Long-read sequencing and de novo genome assembly of Ammopiptanthus nanus, a desert shrub

Fei Gao1, Xue Wang1, Xuming Li2, Mingyue Xu2, Huayun Li3, Merhaba Abla1, Huigai Sun1, Shanjun Wei1, Jinchao Feng1 and Yijun Zhou1,∗

1College of Life and Environment Sciences, Minzu University of China, 27 Zhongguancun South Street, Beijing, 100081, China, 2Biomarker Technologies Corporation, Floor 8, Shunjie Building, 12 Fuqian Road, Nanfaxin Town, Shunyi District, Beijing, 101300, China; and 3Annoroad Genomics, Building B1,Yard 88, Kechuang six Road, Beijing Economic-Technological Development Area, Fengtai District, Beijing, 100176, China

∗Correspondence address. Yijun Zhou, College of Life and Environment Sciences, Minzu University of China, 27 Zhongguancun South Street, Beijing, 100081, China, E-mail: zhouyijun@muc.edu.cn

Abstract

Background: Ammopiptanthus nanus is a rare broad-leaved shrub that is found in the desert and arid regions of Central Asia. This plant species exhibits extremely high tolerance to drought and freezing and has been used in abiotic tolerance research in plants. As a relic of the tertiary period, A. nanus is of great significance to plant biogeographic research in the ancient Mediterranean region. Here, we report a draft genome assembly using the Pacific Biosciences (PacBio) platform and gene annotation for A. nanus.

Findings: A total of 64.72 Gb of raw PacBio sequel reads were generated from four 20-kb libraries. After filtering, 64.53 Gb of clean reads were obtained, giving 72.59× coverage depth. Assembly using Canu gave an assembly length of 823.74 Mb, with a contig N50 of 2.76 Mb. The final size of the assembled A. nanus genome was close to the 889 Mb estimated by k-mer analysis. The gene annotation completeness was evaluated using Benchmarking Universal Single-Copy Orthologs; 1,327 of the 1,440 conserved genes (92.15%) could be found in the A. nanus assembly. Genome annotation revealed that 74.08% of the A. nanus genome is composed of repetitive elements and 53.44% is composed of long terminal repeat elements. We predicted 37,188 protein-coding genes, of which 96.53% were functionally annotated.

Conclusions: The genomic sequences of A. nanus could be a valuable source for comparative genomic analysis in the legume family and will be useful for understanding the phylogenetic relationships of the Thermopsidaceae and the evolutionary response of plant species to the Qinghai Tibetan Plateau uplift.

Keywords: Ammopiptanthus nanus; PacBio sequencing; genome assembly; genome annotation

Data Description

Background information

Ammopiptanthus nanus, a desert shrub and a relic from the tertiary period, is one of two species in the genus Ammopiptanthus. This genus belongs to the tribe Thermopsidaceae and the family Fabaceae (Fig. 1). Ammopiptanthus is the only genus of evergreen broadleaf shrub distributed in the desert and arid regions of Central Asia. The plants in this genus play important ecological roles by fixing moving sands and delaying further desertification [1].

Tribe Thermopsidaceae is considered to be a basal branch in the family Fabaceae. The habitats of the ca. 45 plant species in tribe Thermopsidaceae are interspersed among the Mediterranean Basin, Central Asia, and temperate North America. Studies on the molecular biology of these plant species will promote understanding of the phylogeny of family Fabaceae, as well as some interesting biogeographical topics such as how the Qinghai-
Tibetan Plateau uplift and Tethys retreat affected plant evolution [2, 3]. In addition, the genus *Ammopiptanthus* is a unique and isolated branch in tribe Thermopsidaceae. There are still some debates about the evolution and phylogeny of this genus [9], and more molecular evidence is needed to clarify these issues.

Species in genus *Ammopiptanthus* exhibit extremely high tolerance to drought and freezing and have been used in abiotic tolerance research in plants [4–6]. Although several transcriptome analyses of the response to drought and cold stress have been conducted [1, 7–9], the lack of genome sequence information impedes further investigation into the molecular mechanism underlying the stress tolerance of *Ammopiptanthus* species.

Most of the *de novo* assemblies of plant genomes recently reported have been performed using next-generation sequencing technologies such as Illumina or 454 sequencing platforms [10–12]. However, these assemblies generally contain very fragmented sequences, partly because of the complexity of the plant genome. The newly developed Pacific BioSciences ( PacBio) sequencing platform, a third-generation sequencing technology, has started to address some of the intrinsic challenges in sequencing and assembling large and complex plant genomes by producing tens of thousands of long individual reads (up to ∼40 kb) [13]. Recently, several complicated plant genomes, including those of maize [14], sunflower [15], and *Chenopodium quinoa* [16], have been sequenced using the PacBio sequencing technology. In the present study, we used single-molecule real-time (SMRT) sequencing developed by PacBio to generate a draft genome assembly for *A. nanus*.

