First genome sequence of the Gunnison’s prairie dog (*Cynomys gunnisoni*), a keystone species and player in the transmission of sylvatic plague

Mirian T.N. Tsuchiya¹, Rebecca B. Dikow¹, and Loren Cassin-Sackett²³*

1- Data Science Lab, Office of the Chief Information Officer, Smithsonian Institution, Washington DC

2- Department of Integrative Biology, University of South Florida, Tampa FL

3- Current affiliation: Department of Biology, University of Louisiana, Lafayette LA

*Author for Correspondence: Loren Cassin-Sackett, Department of Biology, University of Louisiana, Lafayette LA, USA, cassin.sackett@gmail.com

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
Data Deposition

The raw data and genome assembly are available on NCBI (BioProject PRJNA573923). VCF and GFF files are available on FigShare (DOI: 10.25573/data.c.4806264) along with a markdown document detailing each step of the post-assembly data processing.

Abstract

Prairie dogs (genus *Cynomys*) are a charismatic symbol of the American West. Their large social aggregations and complex vocalizations have been the subject of scientific and popular interest for decades. A large body of literature has documented their role as keystone species of western North America’s grasslands: They generate habitat for other vertebrates, increase nutrient availability for plants, and act as a food source for mammalian, squamate and avian predators. An additional keystone role lies in their extreme susceptibility to sylvatic plague (caused by *Yersinia pestis*), which results in periodic population extinctions, thereby generating spatio-temporal heterogeneity in both biotic communities and ecological processes. Here, we report the first *Cynomys* genome for a Gunnison’s prairie dog (*C. gunnisoni gunnisoni*) from Telluride, Colorado (USA). The genome was constructed using a hybrid assembly of PacBio and Illumina reads and assembled with MaSuRCA and PBJelly, which resulted in a scaffold N50 of 824 kb. Total genome size was 2.67 Gb, with 32.46% of the bases occurring in repeat regions. We recovered 94.9% (91% complete) of the single copy orthologs using the mammalian BUSCO database and detected 49,377 gene models (332,141 coding regions). PSMC showed support for long-term stable population size followed by a steady decline beginning near the end of the Pleistocene, as well as a recent
population reduction. The genome will aid in studies of chromosome evolution, disease resistance, and the genomic basis of life history traits in ground squirrels.

Key words
Biodiversity genomics, hybrid assembly, repeat evolution, ground squirrels, PSMC

Introduction

Recent years have seen the completion of large scale projects to sequence the genomes of divergent lineages across the tree of life, such as representatives from all neognath avian orders (Jarvis et al., 2014; Zhang et al., 2014), 24 divergent eutherian mammal orders (Lindblad-Toh et al., 2011), diverse squamate species (Tzika et al., 2015), and 159 spider species from diverse lineages (Fernández et al., 2018). Despite these advances, existing genomic resources can be characterized by underrepresentation of the most diverse families and orders. For instance, although they are the most diverse mammalian order—containing 40% of all mammalian species (2,561 out of 6,399 extant species, Burgin et al., 2018)—relatively few rodent genomes have been published (e.g., Kim et al. 2011; Couger et al. 2018; Thybert et al. 2018). For instance, the 84 Rodentia genomes available on GenBank represent <3.3% of the Order’s taxa, in comparison to 15.1% representation of Primates and 18.7% of Carnivora. Rodents are biologically diverse, and some possess medically relevant adaptations (e.g., resistance to cancer and reduced senescence (Buffenstein, 2008; Manov et al., 2013). Among mammals, they provide unparalleled ecological study systems due to the relative ease of catching, housing, and relocating these animals. Rodents vary widely in sociality, longevity, size and life history
traits. In addition, they are thought to be common sources of emerging diseases in humans (Han et al. 2015). Thus, the development of additional genomic resources for rodents would aid in evolutionary, ecological and epidemiological studies.

