Genomics Data 9 (2016) 58–59

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

Data in Brief

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

ARTICLE INFO

Article history:
Received 13 June 2016
Accepted 18 June 2016
Available online 23 June 2016

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

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.

Specifications	
Organism	*Neotoma lepida*
Sex	Male
Sequencer or array type	Illumina HiSeq
Data format	Processed
Experimental factors	Genomic DNA isolated from liver tissue of *N. lepida*
Experimental features	Whole genome sequence of *N. lepida*, assembly and annotation
Consent	Citation
Sample source location	Mojave Desert habitat in Lytle Ranch, Washington Co., UT

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/nuccore/LZPO000000000.

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

* Corresponding author.

http://dx.doi.org/10.1016/j.gdata.2016.06.008

Table 1
Sequencing output.

Insert size (bp)	Read length (bp)	Number of raw reads	Number of quality trimmed reads	Approximate sequencing depth from cleaned reads (based on assembly)	Approximate number of bases sequenced
200	125	1,165,155,028	395,573,529	21.05	49,446,691,125
3000	125	1,037,921,774	382,624,225	20.36	47,828,028,125
5000	101	292,034,662	92,962,239	4.00	9,389,186,139
10,000	101	201,658,776	65,857,602	2.83	6,651,617,802

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 mRNA-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 the 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.

Table 2
Assembly statistic.

Contig minimum size for reporting	1000
Number of contigs	390,383
Number of contigs per Mb	166.2
Number of scaffolds	119,373
Total contig length	2,038,610,551
Total scaffold length, with gaps	2,349,296,578
N50 contig size in kb [N50_contig]	9.1
N50 scaffold size in kb [N50_scaffold]	137
N50 scaffold size in kb, with gaps	151
Number of scaffolds per Mb	50.81
Median size of gaps in scaffolds	681
Median dev of gaps in scaffolds	39
% of bases in captured gaps	12.68
% of bases in negative gaps (after 5 devs)	0.06
%% of ambiguous bases	105.84
Ambiguities per 10,000 bases	26

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

[1] I.Y. Zhbannikov, S.S. Hunter, M.L. Settles, Zhbannikov: SeqyClean User Manual - Google Scholar. 2013.
[2] S. Gnerre, I. Maccallum, D. Przybylski, F.J. Ribeiro, J.N. Burton, B.J. Walker, et al., High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 1513–1518, http://dx.doi.org/10.1073/pnas.1017351108.
[3] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30 (2014) 2114–2120, http://dx.doi.org/10.1093/bioinformatics/btu170.
[4] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29 (2011) 644–652, http://dx.doi.org/10.1038/nbt.1883.
[5] G. Parra, K. Bradnam, I. Korf, CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23 (2007) 1061–1067, http://dx.doi.org/10.1093/bioinformatics/btm071.
[6] W. Bao, K.K. Kojima, O. Kohany, Repbase update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6 (2015) 1, http://dx.doi.org/10.1186/s13100-015-0041-9.
[7] A. Smit, R. Hubley, P. Green, 2010 RepeatMasker Open-3.0, URL: http://www.repeatmasker.org, 1996.
[8] B.L. Cantarel, I. Korf, S.M.C. Robb, G. Parra, E. Ross, B. Moore, et al., MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18 (2008) 188–196, http://dx.doi.org/10.1101/gr.6743907.
[9] M.S. Campbell, C. Holt, B. Moore, M. Yandell, Genome Annotation and Curation Using MAKER and MAKER-P. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2014 http://dx.doi.org/10.1002/0471250953.bi0411s48.
[10] M.S. Campbell, M. Law, C. Holt, J.C. Stein, G.D. Moghe, D.E. Hufnagel, et al., MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 164 (2014) 513–524, http://dx.doi.org/10.1104/pp.113.230144.
[11] C. Holt, M. Yandell, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12 (2011) 1 http://dx.doi.org/10.1186/1471-2105-12-491.
[12] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool. J. Mol. Biol. 215 (1990) 403–410, http://dx.doi.org/10.1016/S0022-2836(05)80360-2.