A draft genome assembly of the eastern banjo frog *Limnodynastes dumerilii dumerilii* (Anura: Limnodynastidae)

Qiye Li[1,2], Qunfei Guo[1,3], Yang Zhou[1], Huishuang Tan[1,4], Terry Bertozzi[5,6], Yuanzhen Zhu[1,7], Ji Li[2,8], Stephen Donnellan[5], Guojie Zhang[2,8,9,10*]

[1] BGI-Shenzhen, Shenzhen 518083, China

[2] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China

[3] College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

[4] Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 611731, China

[5] South Australian Museum, North Terrace, Adelaide 5000, Australia

[6] School of Biological Sciences, University of Adelaide, North Terrace, Adelaide 5005, Australia

[7] School of Basic Medicine, Qingdao University, Qingdao 266071, China

[8] China National Genebank, BGI-Shenzhen, Shenzhen 518120, China

[9] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, 650223, Kunming, China

[10] Section for Ecology and Evolution, Department of Biology, University of Copenhagen, DK-2100 Copenhagen, Denmark

[*] Correspondence: guojie.zhang@bio.ku.dk (G.Z.).

e-mail addresses for all authors:

Qiye Li <liqiye@genomics.cn>, Qunfei Guo <guoqunfei@genomics.cn>, Yang Zhou <zhouyang@genomics.cn>, Huishuang Tan < tanhuishuang@genomics.cn>, Terry Bertozzi <Terry.Bertozzi@samuseum.sa.gov.au>, Yuanzhen Zhu <zhuyuanzhen@genomics.cn>, Ji Li <liji1@genomics.cn>, Stephen Donnellan <Steve.Donnellan@samuseum.sa.gov.au> and Guojie Zhang <guojie.zhang@bio.ku.dk>

ORCIDs:

Qiye Li: 0000-0002-5993-0312; Yang Zhou: 0000-0003-1247-5049; Terry Bertozzi: 0000-0001-6665-3395; Stephen Donnellan: 0000-0002-5448-3226; Guojie Zhang: 0000-0001-6860-1521

Abstract

Amphibian genomes are usually challenging to assemble due to large genome size and high repeat content. The Limnodynastidae is a family of frogs native to Australia, Tasmania and New Guinea. As an anuran lineage that successfully diversified on the Australian continent, it represents an important lineage in the amphibian tree of life but lacks reference genomes. Here we sequenced and annotated the genome of the eastern banjo frog *Limnodynastes dumerilii dumerilii* to fill this gap. The total length of the genome assembly is 2.38 Gb with a scaffold N50 of 285.9 kb. We identified 1.21 Gb of non-redundant sequences as repetitive elements and annotated 24,548 protein-coding genes in the assembly. BUSCO assessment indicated that more than 94% of the expected vertebrate genes were present in the genome assembly and the gene set. We anticipate that this annotated genome assembly will advance the future study of anuran phylogeny and amphibian genome evolution.

Introduction

The recent powerful advances in genome sequencing technology have allowed efficient decoding of the genomes of many species [1, 2]. So far, genome sequences are available publicly for more than one thousand species sampled across the animal branch of the tree of life. These genomic resources have provided vastly improved perspectives on our knowledge of the origin and evolutionary history of metazoans [3, 4], facilitated advances in agriculture [5], enhanced approaches for conservation of endangered species [6], and uncovered the genomic changes underlying the evolutionary successes of some clades such as birds [7] and insects [8]. However, amphibian genomes are still challenging to assemble due to their large genome sizes, high repeat content and sometimes high heterozygosity if specimens are collected from wild populations [9]. This also accounts for the scarcity of reference genomes for Anura (frogs and toads) — the most species-rich order of amphibians including many important models for developmental biology and environmental monitoring [10]. Specifically, despite the existence of more than 7,000 living species of Anura [11], only 10 species have their genomes sequenced and annotated to date [12-21], which cover only 8 out of the 54 anuran families. Moreover, genomes of Neobatrachia, which contains more than 95% of the anuran species [11], are particularly under-represented. Only 5 of the 10 publicly available anuran genomes belong to Neobatrachia [22]. This deficiency of neobatrachian genomes would undoubtedly restrict the study of the genetic basis underlying the great diversification of this amphibian lineage, and our understanding of the adaptive genomic changes that facilitate the aquatic to terrestrial transition of vertebrates and the numerous unique reproductive modes found in this clade.