**Sample collection and genomic DNA sequencing**

The leaf tissues of a single *A. nanus* tree (National Center for Biotechnology Information [NCBI] taxonomy ID, 111851) were collected from Xinjiang, China. After collection, tissues were immediately transferred into liquid nitrogen and stored until DNA extraction. DNA was extracted using the Cetyltrimethyl Ammonium Bromide (CTAB) method according to the protocol “Preparing *Arabidopsis* Genomic DNA for Size-Selected ~20 kb SMRTbell™ Libraries” [17]. The quality of the extracted genomic DNA was checked using 1% agarose gel electrophoresis, and the concentration was quantified using a Qubit fluorimeter (Invitrogen, Carlsbad, CA, USA).

Long-read sequencing was performed at Biomarker Technologies Corporation (Beijing, China) with a PacBio Sequel sequencer (Pacific Biosciences, Menlo Park, CA, USA). The SMRT Bell library was prepared using a DNA Template Prep Kit 1.0 (PacBio p/n 100-259-150), and four 20-kb SMRTbell libraries were constructed. Genomic DNA (10 μg) was mechanically sheared using a Covaris g-Tube (Kbiosciences p/n S20079) with a goal of DNA fragments of approximately 20 kb. A Bioanalyzer 2100 12K DNA Chip assay (Agilent p/n 5067-1508) was used to assess the fragment size distribution. Sheared genomic DNA (5 μg) was DNA-damage repaired and end-repaired using polishing enzymes. A blunt-end ligation reaction followed by exonuclease treatment was conducted to generate the SMRT Bell template. A Blue Pippin device (Sage Science, Inc., Beverly, MA, USA) was used to size select the SMRT Bell template and enrich large fragments (>10 kb). The size-selected library was quality inspected and quantified on an Agilent Bioanalyzer 12 kb DNA Chip (Agilent Technologies, Santa Clara, CA, USA) and a Qubit fluorimeter (Invitrogen, Carlsbad, CA, USA). A ready-to-sequence SMRT Bell-Polymerase Complex was created using a Binding Kit 2.0 (PacBio p/n 100-862-200), according to the manufacturer’s instructions. The Sequel instrument was programmed to load and sequence the sample on PacBio SMRT cells v3.0 (PacBio p/n 100-171-800), acquiring one movie of 360 min per SMRT cell. The MagBead loading (PacBio p/n 100-125-900) method was used to improve the enrichment of the larger fragments. A total of 13 SMRT cells were processed yielding 64.72 G subread sequences.

For Illumina sequencing, paired-end (PE) libraries with insert sizes of 350 bp were constructed with the standard protocol provided by Illumina (San Diego, CA, USA) and sequenced on an Illumina HiSeq X Ten platform. A total of 55.97 Gb of PE (2 × 150 bp) clean sequences were generated (Supplementary Table S1). These data were used for genome size estimation, correction of genome assembly, and assembly evaluation.

**Genome size estimation**

We characterized genome size and heterozygosity using the distribution of k-mers of length 19 from the Illumina HiSeq reads (55.97 Gb clean reads from 350 bp insert size library; NCBI SRA accession number, SRX3286209). This analysis was performed using “kmer_freq_stat” software (developed by Biomarker Technologies). The genome size (G) of *A. nanus* was estimated using the following formula: \[G = \frac{k \times \text{mer number}}{\text{average k-mer depth}}\] where the k-mer number = total k-mers—abnormal k-mers (with too low or too high frequency). The highest peak in the k-mer distribution curve was found at the k-mer depth of 53, with a k-mer number of 47,408,863,457 (Supplementary Fig. S1). The peak at depth of more than 106 was a repetitive peak (k-mers duplicated because of repetition). Finally, the *A. nanus* genome size was estimated to be 888.92 Mb, the heterozygosity was approximately 0.02%, and the data used in 19-mer analysis was approximately 53× coverage of the genome.

