DATA NOTE

The genome of the American groundhog, *Marmota monax* [version 1; peer review: 2 approved]

Daniela Puiu[1,2], Aleksey Zimin[1,2], Alaina Shumate[1,2], Yuchen Ge[1], Jiabin Qiu[3], Manoj Bhaskaran[3], Steven L. Salzberg[1,2,4]

[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, 21211, USA
[2]Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, 21211, USA
[3]Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, IN, USA
[4]Departments of Computer Science and Biostatistics, Johns Hopkins University, Baltimore, MD, USA

v1 **First published:** 16 Sep 2020, **9**:1137
https://doi.org/10.12688/f1000research.25970.1

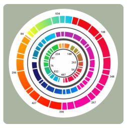
Latest published: 16 Sep 2020, **9**:1137
https://doi.org/10.12688/f1000research.25970.1

Abstract

We sequenced the genome of the North American groundhog, *Marmota monax*, also known as the woodchuck. Our sequencing strategy included a combination of short, high-quality Illumina reads plus long reads generated by both Pacific Biosciences and Oxford Nanopore instruments. Assembly of the combined data produced a genome of 2.74 Gbp in total length, with an N50 contig size of 1,094,236 bp. To annotate the genome, we mapped the genes from another *M. monax* genome and from the closely related Alpine marmot, *Marmota marmota*, onto our assembly, resulting in 20,559 annotated protein-coding genes and 28,135 transcripts. The genome assembly and annotation are available in GenBank under BioProject PRJNA587092.

Keywords

genome assembly, groundhog, woodchuck, genome annotation



This article is included in the Draft Genomes collection.

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 1 16 Sep 2020	✓ report	✓ report

1. **Markus Ralser**, The Francis Crick Institute, London, UK
   Toni Gossmann, Bielefeld University, Bielefeld, Germany
2. **Mihai Pop**, University of Maryland, College Park, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Steven L. Salzberg (salzberg@jhu.edu)

Author roles: Puiu D: Data Curation, Formal Analysis, Investigation, Methodology, Software; **Zimin A**: Data Curation, Formal Analysis, Investigation, Methodology, Software, Writing – Original Draft Preparation; **Shumate A**: Data Curation, Methodology, Software, Writing – Review & Editing; **Ge Y**: Data Curation, Methodology, Software; **Qiu J**: Data Curation, Methodology, Writing – Review & Editing; **Bhaskaran M**: Conceptualization, Project Administration, Writing – Review & Editing; **Salzberg SL**: Conceptualization, Funding Acquisition, Investigation, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Funding was provided in part by a grant from Eli Lilly and Company, and by National Institutes of Health grant R01-HG006677.

How to cite this article: Puiu D, Zimin A, Shumate A *et al.* **The genome of the American groundhog, *Marmota monax* [version 1; peer review: 2 approved]** F1000Research 2020, **9**:1137 https://doi.org/10.12688/f1000research.25970.1

First published: 16 Sep 2020, **9**:1137 https://doi.org/10.12688/f1000research.25970.1

Introduction

Groundhogs (*Marmota monax*), also known as woodchucks, belong to the same family of ground squirrels as the alpine marmot, *Marmota marmota*. Groundhogs are found throughout the eastern United States and across much of Canada. They are small, ground-dwelling rodents that weigh ~4 kg as adults.

The woodchuck is of interest to biomedical science as a model for Hepatitis B virus (HBV) infection in humans, due to endemic infections of woodchucks with woodchuck hepatitis virus (WHV), which is genetically similar to human HBV and causes a similar course of infection[1]. Unlike some animal models of hepatocellular carcinoma (HCC) that require immunocompromised animals, woodchucks can develop HCC spontaneously after WHV infection. This propensity makes the woodchuck a promising model of HBV-induced hepatocellular carcinoma in humans. This in turn motivated our efforts to sequence, assemble, and annotate its genome.

DNA isolation

DNA was collected from a healthy, wild-caught adult male woodchuck (WC2) captured in 2016 near Ithaca, New York by Northeastern Wildlife, Inc. The gDNA was isolated from the left medial lobe of the liver from animal WC2. All DNA used for sequencing came from the same animal.

Results

We generated 3.17 billion paired, 150-bp Illumina reads, for a total of 951 Gbp or approximately 390X genome coverage. We generated 32 million reads using Pacific Biosciences sequencing technology, of which 2.59 million were at least 10,000 bp long. The long PacBio reads contained 42.0 Gbp and had an N50 length of 16,554 bp. We also generated 6.4 million Oxford Nanopore (ONT) reads, of which 1.57 million were at least 10,000 bp long. The long ONT reads totaled 22.2 Gbp and had an N50 length of 13,815 bp. We then assembled the Illumina reads, the PacBio 10Kb+ reads, and the ONT 10Kb+ reads using MaSuRCA v3.2.7[2].

The resulting assembly, Woodchuck_1.0, consists of 8,860 contigs containing 2,737,034,741 bp, with an N50 contig size of 1,094,236. We compared our assembly to a recently published assembly of another woodchuck from the same species, GenBank accession GCA_901343595.1[3]. That assembly (MONAX5) was generated entirely from Illumina reads, and it has a total length of 2,552,052,516 bp in 48,534 scaffolds, with a scaffold N50 of 892 kb and a contig N50 of 74,495 bp. The earlier assembly is thus ~185 Mbp shorter than Woodchuck_1.0.