As a candidate species proposed for genomic analysis by the Genome 10K (G10K) initiative [9], we sequenced and annotated the genome of the Australian banjo frog *Limnodynastes dumerilii* (also called the pobblebonk; NCBI:txid104065) to serve as a representative species of the neobatrachian family Limnodynastidae. This burrowing frog is endemic to Australia and named after its distinctive "bonk" call, which is likened to a banjo string being plucked. It mainly occurs along the southeast coast of Australia, from the coast of New South Wales, throughout Victoria and into the southwest corner of South Australia and Tasmania [23]. Five subspecies of *L. dumerilii* are recognized, including *Limnodynastes dumerilii dumerilii*, *L. dumerilii grayi*, *L. dumerilii fryi*, *L. dumerilii insularis* and *L. dumerilii variegata* [24]. The subspecies chosen for sequencing is the eastern banjo frog *L. dumerilii dumerilii* (NCBI:txid104066), as it is the most widespread among the five subspecies and forms hybrid zones with a number of the other subspecies [23]. We believe that the release of genomic resources from this neobatrachian frog will benefit the future studies of phylogenomics and comparative genomics of anurans, and also facilitate other research related to the evolutionary biology of *Limnodynastes*.

Methods

Sample collection, library construction and sequencing

Genomic DNA was extracted from the liver of an adult female *Limnodynastes dumerilii dumerilii* (Fig. 1) using the Gentra Puregene Tissue Kit (QIAGEN, Hilden, Germany) according to manufacturer's instructions with the following exceptions: following the DNA precipitation step, DNA was spooled onto a glass rod, washed twice in 70% ethanol and dried before dissolving in 100 ul of the recommended elution buffer [25]. The specimen was originally caught in River Torrens, Adelaide, South Australia, Australia, and is archived in the South Australian Museum (registration number: SAMAR66870).

A total of 211 Gb of sequence was generated from four short-insert libraries (170 bp × 1, 250 bp × 1, 500 bp × 1, and 800 bp × 1), and 185 Gb of sequence from ten mate-paired libraries (2 kb × 3, 5 kb × 3, 10 kb × 2, and 20 kb × 2). All the 14 libraries were subject to paired-end sequencing on the HiSeq 2000 platform following the manufacturer's instructions (Illumina, San Diego, CA, USA), using PE100 or PE150 chemistry for the short-insert libraries and PE49 for the mate-paired libraries [26] (Table 1). Low-quality reads, adapter-contaminated reads, and duplicated reads arising from polymerase chain reaction (PCR) amplification during library construction were removed by SOAPnuke (v1.5.3, RRID:SCR_015025) [27] prior to downstream analyses. This yielded a total of 180 Gb of clean sequence for genome assembly, which represents 71 times coverage of the estimated haploid genome size of *L. d. dumerilii* in terms of sequence depth, and 1,198 times in terms of physical depth (Table 1).

Genome size estimation and genome assembly

To obtain a robust estimation of the genome size of *L. d. dumerilii*, we conducted *k*-mer analysis with all of the clean sequence (130 Gb) from the four short-insert libraries using a range of *k* values (17, 19, 21, 23, 25, 27, 29 and 31). The *k*-mer frequencies were counted by Jellyfish (v2.2.6) [28] with the *-C* setting. The genome size of *L. d. dumerilii* was estimated to be around 2.54 Gb (Table 2), which was calculated as the number of effective *k*-mers (i.e. total *k*-mers – erroneous *k*-mers) divided by the homozygous peak depth following Cai *et al* [29]. It is noteworthy that, the presence of a distinct heterozygous peak, which displayed half of the depth of the homozygous peak in the *k*-mer frequency distribution, suggests that the diploid genome of this wild-caught individual has a high level of heterozygosity (Fig. 2). The rate of heterozygosity was estimated to be around 1.17% by GenomeScope (v1.0.0, RRID:SCR_017014) [30] (Table 2).