**Genome assembly**

The Sequel raw bam files were converted into subreads in fasta format using the standard PacBio SMRT software package (read data are available at the NCBI SRA accession number, SRX3262947). Then, subreads of less than 500 bp were filtered out. Finally, 7,918,322 reads and 64,538,018,400 bases (∼73× depth) were produced. The average subread length was 8.15 kb with a N50 length of 12.79 kb (Supplementary Table S2). The genome assembly was conducted using Canu software (v1.5) [18] (correctedErrorRate = 0.045, corOutCoverage = 70). The draft genome was polished with Arrow (SMRT link v5.0.1,
Repeat annotation and gene prediction

For repeat detection, four software packages, i.e., LTR-FINDER (v1.0.5) [20], MITE-Hunter (v1.0.0) [21], PILER (v1.0) [22], and RepeatScout v1.0.5 (RepeatScout, RRID:SCR_014653) [23], were used to build a de novo repeat library on the basis of our assembly with the default settings. Then, the predicted repeats were classified using PASTEClassifier (v1.0) [24] and merged with Repbase (19.06) [25]. Finally, using the resulting repeat database as the final repeat library, RepeatMasker v4.0.5 (RepeatMasker, RRID:SCR_012954) [26] was used to identify repetitive sequences in the A. nanus genome with the following parameters: -noinput -no_in -nomatch -engine wublast. Overall, approximately 610.25 Mb of repetitive sequences (74.08% of the assembly) were detected, containing 440.18 Mb (53.44% of the assembly) long terminal read elements (Supplementary Table S4).

Ab initio-based, homolog-based, and RNA-sequencing (RNA-seq)-based gene prediction methods were used in combination to identify the protein-coding genes in the A. nanus genome assembly. Genscan [27], Augustus v2.4 (Augustus, RRID:SCR_008417) [28], GlimmerHMM v3.0.4 (GlimmerHMM, RRID:SCR_002654) [16], GeneID (v1.4) [29], and SNAP v2006-07-28 (SNAP, RRID:SCR_002127) [30] with the default parameters were used for the ab initio-based gene prediction, and all of these software packages were trained using the Arabidopsis gene model before gene prediction. For gene prediction using Augustus, in addition to the Arabidopsis’s gene model, the Program to Assemble Spliced Alignments (PASA’s) gene model was also used as the initial gene model for training. Finally, the best gene model with higher accuracy and specificity was used. The quality of the gene models was evaluated by aligning transcriptome sequences to the whole genome assembly using TopHat (Supplementary Table S5). GeMoMa (v1.3.1) [31] was used in homolog-based gene annotation, and the protein databases of Cicer arietinum (GCA_00031145.1), Phaseolus vulgaris (GCA_000499845.1), Glycine max (GCA_00004515.3), and Arachis duranensis (GCA_000817695.2) from GenBank were used as the references. For the RNA-seq-based method of gene prediction, TransDecoder (v2.0, [32]), GeneMarkS-T v5.1 (RRID:SCR_011930) [33], and PASA v2.0.2 (RRID:SCR_014656) [34] were used, and the A. nanus transcriptome data were assembled in a previous study (NCBI SRA accession numbers, SRX1409432 and SRX1406652) [35]. Finally, the results from the three methods were integrated using EVM (v1.1.1, RRID:SCR_014659) [36]. Higher weights were assigned to the PASA-predicted transcripts from unigenes and GeMoMa-predicted homologous transcripts than to the ab initio-predicted transcripts when conducting the EVM integration. In total, a gene set with 37,144 protein-coding genes was predicted from the A. nanus genome assembly (Table 1, Supplementary Table S6, Supplementary Fig. S2). These genes were scattered over 1,099 contigs, averaging 33.80 genes per contig. The genes were annotated by aligning to the Non-redundant protein sequences; Nt: Nucleotide collection (NR, Nt), eukaryotic orthologous groups of proteins (KOG) [37], Kyoto Encyclopedia of Genes and Genomes (KEGG) (KEGG, RRID:SCR_001120) [31], SwissProt (Swissprot, RRID:SCR_002380) [38], and TrEMBL [39] databases using the Basic Local Alignment Search Tool (BLAST) with an e-value cutoff of 1E-5 and also aligned to the Pfam (Pfam, RRID:SCR_004726) database [40] using hmmr V3.0 (-E 0.00001 –domE 0.00001 –cpu 2 –noal -acc) [41]. Gene Ontology (GO) terms were assigned to the genes using the BLAST2GO pipeline [42]. In all, 96.71% of the predicted genes could be classified into families according to their putative functions (Table 2).