Some of the most widely-studied wild rodents are North America’s prairie dogs (Sciuridae, genus *Cynomys*). A charismatic emblem of the American frontier, prairie dogs were historically some of the most abundant animals in western grasslands (Merriam 1902). Their large population sizes, diurnal activity, and loud vocalizations have inspired decades of research on social behavior (Dobson et al., 2009; Haynie et al., 2003; Hoogland, 2013, 2001, 1999, 1998, 1981, 1979; Verdolin and Slobodchikoff, 2009), call complexity (Grady and Hoogland, 1986; Perla and Slobodchikoff, 1995; Placer and Slobodchikoff, 2004; Slobodchikoff et al., 1998; Slobodchikoff and Placer, 2006), and the ecosystem consequences of prairie dog activity (Coppock et al., 1983; Davidson et al., 2012; Detling and Whicker, 1987; Kotliar et al., 1999; Whicker and Detling, 1988). Prairie dogs are considered ‘ecosystem engineers’ (VanNimwegen et al. 2008) because their burrows provide shelter for amphibians, burrowing owls and other species (Augustine and Baker, 2013; Ceballos et al., 1999), and their burrow construction aerates the soil, bringing nutrients to the surface where they are available for plants (Coppock et al., 1983; Detling and Whicker, 1987; Whicker and Detling, 1988). The fate of endangered black-footed ferrets (*Mustela nigripes*) is inextricably tied to prairie dogs, as prairie dogs comprise >95% of their diet; prairie dogs are also important prey for golden eagles, ferruginous hawks, coyotes, snakes, and other animals (Davidson et al., 2012; Kotliar et al., 1999). As a result, species composition differs on prairie dog colonies, leading to increased beta diversity across the landscape (Bangert and Slobodchikoff, 2000; Smith and Lomolino, 2004).
In the past two centuries, prairie dogs have declined by 98% as a result of eradication campaigns—due to their public perception as pests (Roemer & Forrest 1996; Reading et al. 1999)—and sylvatic plague (caused by the bacterium *Yersinia pestis*). Plague was introduced to North America from Asia in the early 1900s (Eskey and Haas, 1939; Gage and Kosoy, 2005; Perry and Fetherston, 1997). Plague outbreaks cause 95-99% mortality in prairie dog populations (Cully and Williams, 2001; Cully et al., 1997; Sackett et al., 2013); however, there is increasing evidence from natural populations (Cully et al., 1997; Pauli et al., 2006; Sackett et al., 2013) and experimental studies (Busch et al., 2013; Rocke et al., 2012; Rocke et al., 2015) that resistance to plague may be evolving in at least two species of prairie dogs (*C. ludovicianus* and *C. gunnisoni*). Because the closest relative to have its genome sequenced (*Ictidomys tridecemlineatus*) diverged from *Cynomys* 4.67 [95% HPD 4.18–6.31] Mya (Upham et al. 2019), a reference genome for prairie dogs would aid in our understanding of the genetic basis of evolved resistance.

In summary, Gunnison’s prairie dogs are an important target for the development of a genome for several reasons: 1) They are ecologically important species in North American grasslands; 2) The species has been the object of intense study on life history, behavior, and the consequences of sociality for 4 decades and thus a genome should be of broad interest; 3) Elucidating the genomic basis of plague resistance is of both scientific and conservation interest for prairie dogs and associated species.

**Materials and Methods**

*Sample Preparation*
Several candidate individuals with low heterozygosity were chosen from available frozen DNA (Sackett et al., 2014) to facilitate genome assembly, and a low-heterozygosity individual (microsatellite $H_o = 0.182$) with a large amount of tissue was selected from a roadkill animal found near Telluride, CO (USA). Tissue was stored frozen in a DMSO-EDTA buffer until extraction. DNA was extracted primarily from ear tissue using the Qiagen DNeasy Blood & Tissue Kit, using 40 replicate extractions from the roadkill individual to ensure sufficient DNA. Each DNA aliquot was examined for size distribution on an agarose gel and for purity via Nanodrop and Qubit, and 20 μg of the highest-quality replicates were pooled. Libraries were prepared and samples were sequenced to 20x on a PacBio Sequel and 80x on an Illumina HiSeq 4000 (2x150 bp reads) at Duke University’s Sequencing and Genomic Technologies Shared Resource core facility.