We then employed Platanus (v1.2.1, RRID:SCR_015531) [31] to assemble the genome of *L. d. dumerilii*. Briefly, all the clean sequence from the four short-insert libraries were first assembled into contigs using *platanus assemble* with parameters *-t 20 -k 29 -u 0.2 -d 0.6 -m 150*. Then paired-end reads from the four short-insert and ten mate-paired libraries were used to connect contigs into scaffolds by *platanus scaffold* with parameters *-t 20 -u 0.2 -l 3* and the insert size information of each library. Finally, *platanus gap_close* was employed to close intra-scaffold gaps using the paired-end reads from the four short-insert libraries with default settings. This Platanus assembly was further improved by Kgf (version 1.16) [9] followed by GapCloser (v1.10.1, RRID:SCR_015026) [9] for gap filling with the clean reads from the four short-insert libraries.

Repetitive element annotation

Both homology-based and *de novo* predictions were employed to identify repetitive elements in the *L. d. dumerilii* genome assembly [32]. For homology-based prediction, known repetitive elements were identified by aligning the *L. d. dumerilii* genome sequences against the Repbase-derived RepeatMasker libraries using RepeatMasker (v4.1.0, RRID:SCR_012954; setting *-nolow -norna -no_is*) [33], and against the transposable element protein database using RepeatProteinMask (an application within the RepeatMasker package; setting *-noLowSimple -pvalue 0.0001 -engine ncbi*). For *de novo* prediction, RepeatModeler (v2.0, RRID:SCR_015027) [34] was first executed on the *L. d. dumerilii* assembly to build a *de novo* repeat library for this species. Then RepeatMasker was employed to align the *L. d. dumerilii* genome sequences against the *de novo* library for repetitive element identification. Tandem repeats in the *L. d. dumerilii* genome assembly were identified by Tandem Repeats Finder (v4.09) [35] with parameters *Match=2 Mismatch=7 Delta=7 PM=80 PI=10 Minscore=50 MaxPeriod=2000*.

Protein-coding gene annotation

Similar to repetitive element annotation, both homology-based and *de novo* predictions were employed to build gene models for the *L. d. dumerilii* genome assembly [36]. For homology-based prediction, protein sequences from diverse vertebrate species, including *Danio rerio, Xenopus tropicalis*, *Xenopus laevis*, *Nanorana parkeri*, *Microcaecilia unicolor*, *Rhinatrema bivittatum*, *Anolis carolinensis*, *Gallus gallus* and *Homo sapiens*, were first aligned to the *L. d. dumerilii* genome assembly using TBLASTN (blast-2.2.26, RRID:SCR_011822) [37] with parameters *-F F -e 1e-5*. Then the genomic sequences of the candidate loci together with 5 kb flanking sequences were extracted for exon-intron structure determination, by aligning the homologous proteins to these extracted genomic sequences using GeneWise (wise-2.2.0, RRID:SCR_015054) [38]. For *de novo* prediction, we randomly picked 1,000 homology-derived gene models of *L. d. dumerilii* with complete open reading frames (ORFs) and reciprocal aligning rates exceeding 90% against the *X. tropicalis* proteins to train AUGUSTUS (v3.3.1, RRID:SCR_008417) [39]. The obtained gene parameters were then used by AUGUSTUS to predict protein-coding genes on the repeat-masked *L. d. dumerilii* genome assembly. Finally, gene models derived from the above two methods were combined into a non-redundant gene set using a similar strategy to Xiong *et al*. (2016) [40]. Genes showing BLASTP (blast-2.2.26, RRID:SCR_001010; parameters *-F F -e 1e-5*) hits to transposon proteins in the UniProtKB/Swiss-Prot database (v2019_11), or with more than 70% of their coding regions overlapping repetitive sequences, were removed from the combined gene set.

Results and Discussion

Assembly and annotation of the *L. d. dumerilii* genome

We assembled the nuclear genome of a female eastern banjo frog *L. d. dumerilii* (Fig. 1) with ~180 Gb (71X) clean Hiseq data from four short-insert libraries (170 bp × 1, 250 bp × 1, 500 bp × 1, and 800 bp × 1) and ten mate-paired libraries (2 kb × 3, 5 kb × 3, 10 kb × 2, and 20 kb × 2) (Table 1). The final genome assembly comprised 520,896 sequences with contig and scaffold N50s of 10.2 kb and 286.0 kb, respectively, and a total length of 2.38 Gb, which is close to the estimated genome size of 2.54 Gb by *k*-mer analysis (Table 2 and Fig. 2). There are 242 Mb of regions present as unclosed gaps (Ns), accounting for 10.2% of the assembly. The GC content of the *L. d. dumerilii* assembly excluding gaps was estimated to be 41.0%. The combination of homology-based and *de novo* prediction methods masked 1.21 Gb of non-redundant sequences as repetitive elements, accounting for 56.4 % of the *L. d. dumerilii* genome assembly excluding gaps (Table 3). We also obtained 24,548 protein-coding genes in the genome assembly, of which 67% had complete ORF. Functional annotation by searching the *L. d. dumerilii* proteins against public databases of UniProtKB/Swiss-Prot (v2019_11, RRID:SCR_004426) [41], NCBI nr (v20191030), and KEGG (v93.0, RRID:SCR_012773) [42] with BLASTP (blast-2.2.26; parameters *-F F -e 1e-5*) successfully annotated almost all of the *L. d. dumerilii* gene loci (Table 4).

