Check for updates

GENOME NOTE

A draft genome sequence of the common, or spectacled caiman *Caiman crocodilus* [version 1; peer review: 2 not approved]

Kenichi Okamoto ¹, Nichole Dopkins¹, Elias Kinfu^{1,2}¹Department of Biology, University of Saint Thomas, Saint Paul, MN, 55105, USA²Department of Biochemistry, University of Washington, Seattle, WA, 98195, USA

V1 First published: 02 Dec 2021, 10:1230
<https://doi.org/10.12688/f1000research.73066.1>

Latest published: 02 Dec 2021, 10:1230
<https://doi.org/10.12688/f1000research.73066.1>

Abstract

The common, or spectacled, caiman *Caiman crocodilus* is an abundant, widely distributed Neotropical crocodilian exhibiting notable morphological and molecular diversification. The species also accounts by far for the largest share of crocodilian hides on the global market, with the *C. crocodilus* hide trade alone valued at about US\$86.5 million per year. We obtained 239,911,946 paired-end reads comprising approximately 72 G bases using Illumina™ sequencing of tissue sampled from a single *Caiman crocodilus* individual. These reads were de-novo assembled and progressively aligned against the genomes of increasingly related crocodilians; liftoff was used to annotate the draft *C. crocodilus* genome assembly based on an *Alligator mississippiensis* (a confamilial species) annotation. The draft assembly has been submitted to the National Center for Biotechnology Information and both assembly and annotation are currently available for review at doi.org/10.5281/zenodo.4755063

Keywords

Caiman crocodilus, spectacled caiman, genome, assembly, next-generation sequencing, crocodilian, vertebrate genome

Open Peer Review

Approval Status

	1	2
version 1
02 Dec 2021	view	view
1. Steven L. Salzberg , Johns Hopkins University, Baltimore, USA Johns Hopkins University, Baltimore, USA Johns Hopkins University, Baltimore, USA Johns Hopkins School of Medicine, Baltimore, USA		
2. Marc Tollis , Northern Arizona University, Flagstaff, USA		
Any reports and responses or comments on the article can be found at the end of the article.		

Corresponding author: Kenichi Okamoto (kenichi.okamoto@stthomas.edu)

Author roles: **Okamoto K:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Dopkins N:** Data Curation, Formal Analysis, Investigation, Resources, Software, Writing – Review & Editing; **Kinfu E:** Data Curation, Formal Analysis, Investigation, Software, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This study was funded by a University of St. Thomas CAS Startup Fund.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Okamoto K *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Okamoto K, Dopkins N and Kinfu E. **A draft genome sequence of the common, or spectacled caiman *Caiman crocodilus* [version 1; peer review: 2 not approved]** F1000Research 2021, **10**:1230 <https://doi.org/10.12688/f1000research.73066.1>

First published: 02 Dec 2021, **10**:1230 <https://doi.org/10.12688/f1000research.73066.1>

Introduction

The common, or spectacled, caiman, *Caiman crocodilus*, is one of the most widely distributed and abundant crocodilian species, ranging continuously from Mexico to Argentina (Busack and Pandya 2001; US Fish and Wildlife Service 2018). A generalist predator, *C. crocodilus* is remarkably adaptable, occupying a wide range of habitats from urban to seasonal savannahs to tropical rainforests (Medem 1981, 1983), and has recently been introduced to Cuba, Puerto Rico and Florida where it is considered an invasive species (US Fish and Wildlife Service 2018). The broad distribution and diversity of habitats has facilitated considerable intraspecific diversification within *C. crocodilus*; a recent analysis by Roberto et al. (2020) identified between seven and ten lineages within *C. crocodilus* across differing biogeographic regions and watersheds throughout Central and South America. Within-species diversity is also morphologically apparent, with skull shape in particular exhibiting systematic patterns of regional differentiation (Medem 1955; Gans 1980; Medem 1981, 1983; Ayarzagüena 1984; Escobedo-Galván et al. 2015). These intraspecific patterns of cranial shape variation within *C. crocodilus* have been shown to parallel patterns of interspecific cranial diversity found in extant crocodilians (Okamoto et al. 2015).

