO.

Author contributions

YMG, GA, and JG designed the study. YMG selected and provided the sample. RWM assembled, annotated, and analyzed the genome. MAR provided resources and guidance in support of this work. RWM wrote the initial draft of the manuscript, and all authors contributed to the writing and editing of subsequent drafts.

Funding

This work was supported by NSF grants DEB-1556701, DBI-1725932 to RWM, Natural Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN-2019-04621 to MAR, Rufford Foundation for Nature Conservation (RSG reference 19318-B) to YMG, who was also supported by Mitacs through the Mitacs Elevate Program. The funders did not contribute to the study design, data collection, data analysis, or manuscript preparation.

Data availability

The data described in this Data note can be freely and openly accessed on NCBI under BioProject PRJNA1005273, BioSamples SAMN36978604 [33]. The Supernova genome assembly can be found at NCBI under Accession No. JAVSML000000000 [34]. Please see Table 1 and references [32-34] for details and links to the data.

Declarations

Ethics approval and consent to participate

This sample was previously used by Milián-García et al. [2]. The sample was originally collected and transported under CITES permits C0001166 and C0001455 and an agreement between the Faculty of Biology at the University of Havana and the National Enterprise for the Protection of Flora and Fauna in Cuba.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 17 March 2024 / Accepted: 3 June 2024
Published online: 06 June 2024

References

1. IUCN. The IUCN Red List of Threatened Species. IUCN Red List of Threatened Species. 2023. https://www.iucnredlist.org/en. Accessed 19 Apr 2023.

2. Milián-García Y, Ramos-Targarona R, Pérez-Fleitas E, Sosa-Rodríguez G, Guerra-Manchena L, Alonso-Tabet M, et al. Genetic evidence of hybridization between the critically endangered Cuban crocodile and the American crocodile: implications for population history and in situ/ex situ conservation. Heredity. 2015;114:272–80.
3. Morgan GS, Franz R, Crombie RI. The Cuban crocodile, *Crocodylus rhombifer*, from late Quaternary fossil deposits on Grand Cayman. 1993;:12.
4. Steadman DW, Franz R, Morgan GS, Albury NA, Kakuk B, Broad K, et al. Exceptionally well preserved late quaternary plant and vertebrate fossils from a blue hole on Abaco, the Bahamas. PNAS. 2007;104:19897–902.
5. Targarona RR. Ecologia y conservación del cocodrilo Cubano (*Crocodylus rhombifer*) en la Ciénaga De Zapata, Cuba. Universitat d'Alacant - Universidad de Alicante; 2013. http://purl.org/dc/dcmitype/Text.
6. Ross JP. Crocodiles: status survey and conservation action plan. 1998.
7. Murphy JB, Evans M, Augustine L, Miller K. Behaviors in the Cuban crocodile (*Crocodylus rhombifer*). Herpetological Rev. 2016.
8. Milián-García Y, Castellanos-Labarcena J, Russello MA, Amato G. Mitogenomic investigation reveals a cryptic lineage of *Crocodylus* in Cuba. Bull Mar Sci. 2018;94:329–43.
9. Milián-García Y, Amato G, Gatesy J, Hekkala E, Rossi N, Russello M. Phylogenomics reveals novel relationships among Neotropical crocodiles (*Crocodylus* spp). Mol Phylogenet Evol. 2020;152:106924.
10. Milián-García Y, Russello MA, Castellanos-Labarcena J, Cichon M, Kumar V, Espinosa G, et al. Genetic evidence supports a distinct lineage of American crocodile (*Crocodylus acutus*) in the Greater Antilles. PeerJ. 2018;6:e5836.
11. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. Genome Res. 2017;27:757–67.
12. Alonge M, Lebeigle L, Kirsche M, Aganezov S, Wang X, Lippman ZB. Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. bioRxiv. 2021; 2021.11. 18.469135. 2021.
13. Ghosh A, Johnson MG, Osmanski AB, Louha S, Bayona-Vásquez NJ, Glenn TC, et al. A high-quality reference genome assembly of the saltwater crocodile, *Crocodylus porosus*, reveals patterns of selection in Crocodylidae. Genome Biol Evol. 2020;12:3635–46.
14. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2018;35:543–8.
15. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C et al. RepeatModeler2: automated genomic discovery of transposable element families. preprint. Genomics; 2019.
16. Baril T, Galbraith JG, Hayward A. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. Mol Biol Evol. 2024;41:msae068. https://doi.org/10.1093/molbev/msae068.
17. Baril T, Galbraith JG, Hayward A. Earl Grey. Zenodo. 2023;https://doi.org/10.5281/zenodo.5654615.
18. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP + and AUGUSTUS supported by a protein database. NAR Genomics Bioinf. 2021;3:lqaa108.
19. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32:767–9.
20. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. In: Kollmar M, editor. Gene prediction: methods and protocols. New York, NY: Springer; 2019. pp. 65–95.
21. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
22. Gotoh O. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. Nucleic Acids Res. 2008;36:2630–8.
23. Iwata H, Gotoh O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. Nucleic Acids Res. 2012;40:e161.
24. Levy Karin E, Mirdita M, Söding J. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome. 2020;8:48.
25. Törönen P, Medlar A, Holm L. PANNZER2: a rapid functional annotation web server. Nucleic Acids Res. 2018;46:W84–8.
26. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.
27. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238.
28. McGowan J. jamiemcg/BUSCO_phylogenomics. 2024.
29. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 2018;35:518–22.
30. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–4.
31. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.
32. Meredith RW, Milián-García Y, Gatesy J, Russello MA, Amato G. Datasets of the Cuban crocodile (*Crocodylus rhombifer*) genome. 2024. Figshare, https://doi.org/10.6084/m9.figshare.25388386.
33. Meredith RW, Milián-García Y, Gatesy J, Russello MA, Amato G. NCBI SRA database of the Cuban crocodile (*Crocodylus rhombifer*) genome. NCBI; 2023. https://identifiers.org/ncbi/bioproject:PRJNA1005273.
34. Meredith RW, Milián-García Y, Gatesy J, Russello MA, Amato G. Datasets of the Cuban crocodile (*Crocodylus rhombifer*) genome. NCBI; 2023. https://identifiers.org/nucleotide:JAVSML000000000.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.