Data validation and quality control

Two strategies were employed to estimate the completeness of the *L. d. dumerilii* genome assembly. First, all the clean reads from the short-insert libraries were aligned to the genome assembly using BWA-MEM (BWA, version 0.7.16, RRID:SCR_010910) with default parameters [43]. We observed that 99.6 % of reads could be mapped back to the assembled genome and 85.6 % of the inputted reads were mapped in proper pairs as accessed by samtools flagstat (SAMtools v1.7, RRID:SCR_002105), suggesting that most sequences of the *L. d. dumerilii* genome were present in the current assembly. Secondly, we assessed the *L. d. dumerilii* assembly with Benchmarking Universal Single-Copy Orthologs (BUSCO; v3.0.2, RRID:SCR_015008), a software package that can quantitatively measure genome assembly completeness based on evolutionarily informed expectations of gene content [44], and found that up to 94.7 % of the 2,586 expected vertebrate genes were present in the *L. d. dumerilii* assembly. Furthermore, 85.5% and 84.5 % of the expected genes were identified as complete and single-copy genes, respectively. This BUSCO assessment further highlighted the comprehensiveness of the current *L. d. dumerilii* genome assembly in terms of gene space.

We then evaluated the completeness of the *L. d. dumerilii* protein-coding gene set with BUSCO (v3.0.2) and DOGMA (v3.0, RRID:SCR_015060) [45], a program that measures the completeness of a given transcriptome or proteome based on a core set of conserved domain arrangements (CDAs). BUSCO analysis showed that 97.1 % of the expected vertebrate genes were present in the *L. d. dumerilii* protein-coding gene set with 88.5 % and 84.5% identified as complete and single-copy genes, respectively, close to that estimated for the genome assembly. Meanwhile, DOGMA analysis based on PfamScan Annotations (PfamScan v1.5; Pfam v32.0, RRID:SCR_015060) [46] and the eukaryotic core set identified 95.4 % of the expected CDAs in the annotated gene set. These results demonstrated the high completeness of the *L. d. dumerilii* protein-coding gene set.

Re-use potential

Here, we report a draft genome assembly of the eastern banjo frog *L. d. dumerilii*. It represents the first genome assembly from the family Limnodynastidae (Anura: Neobatrachia). Although the continuity of the assembly in terms of contig and scaffold N50s is modest, probably due to the high repeat content (56%) and heterozygosity (1.17%), the completeness of this draft assembly is demonstrated to be high according to read mapping and BUSCO assessment. Thus, it is suitable for phylogenomics and comparative genomics analyses with other available anuran genomes or phylogenomic datasets. In particular, the high-quality protein-coding gene set derived from the genome assembly will be useful for deducing orthologous relationships across anuran species or reconstructing the ancestral gene content of anurans. Due to evolutionary importance of *Limnodynastes* frogs in Australia, the genomic resources released in this study will also support further research on the biogeography of speciation, evolution of male advertisement calls, hybrid zone dynamics, and conservation of *Limnodynastes* frogs.

Availability of supporting data

The raw sequencing reads are deposited in NCBI under the BioProject accession PRJNA597531 and are also deposited in the CNGB Nucleotide Sequence Archive (CNSA) with accession number CNP0000818. Genome assembly, protein-coding gene and repeat annotations are deposited in the *GigaScience* GigaDB [47] and NCBI under accession number GCA_011038615.1.