Additionally, *C. crocodilus* is a species of commercial importance, chiefly in the leather industry. While the hides of *C. crocodilus* contain osteoderms that render the manufacturing process more difficult than for other crocodilians, a majority of the approximately 1.5 million crocodilian skins traded globally come from *C. crocodilus* (Brazaitis et al. 1998; Caldwell 2015). As with other crocodilians, most legal hides come from commercial farming operations, and the market for caiman hides is estimated to be over US \$85 million (Caldwell 2015). Wild populations of *C. crocodilus* are also hunted for meat and even fishing bait (Da Silveira and Thorbjarnarson 1999; Brum et al. 2015; Pimenta et al. 2018) and provide ecosystem services including nutrient cycling and biological control (Valencia-Aguilar et al. 2013; Marley et al. 2019). Due to its role as an apex predator, *C. crocodilus* exhibits considerable bioaccumulation, with genotoxic analyses demonstrating molecular signatures of pollution on the *C. crocodilus* genome (Oliveira et al. 2021).

Thus, a draft genome sequence for *C. crocodilus* can not only assist with improved husbandry, ecotoxicology and wildlife management, but also has the potential to provide insight into evolutionary processes driving intraspecific diversification in continental systems more broadly.

Materials and methods

DNA was extracted from a tissue sample belonging to a single *Caiman crocodilus* museum specimen (UF-FLMNH 171438) using the DNeasy TM kit from Qiagen (Hilden, Germany). DNA was quantitated using ThermoFisher's (Waltham, MA, USA) Picogreen TM kit (for a final Picogreen concentration of 77.78 ng/ μ L). Tecan's (Männedorf, Switzerland) NuGEN Celero TM kit was then used to construct a paired-end library, which was subsequently sequenced on a single Illumina (San Diego, CA, USA) NovaSeq S4 lane. This yielded 239,911,946 paired-end reads of 2 $\times$ 150 bp each. Nucleic acid isolation, quantitation, library generation and raw-read sequencing were performed at the University of Minnesota Genomics Center.

The paired-end reads (as of October 2021, accessible at doi.org/10.5281/zenodo.5598241) were assembled *de novo* using the Iterative de Bruijn Graph Assembler (IDBA-UD; Peng et al. 2012). To assess the reliability of our pipeline from sequencing to *de novo* assembly using IDBA-UD, we repeated the sequencing and assembly using a museum-derived tissue sample from a single *Alligator mississippiensis* individual (UF-FLMNH 175565). This resulted in 249,325,204 paired-end reads of 2 $\times$ 150 bp each. As was the case for the *C. crocodilus* individual, the reads were then *de novo* assembled using IDBA-UD, and we used QUAST (Gurevich et al. 2013) to determine that the IDBA assembly of *A. mississippiensis* captured approximately 94.2% of a recently published *A. mississippiensis* assembly (GCA_000281125.4; Rice et al. 2017), with an NG50 of 21172 based on *de novo* assembled contigs alone.

We scaffolded the resulting draft *C. crocodilus* contigs using a two-step procedure. First, we scaffolded the caiman's contigs against a *Crocodylus porosus* assembly (GCF_001723895.1; Ghosh et al. 2020) using ragtag (Alonge et al. 2019). We then re-scaffolded the resulting contigs/scaffolds against the confamilial *Alligator mississippiensis* assembly (GCA_000281125.4), again using ragtag. The draft assembly was then submitted to the National Center for Biotechnology Information (NCBI).

Contaminants, mitochondrial DNA, vectors, adapters, and sequences shorter than 200 bp identified by NCBI were manually removed using seqkit (Shen et al. 2016) and custom scripts (available at [http://github.com/kewok/ncbi_scrubber](https://github.com/kewok/ncbi_scrubber)). As of July 2021, this draft assembly can be accessed at doi.org/10.5281/zenodo.4755063.

The resulting scaffold (10.5281/zenodo.4755063) was then masked using RepeatMasker (Smit et al. 2015) relying on the HMMER database (Finn et al. 2011) and with “alligator” specified as species. Finally, liftoff (Shumate and Salzberg 2020) was used to generate a draft annotation based on the masked assembly using the annotations associated with *A. mississippiensis* (GCA_000281125.4; Rice et al. 2017) as a reference.

table 2asn_gff (National Center for Biotechnology Information 2020) was used to generate a Sequin file (National Center for Biotechnology Information (US) 2014), and features flagged as errors were manually removed using custom scripts (available at https://github.com/kewok/ncbi_scrubber); as of July 2021 the draft annotation is available at 10.5281/zenodo.4755063).