**Genome Assembly and Variant Calling**

Genomes were constructed by a hybrid assembly of low-coverage PacBio long read (~mean 9.5 kbp) sequencing for generating scaffolds and high-coverage Illumina short read (150bp) sequencing for inferring the consensus sequence. We performed a hybrid de novo assembly using MaSuRCA (v. 3.2.1, Zimin et al. 2017) and additional scaffolding with SSPACE-longread (Boetzer & Pirovano 2014). Gaps were filled using PBJelly (English et al. 2012), and polishing was performed in Pilon (Walker et al. 2014). We used Kraken (Wood et al. 2019) to filter out scaffolds classified as bacteria and remove them from the final assembly (see Supplementary Material). We used Benchmarking Universal Single-Copy Orthologs (BUSCO v. 3.0.2, Simao et al. 2015) to assess the assembly completeness by comparing it to 4,104 orthologs from 50 species contained in the mammalia_odb9 gene
database (Zdobnov et al. 2017). We used Bowtie2 (Langmead & Salzberg 2012) to align the raw reads to the final assembly, and samtools v1.9 (Li et al. 2009) to generate a sorted bam file. Then, we removed PCR duplicates with picard-tools v2.5 (http://broadinstitute.github.io/picard/), and realigned indels and called variants using the GATK v4 (McKenna et al. 2010) following standard pipelines (e.g., DePristo et al. 2011; Cassin-Sackett et al. 2018).

To assemble the mitogenome, we imported the final whole genome assembly into Geneious Prime (Biomatters, 2019.1.3), and then mapped the scaffolds to the C. gunnisoni mitochondrial reference genome, available on Genbank (accession number MG450794, Streich et al. 2019).

Genome Structural Contents

We estimated genome-wide heterozygosity of the Gunnison’s prairie dog using jellyfish v2.3.0 (Marçais & Kingsford 2011) with both the default settings (removing kmers with coverage > 1,000x) and with the removal of kmers with coverage > 10,000x. Finally, we obtained the genome sequences of four high-quality ground squirrel genomes from Genbank (Marmota flaviventris, estimated 7.59 [95% HPD 6.40 –9.33] Million years divergence from Cynomys) (M. marmota (7.59 [95% HPD 6.40 –9.33] My divergence), Urocitellus parryi (5.66 [95% HPD 4.98–7.34] My divergence), and I. tridecemlineatus (4.67 [95% HPD 4.18–6.31] My divergence; Upham et al. 2019) and analyzed both repeat content and the relative proportion of CG sites (see Supplementary Materials) in each genome.
**Genome Annotation**

The genome was annotated using a multi-pronged approach that included repeat identification, a combination of *ab initio* and evidence-driven gene prediction using AUGUSTUS (v. 3.3.2; Stanke *et al.* 2006), and functional gene annotation using Blast2GO (Götz *et al.* 2008). First, we used RepeatMasker (open-4.0.6, Smit *et al.* 2013-2015) with the Rodentia database to identify repetitive elements in the genome, and soft-mask the assembly. Next, we generated a hints file for AUGUSTUS from two different lines of evidence: 1) alignment of the *Ictidomys tridecemlineatus* transcriptome (Hampton *et al.* 2011) to our assembly using BLAT (Kent 2002); and 2) conversion of the RepeatMasker .out to GFF (RepeatMasker script rmOutToGFF3.pl) and then GFF to hints (available at http://arthropods.eugenes.org/EvidentialGene/evigene/scripts/gff2hints.pl). AUGUSTUS training was performed during the BUSCO run using the --long flag. To speed up the analysis, we partitioned our assembly into scaffolds using the script partition_EVM_inputs.pl from EVM (Evidence Modeler, Haas *et al.* 2008). We ran AUGUSTUS in each scaffold individually, allowing genes to be predicted independently on both strands. We concatenated the results using the script join_aug_pred.pl, and extracted both the protein and nucleotide sequences of the gene models identified, as well as the individual CDS, using the AUGUSTUS script getAnnoFasta.pl. Finally, we used Blast2GO (v5.2.5, Gotz *et al.* 2008) to functionally annotate the genome. To do so, we ran Blast (v2.6.0+, Altschul *et al.* 1990) on the gene models identified by AUGUSTUS, and used the final .xml file as an input to Blast2GO.

We used Blobtools to assess the degree of microbial contamination in the *de novo* genome assembly. To do so, we subsetted the assembly into multiple fasta files and ran
blastn on each. Matches were categorized according to species at the lowest taxonomic level and according to phylum at the highest taxonomic level.

Demographic Inference

All species of prairie dogs are thought to have experienced drastic population declines in the past two centuries as a result of persecution and disease. To infer whether we could detect such changes in historical population size, we estimated the effective population size history using the Pairwise Sequentially Markovian Coalescent implemented in PSMC (Li & Durbin 2011). We generated the input file according to the recommendations of the author (described here [https://github.com/lh3/psmc](https://github.com/lh3/psmc)), and ran the analysis using the default settings, performing 100 bootstrap replicates. We scaled the PSMC plots assuming a mean generation time of two years, and compared two different mutation rates based on estimates from the literature: (1) 2.2 x10^{-9} per site per year (Kumar & Subramanian 2002), an estimated genome-wide rate for all mammals (“mammal rate”) and (2) 8.8 x10^{-10} per site per year (Nabholz et al. 2008), which is the estimated rate for a single nuclear gene (IRBP) in *Cynomys* (“*Cynomys* rate”).