List of abbreviations

BUSCO: Benchmarking Universal Single-Copy Orthologs; G10K: Genome 10K; NCBI: National Center for Biotechnology Information; PCR: Polymerase Chain Reaction; ORF: Open Reading Frame; KEGG: Kyoto Encyclopedia of Genes and Genomes; DOGMA: DOmain-based General Measure for transcriptome and proteome quality Assessment; CDA: Conserved Domain Arrangement; CNGB: China National GeneBank; CNSA: CNGB Sequence Archive.

Funding

This work was funded by the Strategic Priority Research Program of Chinese Academy of Sciences (No. XDB31020000), a National Key R&D Program of China (MOST) grant (No. 2018YFC1406901), the International Partnership Program of Chinese Academy of Sciences

(No. 152453KYSB20170002), a Carlsberg Foundation grant (No. CF16-0663) and the Villum Foundation (No. 25900).

Competing interests

The authors declare that they have no competing interests.

Author contributions

G.Z. and Q.L. conceived and supervised the study; T.B. and S.D. prepared the DNA samples; Y.Z. and Q.G. performed *k*-mer analysis and genome assembly; Q.G. and J.L. conducted assessment of assembly quality; H.T. performed protein-coding gene annotation; Y.Z. performed repeat annotation; G.Z. and S.D. contributed reagents/materials/analysis tools; Q.L. wrote the manuscript with the inputs from all authors. All authors read and approved the final manuscript.

References

1. Goodwin S, McPherson JD and McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature reviews Genetics. 2016;17 6:333-51. doi:10.1038/nrg.2016.49.

2. van Dijk EL, Jaszczyszyn Y, Naquin D and Thermes C. The Third Revolution in Sequencing Technology. Trends in genetics : TIG. 2018;34 9:666-81. doi:10.1016/j.tig.2018.05.008.

3. Sebe-Pedros A, Degnan BM and Ruiz-Trillo I. The origin of Metazoa: a unicellular perspective. Nature reviews Genetics. 2017;18 8:498-512. doi:10.1038/nrg.2017.21.

4. Laumer CE, Fernandez R, Lemer S, Combosch D, Kocot KM, Riesgo A, et al. Revisiting metazoan phylogeny with genomic sampling of all phyla. Proceedings Biological sciences / The Royal Society. 2019;286 1906:20190831. doi:10.1098/rspb.2019.0831.

5. Beiki H, Eveland AL and Tuggle CK. Recent advances in plant and animal genomics are taking agriculture to new heights. Genome Biol. 2018;19 1:48. doi:10.1186/s13059-018-1427-z.

6. Supple MA and Shapiro B. Conservation of biodiversity in the genomics era. Genome Biol. 2018;19 1:131. doi:10.1186/s13059-018-1520-3.

7. Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science. 2014;346 6215:1311-20. doi:10.1126/science.1251385.

8. Thomas GWC, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K, et al. Gene content evolution in the arthropods. Genome Biol. 2020;21 1:15. doi:10.1186/s13059-019-1925-7.

9. Koepfli KP, Paten B, Genome KCoS and O'Brien SJ. The Genome 10K Project: a way forward. Annu Rev Anim Biosci. 2015;3:57-111. doi:10.1146/annurev-animal-090414-014900.

10. Carroll R. The rise of amphibians: 365 million years of evolution. Johns Hopkins University Press. 2009.

11. AmphibiaWeb. <https://amphibiaweb.org> University of California, Berkeley, CA, USA. (Accessed 18 Feb 2020).

12. Li J, Yu H, Wang W, Fu C, Zhang W, Han F, et al. Genomic and transcriptomic insights into molecular basis of sexually dimorphic nuptial spines in *Leptobrachium leishanense*. Nat Commun. 2019;10 1:5551. doi:10.1038/s41467-019-13531-5.

13. Li Y, Ren Y, Zhang D, Jiang H, Wang Z, Li X, et al. Chromosome-level assembly of the mustache toad genome using third-generation DNA sequencing and Hi-C analysis. Gigascience. 2019;8 9 doi:10.1093/gigascience/giz114.

14. Seidl F, Levis NA, Schell R, Pfennig DW, Pfennig KS and Ehrenreich IM. Genome of *Spea multiplicata*, a Rapidly Developing, Phenotypically Plastic, and Desert-Adapted Spadefoot Toad. G3 (Bethesda). 2019;9 12:3909-19. doi:10.1534/g3.119.400705.