Results and conclusions

Our assembly yielded a draft assembly of length 2,341,057,913 bp with 465,471 scaffolds and contigs, and an N50 of 70,464,410 bp (Proch::N50 - Telatin 2018). A total of 297,374 gene features were predicted.

Here we have described the first draft assembly and annotation of the *C. crocodilus* genome. We feel these data can assist natural resource management, ecotoxicology, agriculture, as well as research into broader questions about the interplay between microevolutionary and macroevolutionary processes across broad biogeographic scales.

Competing interests

The authors declare no competing interests.

Data availability

The draft assembly has been submitted to the National Center for Biotechnology Information and both assembly and annotation are currently available for review at (doi.org/10.5281/zenodo.4755063).

Acknowledgments

We are especially indebted to Dr. P. S. Soltis, T. A. Lott and the Genetic Resources Repository at the University of Florida - Florida Natural History Museum (UF-FLNHM) for generously providing us with tissue samples. We would like to thank the University of Minnesota Genomics Center (Minneapolis, MN, USA) for their guidance and for isolating DNA from museum samples, and for performing library preparation and raw sequencing. We wish to thank the Minnesota Supercomputing Institute (MSI) at the University of Minnesota and the Department of Chemistry at the University of St. Thomas for providing critical computational resources that contributed to the research results reported within this paper. Finally, we are very grateful to Dr. S. Pirro and Dr. M. Kieras at Iridian Genomes (Bethesda, MD, USA) for valuable insight on scaffolding the draft assemblies.

References

- Alonge M, Soyk S, Ramakrishnan S, et al.: **RaGOO: Fast and accurate reference-guided scaffolding of draft genomes.** *Genome Biol.* 2019; **20**: 1–17.
[Publisher Full Text](#)
- Ayarzagüena J: **Variaciones en la dieta de *Caiman sclerops*. La relación entre morfología bucal y dieta.** *Memoria De La Sociedad De Ciencias Naturales La Salle.* 1984; **44**: 123–140.
- Brazaitis P, Watanabe ME, Amato G: **The Caiman Trade.** *Sci. Am.* 1998; **278**: 70–76.
[Publisher Full Text](#)
- Brum SM, da Silva VM, Rossoni F, et al.: **Use of dolphins and caimans as bait for *Calophrys macropterus* (Lichtenstein, 1819) (Siluriforme: Pimelodidae) in the Amazon.** *J. Appl. Ichthyol.* 2015; **31**: 675–680.
[Publisher Full Text](#)
- Busack SD, Pandya S: **Geographic variation in *Caiman crocodilus* and *Caiman yacare* (Crocodylia: Alligatoridae): Systematic and legal implications.** *Herpetologica.* 2001; **57**: 294–312.
- Caldwell J: **World Trade in Crocodilian Skins 2013–2015. Technical report.** UN Environment Programme World Conservation Monitoring Centre; 2015.
- da Silveira R, Thorbjarnarson JB: **Conservation implications of commercial hunting of black and spectacled caiman in the Mamirauá Sustainable Development Reserve, Brazil.** *Biol. Conserv.* 1999; **88**: 103–109.
[Publisher Full Text](#)
- Escobedo-Galván AH, Velasco JA, González-Maya JF, et al.: **Morphometric analysis of the Rio Apaporis Caiman (Reptilia, Crocodylia, Alligatoridae).** *Zootaxa.* 2015; **4059**: 541–554.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Finn RD, Clements J, Eddy SR: **HMMER web server: Interactive sequence similarity searching.** *Nucleic Acids Res.* 2011; **39**: W29–W37.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Gans C: **Allometric Changes in the Skull and Brain of *Caiman crocodilus*.** *J. Herpetol.* 1980; **14**: 297–301.
[Publisher Full Text](#)
- Ghosh A, Johnson MG, Osmanski AB, 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**: 3635–3646.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Gurevich A, Saveliev V, Vyahhi N, et al.: **QUAST: Quality assessment tool for genome assemblies.** *Bioinformatics.* 2013; **29**: 1072–1075.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Marley G, Lawrence AJ, Phillip DA, et al.: **Mangrove and mudflat food webs are segregated across four trophic levels, yet connected by highly mobile top predators.** *Mar. Ecol. Prog. Ser.* 2019; **632**: 13–25.
[Publisher Full Text](#)
- Medem F: **A new subspecies of *Caiman sclerops* from Colombia.** *Fieldiana: Zoology.* 1955; **37**: 339–343.
[Publisher Full Text](#)

Medem F: *Los Crocodylia de Sur America Volumen I*. Bogotá, Colombia: Ministerio de Educación Nacional; 1981.

