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-Sackett 2,3,*

1Data Science Lab, Office of the Chief Information Officer, Smithsonian Institution, Washington, District of Columbia
2Department of Integrative Biology, University of South Florida, Tampa, FL
3Present address: Department of Biology, University of Louisiana, Lafayette, LA

*Corresponding author: E-mail: cassin.sackett@gmail.com.

Accepted: April 5, 2020

Data deposition: The raw data and genome assembly have been deposited at NCBI under BioProject PRJNA573923. The VCF and GFF files, along with a markdown document detailing each step of the postassembly data processing, have been deposited at FigShare under doi:10.25573/data.c.4806264.

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 spatiotemporal 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 Benchmarking Universal Single-Copy Orthologs database and detected 49,377 gene models (332,141 coding regions). Pairwise Sequentially Markovian Coalescent 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 mammalian 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 (Tzaika 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% 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 (Hoogland 1979, 1981, 1998, 1999, 2001, 2013; Haynie et al. 2003; Dobson et al. 1998; Verdoolin and Slobodchikoff 2009), call complexity (Grady and Hoogland 1986; Perla and Slobodchikoff 2002; Slobodchikoff et al. 1998; Placer and Slobodchikoff 2004; Slobodchikoff and Placer 2006), and the ecosystem consequences of prairie dog activity (Coppock et al. 1983; Detting and Whicker 1987; Whicker and Detting 1988; Kotliar et al. 1999; Davidson et al. 2012). Prairie dogs are considered “ecosystem engineers” (Van Nimwegen et al. 2008) because their burrows provide shelter for amphibians, burrowing owls, and other species (Ceballos et al. 1999; Augustine and Baker 2013), and their burrow construction aerates the soil, bringing nutrients to the surface where they are available for plants (Coppock et al. 1983; Detting and Whicker 1987; Whicker and Detting 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 (Kotliar et al. 1999; Davidson et al. 2012). 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 (Burnett and McCampbell 1926; Roemer and 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; Perry and Fetherston 1997; Gage and Kosoy 2005). Plague outbreaks cause 95–99% mortality in prairie dog populations (Cully et al. 1997; Cully and Williams 2001; 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 (Rocke et al. 2012, 2015; Busch et al. 2013) that resistance to plague may be evolving in at least two species of prairie dogs (Cynomys ludovicianus and Cynomys gunnisoni). Because the closest relative to have its genome sequenced (Ictidomys tridecemlineatus) diverged from Cynomys 4.67 (95% highest posterior density (HPD) 4.18–6.31) Ma (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 decades and thus a genome should be of broad interest; and 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 H9 = 0.182) with a large amount of tissue was selected from a roadkill animal found near Telluride, CO. Tissue was stored frozen in a dimethyl sulfoxide–ethylenediaminetetraacetic acid buffer until extraction. DNA was extracted primarily from ear tissue using the QiaNeasy 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 using Nanodrop and Qubit, and 20 μg of the highest-quality replicates were pooled. Libraries were prepared and samples were sequenced to 20× on a PacBio Sequel and 80× on an Illumina HiSeq 4000 (2 × 150-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 kb) sequencing for generating scaffolds and high-coverage Illumina short read (150 bp) sequencing for inferring the consensus sequence. We performed a hybrid de novo assembly using MaSuRA (v. 3.2.1, Zimin et al. 2017) and additional scaffolding with SSPACE-LongRead (Boetzer and 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 online). 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 and 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 polymerase chain reaction
duplicates with picard-tools v2.5 (http://broadinstitute.github.io/picard/, last accessed December 28, 2019) 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. 2019).

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 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 and Kingsford 2011) with both the default settings (removing kmers with coverage > 1,000×) and with the removal of kmers with coverage > 10,000×. Finally, we obtained the genome sequences of four high-quality ground squirrel genomes from GenBank (Marmota flaviventris, estimated coverage 7.59 [95% HPD 6.40–9.33] Myr divergence from Cynomys; M. marmota, 7.59 [95% HPD 6.40–9.33] Myr divergence; Urocitellus parryi, 5.66 [95% HPD 4.98–7.34] Myr divergence; and I. tridecemlineatus, 4.67 [95% HPD 4.18–6.31] Myr divergence; Upham et al. 2019) and analyzed both repeat content and the relative proportion of CG sites (see Supplementary Material online) in each genome.