15. Edwards RJ, Tuipulotu DE, Amos TG, O'Meally D, Richardson MF, Russell TL, et al. Draft genome assembly of the invasive cane toad, *Rhinella marina*. Gigascience. 2018;7 9 doi:10.1093/gigascience/giy095.

16. Rogers RL, Zhou L, Chu C, Marquez R, Corl A, Linderoth T, et al. Genomic Takeover by Transposable Elements in the Strawberry Poison Frog. Mol Biol Evol. 2018;35 12:2913-27. doi:10.1093/molbev/msy185.

17. Denton RD, Kudra RS, Malcom JW, Du Preez L and Malone JH. The African Bullfrog (*Pyxicephalus adspersus*) genome unites the two ancestral ingredients for making vertebrate sex chromosomes. bioRxiv. 2018:329847.

18. Hammond SA, Warren RL, Vandervalk BP, Kucuk E, Khan H, Gibb EA, et al. The North American bullfrog draft genome provides insight into hormonal regulation of long noncoding RNA. Nat Commun. 2017;8 1:1433. doi:10.1038/s41467-017-01316-7.

19. Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, et al. Genome evolution in the allotetraploid frog *Xenopus laevis*. Nature. 2016;538 7625:336-43. doi:10.1038/nature19840.

20. Sun YB, Xiong ZJ, Xiang XY, Liu SP, Zhou WW, Tu XL, et al. Whole-genome sequence of the Tibetan frog *Nanorana parkeri* and the comparative evolution of tetrapod genomes. Proceedings of the National Academy of Sciences of the United States of America. 2015;112 11:E1257-62. doi:10.1073/pnas.1501764112.

21. Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, Kapitonov V, et al. The genome of the Western clawed frog *Xenopus tropicalis*. Science. 2010;328 5978:633-6. doi:10.1126/science.1183670.

22. The NCBI Assembly database: https://www.ncbi.nlm.nih.gov/assembly/?term=Anura; access on February 18, 2020.

23. Martin A. Studies in Australian amphibia III. The limnodynastes dorslis complex (Anura: Leptodactylidae). Australian Journal of Zoology. 1972;20 2:165-211.

24. Schauble CS, Moritz C and Slade RW. A molecular phylogeny for the frog genus Limnodynastes (Anura: myobatrachidae). Mol Phylogenet Evol. 2000;16 3:379-91. doi:10.1006/mpev.2000.0803.

25. Bertozzi T and Donnellan S. DNA extraction protocol for the eastern banjo frog using the Gentra Puregene Tissue Kit. protocols.io 2020; doi:dx.doi.org/10.17504/protocols.io.bcy6ixze.

26. Li Q, Guo Q, Zhou Y, Tan H, Bertozzi T, Zhu Y, et al. Construction and sequencing of DNA libraries on Hiseq 2000 platform for the eastern banjo frog. protocols.io 2020; doi:dx.doi.org/10.17504/protocols.io.bc22iyge.

27. Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, et al. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience. 2018;7 1:1-6. doi:10.1093/gigascience/gix120.

28. Marcais G and Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27 6:764-70. doi:10.1093/bioinformatics/btr011.

29. Cai H, Li Q, Fang X, Li J, Curtis NE, Altenburger A, et al. A draft genome assembly of the solar-powered sea slug *Elysia chlorotica*. Sci Data. 2019;6:190022. doi:10.1038/sdata.2019.22.

30. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33 14:2202-4. doi:10.1093/bioinformatics/btx153.

31. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 2014;24 8:1384-95. doi:10.1101/gr.170720.113.

32. Li Q, Guo Q, Zhou Y, Tan H, Bertozzi T, Zhu Y, et al. Repetitive element annotation protocol for the eastern banjo frog. protocols.io 2020; doi:dx.doi.org/10.17504/protocols.io.bc4niyve.

33. Smit AF, Hubley R and Green P. Available fom http://www.repeatmasker.org. 20 September 2019 date last accessed.

34. Smit A and Hubley R. Available fom http://www.repeatmasker.org/RepeatModeler/. 20 September 2019 date last accessed.

35. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research. 1999;27 2:573-80.

36. Li Q, Guo Q, Zhou Y, Tan H, Bertozzi T, Zhu Y, et al. Protein-coding gene annotation protocol for the eastern banjo frog. protocols.io 2020; doi:dx.doi.org/10.17504/protocols.io.bc38iyrw.

37. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215 3:403-10. doi:10.1016/S0022-2836(05)80360-2.

38. Birney E, Clamp M and Durbin R. GeneWise and Genomewise. Genome Res. 2004;14 5:988-95. doi:10.1101/gr.1865504.

39. Stanke M, Diekhans M, Baertsch R and Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24 5:637-44. doi:10.1093/bioinformatics/btn013.

40. Xiong Z, Li F, Li Q, Zhou L, Gamble T, Zheng J, et al. Draft genome of the leopard gecko, *Eublepharis macularius*. Gigascience. 2016;5 1:47. doi:10.1186/s13742-016-0151-4.

41. UniProt Consortium T. UniProt: the universal protein knowledgebase. Nucleic acids research. 2018;46 5:2699. doi:10.1093/nar/gky092.

42. Kanehisa M and Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28 1:27-30. doi:10.1093/nar/28.1.27.

43. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013.

44. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31 19:3210-2. doi:10.1093/bioinformatics/btv351.

45. Dohmen E, Kremer LP, Bornberg-Bauer E and Kemena C. DOGMA: domain-based transcriptome and proteome quality assessment. Bioinformatics. 2016;32 17:2577-81. doi:10.1093/bioinformatics/btw231.

46. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic acids research. 2014;42 Database issue:D222-30. doi:10.1093/nar/gkt1223.

47. Li Q, Guo Q, Zhou Y, Tan H, Bertozzi T, Zhu Y, et al. Genomic data from the Eastern banjo frog *Limnodynastes dumerilii dumerilii* (Anura: Limnodynastidae). GigaScience Database. 2020; doi:http://dx.doi.org/10.5524/100717.

Figures

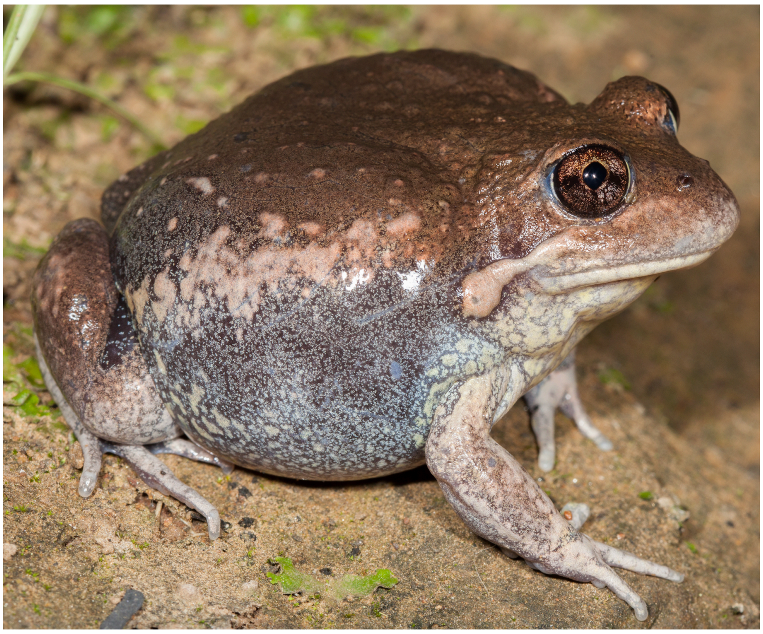

Figure 1. Photograph of an adult *Limnodynastes dumerilii dumerilii* from the Adelaide region (image from Stephen Mahony).

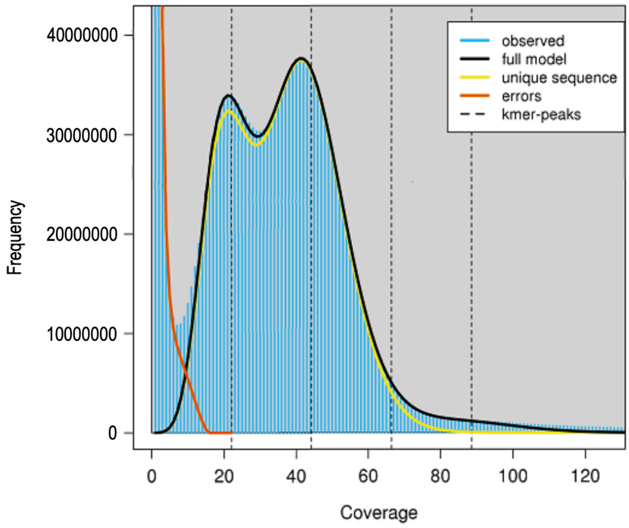

Figure 2. A 21-mer frequency distribution of the *L. d. dumerilii* genome data. The first peak at coverage 21X corresponds to the heterozygous peak. The second peak at coverage 42X corresponds to the homozygous peak.