Medem F: *Los Crocodylia de Sur America Volumen II*. Bogotá, Colombia: Ministerio de Educación Nacional; 1983.

National Center for Biotechnology Information (US): **Submitting Sequences using Specific NCBI Submission Tools**. *The GenBank Submissions Handbook*. Bethesda, MD: National Center for Biotechnology Information (US); 2014; p. NBK566995

National Center for Biotechnology Information: **table 2asn_gff**. 2020.

Okamoto KW, Langerhans RB, Rashid R, *et al.*: **Microevolutionary patterns in the common caiman predict macroevolutionary trends across extant crocodilians**. *Biol. J. Linn. Soc.* 2015; **116**: 834–846. [Publisher Full Text](#)

Oliveira VCS, Viana PF, Gross MC, *et al.*: **Looking for genetic effects of polluted anthropized environments on *Caiman crocodilus crocodilus* (Reptilia, Crocodylia): A comparative genotoxic and chromosomal analysis**. *Ecotoxicol. Environ. Saf.* 2021; **209**: 111835. [PubMed Abstract](#) | [Publisher Full Text](#)

Peng Y, Leung HC, Yiu SM, *et al.*: **IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth**. *Bioinformatics*. 2012; **28**: 1420–1428. [PubMed Abstract](#) | [Publisher Full Text](#)

Pimenta NC, Barnett AA, Botero-Arias R, *et al.*: **When predators become prey: Community-based monitoring of caiman and dolphin hunting for the catfish fishery and the broader implications on Amazonian**

human-natural systems. *Biol. Conserv.* 2018; **222**: 154–163.

[Publisher Full Text](#)

Rice ES, Kohno S, St John J, *et al.*: **Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling**. *Genome Res.* 2017; **27**: 686–696. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Roberto IJ, Bittencourt PS, Muniz FL, *et al.*: **Unexpected but unsurprising lineage diversity within the most widespread Neotropical crocodilian genus *Caiman* (Crocodylia, Alligatoridae)**. *Syst. Biodivers.* 2020; **18**: 377–395. [Publisher Full Text](#)

Shen W, Le S, Li Y, *et al.*: **SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation**. *PLoS ONE*. 2016; **11**: e0163962. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Shumate A, Salzberg SL: **Liftoff: accurate mapping of gene annotations**. *Bioinformatics*. 2020. In press.

Smit A, Hubley R, Grenn P: **RepeatMasker Open-4.0**. 2015.

Telatin A: **Proch::N50**. 2018.

US Fish and Wildlife Service: *Common Caiman (Caiman crocodilus) Ecological Risk Screening Summary. Technical report*. US Fish and Wildlife Service; 2018.

Valencia-Aguilar A, Cortés-Gómez AM, Ruiz-Agudelo CA: **Ecosystem services provided by amphibians and reptiles in Neotropical ecosystems**. *Int. J. Biodivers. Sci. Ecosyst. Serv. Manag.* 2013; **9**: 257–272. [Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 28 April 2022

<https://doi.org/10.5256/f1000research.76689.r128753>

© 2022 Tollis M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Marc Tollis

School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, AZ, USA

A spectacled caiman genome is a welcome contribution to the field, however I believe some additional steps need to be taken before this manuscript can be acceptable for indexing. Fortunately, these steps are standard for most genome reports so the authors should have little problem addressing these issues.

The authors give a convincing rationale for sequencing the spectacled caiman genome - it is economically and ecologically important - but I believe they left out a key point. This is the first publicly available caiman genome, and the third member of the Alligatoridae (American and Chinese alligator genomes are available). There are relatively few other crocodilian genomes available (a crocodile and a gharial), so this assembly helps with the lack of crocodilian genomes and fills in an important taxonomical gap for vertebrate comparative genomics. I think the authors can make this point rather easily.

