The draft genome sequence and annotation of the desert woodrat Neotoma lepida

Michael Campbell a, Kelly F. Oakeson b,⁎, Mark Yandell c, James R. Halpert d, Denise Dearing b

a Cold Spring Harbor Laboratory, 1 Bungtown Rd, Cold Spring Harbor, NY 11724, USA
b Department of Biology, University of Utah, 257 South 1400 East, Salt Lake City, UT 84112, USA
c Department of Human Genetics, University of Utah, 15 North 2030 East, Salt Lake City, UT 84112, USA
d School of Pharmacy, University of Connecticut, 69 N Eagleville Rd Storrs, CT 06269, USA

Abstract

We present the de novo draft genome sequence for a vertebrate mammalian herbivore, the desert woodrat (Neotoma lepida). This species is of ecological and evolutionary interest with respect to ingestion, microbial detoxification and hepatic metabolism of toxic plant secondary compounds from the highly toxic creosote bush (Larrea tridentata) and the juniper shrub (Juniperus monosperma). The draft genome sequence and annotation have been deposited at GenBank under the accession LZPO01000000.

Keywords: Mammalian herbivore, Desert woodrat, Microbial detoxification, Toxic plant secondary compounds, Draft genome.

1. Direct link to deposited data


2. Sequencing and quality trimming

Paired end libraries were prepared with a 200 bp insert size using the Illumina TruSeq DNA PCR-Free Sample Preparation Kit (Illumina, Inc., San Diego, CA). Matepair libraries with 3 kb, 5 kb, and 10 kb insert sizes were prepared using the Illumina Nextera Mate Pair Sample Preparation Kit with some modifications (Illumina, Inc., San Diego, CA). The 200 bp and 3 kb libraries were sequenced utilizing the Illumina v4 chemistry generating 125 bp reads with two sequencing lanes dedicated to each library. The remaining libraries were sequenced using illumina v3 chemistry generating 101 bp reads. These libraries were barcoded and multiplexed on a single sequencing lane. Reads were trimmed for quality at a cutoff of phred 30 and remaining sequencing adapter fragments were removed using SeqyClean [1]. Sequencing output is summarized in Table 1.

3. Genome assembly

The cleaned genomic reads were assembled with the ALLPATHS assembler using default parameters [2]. Summary statistics of the assembled genome are reported in Table 2.

4. Transcriptome sequencing and assembly

Total RNA was isolated from frozen liver tissue samples of Neotoma lepida with a Qiagen RNeasy kit according to manufacturer’s instructions (Qiagen, Valencia, CA) and used to construct strand specific paired end sequencing libraries using the Illumina TruSeq Stranded mRNA sample Preparation Kit (Illumina, Inc., San Diego, CA). Libraries were then multiplexed together and sequenced on a single lane of the Illumina HiSeq platform, which generated 83,456,961 total paired-end reads.
reads of 101 bp in length. Paired-end reads were quality filtered and trimmed using Trimmomatic [3]. Quality filtered reads were then de novo assembled using Trinity [4].

5. Protein coding gene annotation

We assessed the completeness of gene space in the assembly using CEGMA [5]. 98.39% of the core eukaryotic genes were identifiable in the genome with 92.34% identified as complete. To annotate the whole genome, MAKER version 3.1 was run on Neotoma lepida using Trinity assembled miRNA-seq reads (described above), and all annotated mouse and rat proteins available from NCBI (ftp://ftp.ncbi.nih.gov/genomes/). Known rodent repetitive elements in RepBase [6] were masked using RepeatMasker [7]. Additional masking was done using a library of known transposable element protein product provided by MAKER [8]. Genes were predicted using SNAP and Augustus trained for Neotoma lepida using MAKER in an iterative fashion as described previously [8,9].

The final annotation set consisted of all MAKER generated annotations with protein or mRNA-seq support, and the subset of unsupported gene predictions that contained one or more protein family domains as detected by IPRscan and is described as the MAKER standard build [9, 10]. This annotation contained 24,574 protein coding genes, 75% of which contained a protein domain as detected by IPRscan, and 83% have an annotation edit distance <0.5 (consistent with a reasonably well annotated genome [11]). 95% of the annotated genes have similarity to proteins in SwissProt as identified by BLAST [12] (E < 0.000001).

The median gene length is 9324 bp with median exon and intron lengths of 130 bp and 1020 bp respectively. The average gene length is 19,733 bp. The high gene count and preponderance of short genes in the annotation suggests that many of the genes in the assembly are split between scaffolds. This result is in contrast with the CEGMA results.

However, the conserved core eukaryotic genes CEGMA uses are short and more likely to be found in full length in a fragmented genome assembly thereby providing an upper limit of complete genes in the assembly.

6. Nucleotide sequence accession number

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession LZPO00000000. The version described in this paper is version LZPO01000000.

Conflict of interest

The authors declare that there is no conflict of interests with respect to the work published in this paper.

Acknowledgements

The study was supported by grants from the National Science Foundation (DEB 1342615 & IOS 1256383).

References