**Genome Annotation**

The genome was annotated using a multipronged 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 removal of kmers from GenBank (Marmota flaviventris, estimated divergence 7.59 [95% HPD 6.40–9.33] Myr divergence from Cynomys; M. marmota, 7.59 [95% HPD 6.40–9.33] Myr divergence; Urocitellus parryi, 5.66 [95% HPD 4.98–7.34] Myr divergence; and I. tridecemlineatus, 4.67 [95% HPD 4.18–6.31] Myr divergence; Upham et al. 2019) and analyzed both repeat content and the relative proportion of CG sites (see Supplementary Material online) in each genome.

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 gunnisoni mitochondrial reference genome, available on GenBank (accession number MG450794, Streich et al. 2019).

**Results and Discussion**

Long-read sequencing resulted in 52.5 GB of data from 14 PacBio SMRT cells, with an average read length of 9 kb. The genome was estimated to be 2.67 Gb in length (supplementary table S1, Supplementary Material online), similar to other rodents, particularly other ground squirrels (e.g., Accessions PRJNA399425, PRJNAS158936, and PRJNA477386). The assembly resulted in 15,346 contigs (with a contig N50 of 824,613 bp; supplementary table S1, Supplementary Material online). In comparison with other ground squirrel genomes available on GenBank, this assembly resulted in the highest 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 Pairwise Sequentially Markovian Coalescent (PSMC) (Li and Durbin 2011). We generated the input file according to the recommendations of the author (described here https://github.com/lh3/psmc, last accessed December 28, 2019) and ran the analysis using the default settings, performing 100 bootstrap replicates. We scaled the PSMC plots assuming a mean generation time of 2 years and compared two different mutation rates based on estimates from the literature: 1) 2.2 × 10⁻⁹ per site per year (Kumar and Subramanian 2002), an estimated genome-wide rate for all mammals (“mammal rate”) and 2) 8.8 × 10⁻¹⁰ per site per year (Nabholz et al. 2008), which is the estimated rate for a single nuclear gene (IRBP) in Cynomys (“Cynomys rate”).
We averaged 66x. We recovered 3,811 (91%) complete and 148 (3.6%) fragmented BUSCOs out of 4,104 mammalian orthologs searched (fig. 1). A single scaffold (~29 kb) mapped to the reference mitochondrial genome (Streich et al. 2019) with 99.66% similarity. Variant calling produced a set of 2,336,054 single-nucleotide polymorphisms.

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 C. gunnisoni ranging from 9.1–13.4 Myr, both in terms of total repeat content and the proportion of each type of repeat (fig. 2 and supplementary table S5, Supplementary Material online). 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 coding sequences 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 (supplementary fig. S2, Supplementary Material online).

Our assessment of contamination in Blobtools indicated that 92.02% of the Illumina reads mapped to the assembly were classified as Chordata, whereas 0.63% of reads mapped to microbial taxa, including bacteria (Proteobacteria, 0.03% and Bacteroidetes, 0.05%), fungi (Ascomycota, 0.10%) and viruses (0.45%; supplementary fig. S3a, Supplementary Material online). 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%) (supplementary fig. S3b, Supplementary Material online). *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 (fig. 1). Using the *Cynomys* rate, population decline occurred from ~127 to 13 thousand years ago (ka), and with the mammal rate, populations declined from ~51 to 9 ka. This time period corresponds approximately to increased glaciation experienced across the planet beginning ~115 ka (potentially causing population declines). Under the Cynomys rate scenario, population size recovered slightly around 8 ka (a smaller recovery was inferred with the mammal rate at 3 ka), 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 and Durbin 2011; Cahlil 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 and 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 and 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 (Venesy et al. 2012; Strauss et al. 2017).

**Supplementary Material**

Supplementary data are available at Genome Biology and Evolution online.

**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 L.C.-S. by the University of South Florida and by a crowdfunding campaign through Instrumentl. Computing was conducted by the Smithsonian Institution High Performance Cluster (SI/HPC), Smithsonian Institution. https://doi.org/10.25572/SIHPC

Literature Cited


Associate editor: Mar Alba