The sequencing and assembly steps are well described, as well as the annotation process, however the results of these methods are not well described and thus it is difficult to place the importance of this resource in further proper context. How does the contiguity in terms of N50 compare to other crocodilian (or reptile) genomes? A BUSCO analysis would help provide an estimate of the gene completeness of the genome, and is part of the standard battery of tests for new genome assemblies, but has not been done.

The authors report almost 300,000 gene features, but it is not elaborated upon what these features are. How many contain coding sequences? How many contain a Pfam domain? How many contain a reciprocal blast hit with another crocodilian genome? What is the average length of exons and introns, average number of exons per gene? Answers to these questions and a comparison to annotations from other genomes would be essential for understanding the usefulness of this annotation.

Finally, the authors need to provide the accession numbers or BioProject number that indicates

the submissions to NCBI. If the assembly has been submitted as stated in the manuscript, this should be relatively straightforward as the BioProject already exists or the accession number has been issued. The same goes for the raw reads, they should be deposited in the NCBI Short Read Archive and accession numbers should be reported. Zenodo is a good place to keep the assembly that was used for the analyses, but NCBI submission is standard for the field.

Are the rationale for sequencing the genome and the species significance clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Partly

Are sufficient details of the sequencing and extraction, software used, and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a usable and accessible format, and the assembly and annotation available in an appropriate subject-specific repository?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Comparative genomics, phylogenetics, vertebrates

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 24 December 2021

<https://doi.org/10.5256/f1000research.76689.r101899>

© 2021 Salzberg S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Steven L. Salzberg

¹ Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

² Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA

³ Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

⁴ Department of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA

This paper describes the assembly and annotation of a crocodile genome. It's a very brief note that provides some useful background about the species, *Caiman crocodilus*, but it needs some additional work before it would be acceptable for indexing. All of the following requests should be

quite easy to satisfy, but they are all essential.

First, the data are available on a Zenodo website, but that is unacceptable. All genomic data must be deposited in GenBank, EMBL, or DDBJ, and made available for any scientific publication. This practice has been near-universal for the past 25 years. The authors seem to know this, because they state that the assembly has been submitted to NCBI (the home of GenBank). However, the paper fails to provide GenBank accession numbers or a BioProject identifier. The authors need to get these identifiers and put them in the manuscript before the paper can be accepted. NCBI routinely provides such identifiers prior to publication.

Second, the raw reads are also at their Zenodo site, which again is not adequate. They need to deposit them either in SRA or ENA, and provide the accession numbers for those as well.

Third, because the paper does not provide anything other than a brief description of the methods used for assembly and annotation, it needs at least a cursory effort at quality evaluation. For the assembly, the authors could run Merqury to get an overall base-level quality value. That's a really minimal step. A more thorough step would be to run QUAST to estimate the number of mis-assemblies. This is especially important with an assembly based only upon short Illumina reads.

The other missing part is at least a short discussion of the annotation results. They mapped over the annotation from an alligator genome, which is fine, but they don't report any numbers other than the total number of "features," which is nearly 300,000. How many protein-coding genes did they annotate? How many non-coding RNAs? How many total protein-coding transcripts, and total transcripts of all types? These numbers are very easy to extract and the authors should at least comment on them.

Many of the references are garbled or nonsensical:

- For example, this is supposedly a reference about the term N50:

Telatin A: Proch::N50.2018.

I don't know what this means.

- Another example is this reference to the RepeatMasker software:

Smit A, Hubley R, Grenn PRepeatMasker Open-4.0.2015.

I can't figure out what that is, but the RepeatMasker program is well known and they should reference it properly.

- This reference refers to a piece of unpublished software from NCBI, apparently, and note that it misspells "Biotechnology" in the NCBI name:

National Center for Biotechnology Information: table 2asn_gff.2020.

- The reference to Shumate and Salzberg (2020) says "in press" but that paper was published in June 2021: <https://academic.oup.com/bioinformatics/article-abstract/37/12/1639/6035128>.

Are the rationale for sequencing the genome and the species significance clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Partly

Are sufficient details of the sequencing and extraction, software used, and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a usable and accessible format, and the assembly and annotation available in an appropriate subject-specific repository?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: genomics, bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research