We aligned all contigs and scaffolds between the two assemblies, and found that 3791 scaffolds in MONAX5 were contained within longer contigs in Woodchuck_1.0, with an average identity of 99.24%. In contrast, only 84 contigs from Woodchuck_1.0 were contained in MONAX5 scaffolds, consistent with the much larger contig sizes in our assembly.

We mapped the annotation from MONAX5 to Woodchuck_1.0 using Liftoff[4]. To assign functions to the mapped transcripts, we aligned them to transcripts annotated in the Alpine marmot (*M. marmota*, GenBank accession GCA_001458135.1[5]. This yielded 20,559 protein-coding genes with 28,135 transcripts (including alternative splice variants). 10,664 of the genes were assigned functions based on near-identical matches with the Alpine marmot annotation, and the rest were labeled as hypothetical proteins. The average transcript contains 7.9 exons.

Data availability

Data from *Marmota monax* is available at NCBI under BioProject PRJNA587092, including the assembly with annotation at GenBank accession WJEC00000000, and the read data in the Sequence Read Archive under the same BioProject. The assembly and annotation are also available at ftp://ftp.ccb.jhu.edu/pub/data/Groundhog.

References

1. Menne S, Cote PJ: **The woodchuck as an animal model for pathogenesis and therapy of chronic hepatitis B virus infection.** *World J Gastroenterol.* 2007; **13**(1): 104–24.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

2. Zimin AV, Puiu D, Luo MC, *et al.*: **Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm.** *Genome Res.* 2017; **27**(5): 787–92.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

3. Alioto TS, Cruz F, Gomez-Garrido J, *et al.*: **The Genome Sequence of the Eastern Woodchuck (*Marmota monax*) - A Preclinical Animal Model for Chronic Hepatitis B.** *G3: Genes, Genomes, Genetics.* 2019; **9**(12): 3943–52.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

4. Shumate A, Salzberg SL: **Liftoff: an accurate gene annotation mapping tool.** *bioRxiv.* 2020.
Publisher Full Text

5. Gossmann TI, Shanmugasundram A, Borno S, *et al.*: **Ice-Age Climate Adaptations Trap the Alpine Marmot in a State of Low Genetic Diversity.** *Curr Biol.* 2019; **29**(10): 1712–1720.e7.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 1

Reviewer Report 20 November 2020

https://doi.org/10.5256/f1000research.28660.r74309

✓ **Mihai Pop**

Center for Bioinformatics and Computational Biology, Institute for Advanced Computer Sciences, University of Maryland, College Park, MD, USA

This manuscript describes the sequencing, assembly, and annotation of the groundhog genome. Overall the manuscript is clearly written and provides a good level of detail about the experimental setup and validation.

It would be beneficial for the readers if full details of the analytical pipeline (including specific parameters used and quality control/trimming steps) were made available either as supplementary material or through some other online platform (e.g. the authors' FTP server).

Given the substantial investment in the sequencing of the genome I am surprised that the project did not also generate RNA-seq data from the same animal. Since the rationale for this project is to support the use of the groundhog as a model organism for HCC, it's likely that animals are sacrificed during research making available a range of tissues from which a fairly comprehensive picture of the transcriptome can be gleaned.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Partly

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics, genome assembly.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response ( ) 25 Nov 2020

Steven Salzberg, Johns Hopkins University, Baltimore, USA

Response to reviewer 2: we produced the assembly using the default parameters for MaSuRCA 3.2.7, utilizing ~120x coverage from Illumina reads (a subset of the full data set) and all of the PacBio (15x) and Nanopore (8x) reads. Trimming and error correction steps are built into the MaSuRCA pipeline. The long reads were all labeled as the “NANOPORE” type for input to MaSuRCA. Our annotation process used the default settings for the Liftoff program.

Competing Interests: No competing interests were disclosed.

Reviewer Report 09 November 2020

https://doi.org/10.5256/f1000research.28660.r71457

✓ **Markus Ralser**
Molecular Biology of Metabolism Laboratory, The Francis Crick Institute, London, UK
Toni Gossmann
Bielefeld University, Bielefeld, Germany

The manuscript by Puiu *et al*. announces an improved, better annotated, and more complete genome of the American groundhog (*M. monax*), a ground squirrel species. The authors justify the importance of the genome as a model for Hepatitis infections, which sounds plausible. Obviously, such a genome is also important for other disciplines like evolutionary genetics, and as Marmots are exquisitely niche adapted species, a better understanding of their biology helps to understand problems of global importance, i.e. climate change, which warrants highlighting.

The paper is well written and technically of high quality. I have some questions concerning the relationship to the annotated *M. monax* genome deposited in GenBank (MONAX5), which the authors use as comparison of their genome quality. The authors conclude that their genome is more complete than the deposited genome. As the better quality is a key aspect of this submission, the manuscript would profit to disentangle the relative contribution of technical differences (e.g. due to novel sequencing technologies and assembly strategies) and biological differences (e.g. species variation) or what the relative contribution of each of the parts are.

Also, the improved genome annotation builds upon comparative mapping of the Alpine marmot and the deposited *M. monax* genome - meaning that “novel” regions in the new assembly are not covered in a similar quality (potential annotation bias). Also the genome assemblies’ DNA stems from the same individual, which means that the authors may want to include information on heterozygosity to conclude how representative the chosen individuals’ genome is of the species.

Minor remark: I do not think that a 4 kg rodent would be referred to as “small”; that perception may come from a comparison to some other marmots, but it may feel different in the context of other rodents.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Metabolism

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.