**Results and Discussion**

**Genome Assembly and Variant Calling**

Long-read sequencing resulted in 52.5 GB of data from 14 PacBio SMRT cells, with an average read length of 9kb. The genome was estimated to be 2.67 Gb in length (Supplementary Table 1), similar to other rodents, particularly other ground squirrels (e.g., Accessions PRJNA399425, PRJNA516936, PRJNA477386). The assembly resulted in
15,346 contigs (with a contig N50 of 686,670bp) and 12,628 scaffolds (with a scaffold N50 of 824,613bp; Table S1). In comparison with other ground squirrel genomes available on GenBank, this assembly resulted in the second highest scaffold N50 and L50 (after *I. tridecemlineatus*) and the third fewest number of scaffolds (after *M. himalayana* and *I. tridecemlineatus*). Final coverage averaged 66x. We recovered 3,811 (91%) complete and 148 (3.6%) fragmented BUSCOs out of 4,104 mammalian orthologs searched (Figure 1). A single scaffold (~29kb) mapped to the reference mitochondrial genome (Streich et al. 2019) with 99.66% similarity. Variant calling produced a set of 2,336,054 SNPs.

**Genome Structural Contents**

Genome-wide heterozygosity was low, estimated at 0.315% under both kmer settings; this inference is consistent with previously estimated microsatellite heterozygosity (0.18;
Sackett et al. 2014). Repeat Masking indicated that 32.47% of the genome consisted of repetitive sequences, primarily LINEs (15.17%), SINEs (5.69%), and LTR elements (6.57%). Repeat content was nearly identical to four other ground squirrel species with divergence times to *Cynomys gunnisoni* ranging from 9.1 - 13.4 million years, both in terms of total repeat content and the proportion of each type of repeat (Figure 2, Table S5). In all five species, repeat sequences comprised approximately one third of the genome.

**Genome Annotation**

AUGUSTUS identified 332,141 coding DNA sequences/exons and a total of 49,377 gene models. The number of CDS identified for *C. gunnisoni* was within the range of those found for the other four ground squirrel species, which varied from 324,927 for *M. marmota* to 463,195 for *I. tridecemlineatus*. Out of the total number of gene models analyzed, ~1% (559) returned with Blast hits but without associated Gene Ontology entries. Blast2GO assigned functional labels to ~82% (40,255), with enzyme codes assigned to 17.32% (8,553) of the sequences (Figure S2).
Our assessment of contamination in Blobtools indicated that 92.02% of the Illumina reads mapped to the assembly were classified as Chordata while 0.63% of reads mapped to microbial taxa, including bacteria (Proteobacteria - 0.03%, Bacteroidetes - 0.05%), fungi (Ascomycota - 0.10%) and viruses (0.45%; Figure S3a). The remaining reads either had no blast hits (0.92%), or did not map to the assembly (6.41%). At the lowest taxonomic level, 85.53% of reads mapped to ground squirrels and 5.11% to Hominidae (4.79% human and 0.32% to the genus Pan), likely a function of the completeness of the blast database, which contains more complete human than squirrel sequences. Two microbial taxa present in the assembly were identified to genus: Pseudogymnoascus (0.09%) and Orthohepadnavirus (0.44%; Figure S3b). Pseudogymnoascus are a genus of fungi typically found in soil and rotting wood; thus, it is likely that this taxon is a contaminant present on the substrate on which the prairie dog was collected that was isolated along with the specimen. Orthohepadnavirus is a genus of viruses naturally hosted by humans and other mammals.