For pseudogene prediction, GenBlastA [43] was used to scan the A. nanus genome for sequences homologous to the known protein-coding genes it contained. Then, GeneWise (GeneWise, RRID:SCR_015054) [44] was adopted to search the premature stop codons or frameshift mutations in those sequences and, consequently, to identify pseudogenes. In total, 7,891 pseudogenes were identified from the A. nanus genome (Supplementary Table S7).
Second, the A. nanus unigenes assembled in a previous study (NCBI SRA accession numbers, SRX1409432 and SRX1406652) were aligned to the A. nanus genome using the BLAST-like alignment tool v0.36 (BLAT, RRID:SCR_011919) with default parameters. The alignment indicated that 100% of unigene (>500 bp in length) assemblies were mapped to the A. nanus genome assembly (Table 3).

We also evaluated the completeness of the genome assembly of A. nanus using Benchmarking Universal Single-Copy Orthologs (BUSCO) v2.0 (BUSCO, RRID:SCR_015008) [49]. The results showed that 9,215% (1,327 out of 1,440 BUSCOs) of plant sets (embryophytaodb9, download from [50]) were identified as complete in the A. nanus assembly (Supplementary Table S9). Together, the results indicated that our dataset represented a genome assembly with a high level of coverage.

Conclusions

In summary, the draft genome sequence of A. nanus that we obtained demonstrated that third-generation sequencing technology, such as the PacBio platform, could be useful in deciphering complex plant genomes. The availability of the A. nanus genome sequence should facilitate de novo genome assembly of other species in this genus. The datasets from the present study could not only provide a valuable source for further comparative genomics analysis in the legume family and help to answer some important questions related to the biogeography in the ancient Mediterranean region but also could facilitate our understanding of how plants adapt to the stressful conditions in temperate deserts in Central Asia.

Availability of supporting data

Raw genomic sequence reads are available in the NCBI Sequence Read Archive under project number PRJNA413722. Supporting data are available from the GigaScience GigaDB database [51].

Additional files

Table S1. Distribution of k-mers of length 19 from the Illumina HiSeq reads
Table S2. Venn diagram plot using UpSetR showing the overlap of the gene annotation results of the three gene prediction methods
Table S3. Illumina sequencing reads used for genome size estimation and correction of genome assembly
Table S4. Size distribution of the PacBio subreads
Table S5. Statistics of the identified repeat sequences
Table S6. Statistics of the genome annotation
Table S7. Statistics of the predicted pseudogenes
Table S8. The alignment of the Illumina reads to the A. nanus genome assembly
Table S9. BUSCO assessment of the A. nanus genome assembly

Abbreviations

BLAST: Basic Local Alignment Search Tool; BUSCO: Benchmarking Universal Single-Copy Orthologs; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; KOG: eukaryotic orthologous groups of proteins; PacBio: Pacific Biosciences; PASA: PE: paired-end; QV: quality value; RNA-seq: RNA sequencing; SMRT: single-molecule real-time.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was financially supported by the National Natural Science Foundation of China (31670335 and 31770363) and the Ministry of Education of China through 111 and 985 projects (B08044, YLDX01013).

Author contributions

Y.Z. and J.F. oversaw the project. M.A., X.W., and S.W. collected the sample and extracted the genomic DNA. X.L., H.L., M.X., and H.S. performed the genome assembly, annotated the genome, and analyzed data. F.G. analyzed data. F.G. and Y.Z. wrote the manuscript.

References

1. Gao F, Wang J, Wei S, et al. Transcriptomic analysis of drought stress responses in Ammopiptanthus mongolicus leaves using the RNA-seq technique. PLoS One 2015;10(4):e0124382.
2. Zhang ML, Huang JF, Sanderson SC, et al. Molecular biogeography of tribe Thermopsideae (Leguminosae): a Madrean-Tethyan disjunction pattern with an African origin of core Genistoides. Biomed Res Int 2015;2015:864804.
3. Shi W, Liu PL, Duan L, et al. Evolutionary response to the Qinghai-Tibetan Plateau uplift: phylogeny and biogeography of Ammopiptanthus and tribe Thermopsideae (Fabaceae). Peer J 2017;5:e3607.
4. Xu S, An L, Feng H, et al. The seasonal effects of water stress on Ammopiptanthus mongolicus in a desert environment. J Arid Environ 2002;51(3):437–47.
5. Wang W, Chen Y, Liu M, et al. Effects of cold-hardening on compatible solutes and antioxidant enzyme activities related to freezing tolerance in Ammopiptanthus mongolicus seedlings. For Stud China 2008;10(2):101–6.
6. Gao T-F, Chen T, Feng H-Y, et al. Seasonal and annual variation of osmotic solute and stable carbon isotope composition in leaves of endangered desert evergreen shrub Ammopiptanthus mongolicus. S Afr J Bot 2006;72(4):570–8.
7. Zhou Y, Gao F, Liu R, et al. De novo sequencing and analysis of root transcriptome using 454 pyrosequencing to discover putative genes associated with drought tolerance in Ammopiptanthus mongolicus. BMC Genomics 2012;13:266.
8. Wu Y, Wei W, Pang X, et al. Comparative transcriptome profiling of a desert evergreen shrub, Ammopiptanthus mongolicus, in response to drought and cold stresses. BMC Genomics 2014;15:671.
9. Pang T, Ye CY, Xia X, et al. De novo sequencing and transcriptome analysis of the desert shrub, *Ammopiptanthus mongolicus*, during cold acclimation using Illumina/ Solexa. BMC Genomics 2013;14:488.