Tables

Table 1. Statistics of DNA reads produced for the *L. d. dumerilii* genome.

Insert size (bp)	No. of Libraries	Read length (bp)	Raw data			Clean data		
			Total bases (Gb)	Sequence depth (X)	Physical depth (X)	Total bases (Gb)	Sequence depth (X)	Physical depth (X)
170	1	100	43.45	17.11	14.54	36.20	14.25	12.75
250	1	150	67.56	26.60	22.17	46.22	18.20	15.69
500	1	150	61.47	24.20	40.33	31.02	12.21	24.43
800	1	150	38.34	15.09	40.25	16.73	6.59	21.08
2,000	3	49	59.90	23.58	481.28	29.85	11.75	301.33
5,000	3	49	52.11	20.52	1046.72	11.88	4.68	299.82
10,000	2	49	36.81	14.49	1478.79	5.27	2.07	266.00
20,000	2	49	36.02	14.18	2894.10	2.54	1.00	256.41
Total	14		395.66	155.77	6018.18	179.71	70.75	1197.50

Note: Coverage calculation was based on the estimated haploid genome size of 2.54 Gb according to *k*-mer analysis. Sequence coverage is the average number of times a base is read, while physical coverage is the average number of times a base is spanned by sequenced fragments.

Table 2. Estimation of genome size and heterozygosity of *L. d. dumerilii* by *k*-mer analysis.

k	Total number of *k*-mers	Minimum coverage (X)	Number of erroneous *k*-mers	Homozygous peak	Estimated genome size (Gb)	Estimated heterozygosity (%)
17	112,401,363,509	9	1,418,748,938	45	2.47	1.10
19	110,136,516,133	8	2,588,664,358	43	2.50	1.23
21	107,871,808,889	7	3,023,604,282	42	2.50	1.24
23	105,607,392,491	7	3,286,834,146	40	2.56	1.22
25	103,343,108,760	7	3,501,481,190	39	2.56	1.19
27	101,078,882,097	7	3,689,197,189	38	2.56	1.16
29	98,815,880,190	6	3,839,002,752	37	2.57	1.14
31	96,552,885,503	6	3,986,778,359	36	2.57	1.11

Note: *k*-mer frequency distributions were generated by Jellyfish (v2.2.6) using 130 Gb clean sequences as input. Minimum coverage was the coverage depth value of the first trough in *k*-mer frequency distribution. *k*-mers with coverage depth less than the minimum coverage were regarded as erroneous *k*-mers. Estimated genome size was calculated as (Total number of *k*-mers – Number of erroneous *k*-mers) / Homozygous peak.

Table 3. Statistics of repetitive sequences identified in the *L. d. dumerilii* genome.

Category	Total repeat length (bp)	% of assembly
DNA	155,988,597	7.30%
LINE	242,754,702	11.36%
SINE	11,761,904	0.55%
LTR	97,615,246	4.57%
Tandem repeats	178,355,571	8.35%
Unknown	704,263,255	32.96%
Combined	1,205,873,056	56.43%

Note: DNA: DNA transposon; LINE: long interspersed nuclear element; SINE: short interspersed nuclear elements; LTR: long terminal repeat.

Table 4. Summary of protein-coding genes annotated in the *L. d. dumerilii* genome.

Characteristics of protein-coding genes	
Total number of protein-coding genes	24,548
Gene space (exon + intron; Mb)	634.6 (26.7 % of assembly)
Mean gene size (bp)	25,851
Mean CDS length (bp)	1,552
Exon space (Mb)	38.1 (1.6 % of assembly)
Mean exon number per gene	8.6
Mean exon length (bp)	181
Mean intron length (bp)	3,217
Functional annotation by searching public databases	
% of proteins with hits in UniProtKB/Swiss-Prot	95.8
% of proteins with hits in NCBI nr database	99.6
% of proteins with KO assigned by KEGG	71.3
% of proteins with functional annotation (combined)	99.9