**Demographic Inference**

PSMC showed support for long-term stable population size followed by a steady decline beginning during the late Pleistocene and continuing into the present (Figure 1). Using the *Cynomys* rate, population decline occurred from ~127k – 13k years ago (ya), and with the mammal rate, populations declined from ~51k – 9k ya. This time period corresponds approximately to increased glaciation experienced across the planet beginning ~115kya (potentially causing population declines). Under the *Cynomys* rate scenario, population size recovered slightly around 8kya (a smaller recovery was inferred with the mammal rate at 3kya), a time marked by the widespread expansion of grasslands across North America,
which facilitated grassland specialists (Wisely et al. 2008, Oh et al. 2019) such as prairie dogs. This small increase in effective size may also correspond to divergence (Li & Durbin 2011; Cahill et al. 2016) between subspecies of Gunnison’s prairie dogs. Although the exact magnitude of effective population size inferred by using the genome of a low-heterozygosity individual may not be exact throughout all historical time periods, the patterns (i.e., shape of the curve) of changing population size should be robust to genome-wide heterozygosity levels (Li & Durbin 2011).

The assembly and annotation of the Gunnison’s prairie dog genome will facilitate future study on the genetic basis of social (Wilson-Henjum et al. 2019) and mating behavior (Hoogland et al. 2019), disease resistance (Busch et al. 2011, 2013), divergence and introgression (Sackett et al. 2014), coevolution (Holding et al. 2016), hibernation ecology (Lane et al. 2011, 2012), landscape genetics (Anderson et al. 2015; Kierepka & Latch 2016), phylogeography (Castellanos-Morales et al. 2016), keystone roles (Lindtner et al. 2018) and genomic variation in ground squirrels (Gedeon et al., 2017). A deeper understanding of genomic variation will enable scientists to inform management of threatened and endangered species, for instance, by lending insight into the optimal degree of gene flow among populations in the presence of disease (Sackett et al. 2013), or by identifying populations with ‘resistance’ alleles or high genetic diversity as potential sources for the reintroduction of diversity (Venesky et al. 2012; Strauss et al. 2017).

Acknowledgments

We are grateful to Ramona and Kent Gaylord for providing the prairie dog specimen that was used to obtain the genome sequence, and to Erin Arnold and Jeanette Calarco for
preparing the sample for sequencing. This work was supported by startup funds to LCS by the University of South Florida, and by a crowdfunding campaign through Instrumentl. Computing was conducted on the Smithsonian Institution High Performance Cluster (SI/HPC), Hydra.

References


Relative of House Mice with Unique Social and Ecological Adaptations. Genes, Genomes, Genetics, 8, 2145–2152.


20152841.


Streich SP, Keepers KG, Griffin KA, Kane NC, Martin AP (2019) The complete mitochondrial genome of Gunnison’s prairie dog subspecies (Cynomys gunnisoni gunnisoni) and phylogenetic relationship within the genus Cynomys. Mitochondrial DNA Part B: Resources, 4, 397–398.

Thybert D, Roller M, Navarro FCP et al. (2018) Repeat associated mechanisms of genome
evolution and function revealed by the Mus caroli and Mus pahari genomes. *Genome Research*, 28, 448–459.


Wilson-Henjum GE, et al. (2019) Alarm call modification by prairie dogs in the presence of


**Figure legends**

Fig. 1: A) Assembly statistic visualization ([https://github.com/rjchallis/assembly-stats](https://github.com/rjchallis/assembly-stats)), showing the genome N50 (dark orange), N90 (light orange), base composition (percentage of GC in dark blue, AT in light blue, and N in light grey), and BUSCO results (top right, in shades of green). B) Pairwise Sequentially Markovian Coalescent (PSMC) reconstruction of population size estimates over time, estimated using generation time of two years (g=2) and two mutation rates: $\mu = 2.2 \times 10^{-9}$ (green; “mammal rate”), and $\mu = 8.8 \times 10^{-10}$ (orange; “*Cynomys* rate”). Shaded lines correspond to 100 bootstrap estimates. The $\Delta$Temp ($^\circ$C) was calculated using benthic d18O records (Lisiecki and Raymo 2005), and extrapolated using the formula from Epstein et al. (1953). C) Map depicting the species distribution of *C. gunnisoni* (blue) in the western United States, with a star denoting the location where the sample was collected (Sackett et al. 2014). D) Image of *C. gunnisoni* (LCS).

Fig. 2: Percent repeat content (repeat classes, left; repeat subclasses, right) in ground squirrel genomes. Top to bottom, and pictured left to right: *Ictidomys tridecemlineatus*, *Marmota flaviventris*, *M. marmota*, *Urocitellus parryii*, *Cynomys gunnisoni*. *M. flaviventris* and *C.
gunnisoni images copyright Loren Cassin-Sackett; others publicly available from Wiki Commons.