10. Fu Y, Li L, Hao S, et al. Draft genome sequence of the Tibetan medicinal herb Rhodiola crenulata. GigaScience 2017;6(6):1–5.

11. Zhao D, Hamilton JP, Phar M, et al. De novo genome assembly of *Camptotheca acuminata*, a natural source of the anti-cancer compound camptothecin. GigaScience 2017;6(9):1–7.

12. Xia EH, Zhang HB, Sheng J, et al. The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. Mol Plant 2017;10(6):866–77.

13. Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. Genome Biol 2013;14(7):405.

14. Jiao Y, Peluso P, Shi J, et al. Improved maize reference genome with single-molecule technologies. Nature 2017;546(7659):524–7.

15. Badouin H, Gouzy J, Grassa CJ, et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature 2017;546(7656):148–52.

16. Jarvis DE, Ho YS, Lightfoot DJ, et al. The genome of *Chenopodium quinoa*. Nature 2017;542(7641):307–12.

17. PacBio shared protocol preparing Arabidopsis DNA for 20kb SMRTbell libraries , Preparing Arabidopsis Genomic DNA for Size-Selected ~20 kb SMRTbell™ Libraries, Accessed 15 Jul 2017. http://www.pacbio.com/wp-content/uploads/2015/09/Shared-Protocol-Preparing-Arabidopsis-DNA-for-20-kb-SMRTbell-Libraries.pdf

18. Koren S, Walenz BP, Berlin K, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 2017;27(5):722–36.

19. Walker BJ, Abeel T, Shea T, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 2014;9(11):e112963.

20. Xu Z, Wang H. LTRoFINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 2007;35(web server issue):W265–8.

21. Han Y, Wessler SR. MITe-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 2010;38(22):e199.

22. Edgar RC, Myers EW. PILER: identification and classification of genomic repeats. Bioinformatics 2005;21:1152–8.

23. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics 2005;21(suppl 1):i351–8.

24. Hoebe C, Arnoux S, Moisset M, et al. PASTEC: an automatic transposable element classification tool. PLoS One 2014;9:e91929.

25. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 2015;6:11.

26. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics, Hoboken: John Wiley & Sons 2009; Chapter 4: Unit 4.10.

27. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997;268:78–94.

28. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 2003;19(Suppl 2):ii215–25.

29. Blanco E, Parra G, Guió R. Using geneid to identify genes. Curr Protoc Bioinformatics, Hoboken: John Wiley & Sons, 4, 3.4.3.1–4.3.28. 2007.

30. Korf I. Gene finding in novel genomes. BMC Bioinformatics 2004;5:59.

31. Keilwagen J, Wenk M, Erickson JL, et al. Using intron position conservation for homology-based gene prediction. Nucleic Acids Res 2016;44:e89.

32. Transdecoder http://transdecoder.github.io

33. Tang S, Lomsadze A, Borodovsky M. Identification of protein coding regions in RNA transcripts. Nucleic Acids Res 2015;43(12):e78.

34. Campbell MA, Haas BJ, Hamilton JP, et al. Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genomics 2006;7:327.

35. Gao F, Li H, Xiao Z, et al. De novo transcriptome analysis of *Ammopiptanthus nanus* and its comparative analysis with *A. mongolicus*. Trees 2018;32(1):287–300.

36. Haas BJ, Salzberg SL, Zhu W, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol 2008;9(1):R7.

37. Tatusov RL, Natale DA, Garkavtsev IV, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 2001;29(1):22–8.

38. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000;28(1):27–30.

39. Bickhart DM, Rosen BD, Koren S, et al. Single-molecule sequencing and de novo genome assembly of *Ammopiptanthus nanus*, a desert shrub.” GigaScience Database 2018. http://dx.doi.org/10.5524/100466