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.

© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
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 all MAKER generated annotated genes 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

[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] 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 organisms genomes. Genome Res. 18 (2008) 188–196, http://dx.doi.org/10.1101/gr.6743907.
[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 organisms 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/10.1002/0470059533.btm011635.
[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/10.1016/0022-2836(90)90300-2.