

1 **Draft genome of the Reindeer (*Rangifer tarandus*)**

2 **Zhipeng Li<sup>1\*</sup>, Zeshan Lin<sup>2\*</sup>, Lei Chen<sup>2\*</sup>, Hengxing Ba<sup>1\*</sup>, Yongzhi Yang<sup>2</sup>, Kun**  
3 **Wang<sup>2</sup>, Wen Wang<sup>2#</sup>, Qiu Qiang<sup>2#</sup>, Guangyu Li<sup>1#</sup>**

4 <sup>1</sup> Jilin Provincial Key Laboratory for Molecular Biology of Special Economic  
5 Animals, Institute of Special Animal and Plant Sciences, Chinese Academy of  
6 Agricultural Sciences, Changchun, 130112, China

7 <sup>2</sup> Center for Ecological and Environmental Sciences, Northwestern Polytechnical  
8 University, Xi'an 710072, China

9 \* These authors contributed equally to this work.

10 # Corresponding authors: [tcslyg@126.com](mailto:tcslyg@126.com) (GL), [qiuqiang@lzu.edu.cn](mailto:qiuqiang@lzu.edu.cn) (QQ),  
11 [wwang@wangwen-lab.org](mailto:wwang@wangwen-lab.org) or [wwang@mail.kiz.ac.cn](mailto:wwang@mail.kiz.ac.cn) (WW)

12

13 **Abstract**

14 **Background:** Reindeer (*Rangifer tarandus*) is the only fully domesticated species in  
15 the Cervidae family, and is the only cervid with a circumpolar distribution. Unlike all  
16 other cervids, female reindeer regularly grow cranial appendages (antlers, the defining  
17 characteristics of cervids), as well as males. Moreover, reindeer milk contains more  
18 protein and less lactose than bovids' milk. A high quality reference genome of this  
19 specie will assist efforts to elucidate these and other important features in the reindeer.

20 **Findings:** We obtained 723.2 Gb (Gigabase) of raw reads by an Illumina Hiseq 4000  
21 platform, and a 2.64 Gb final assembly, representing 95.7% of the estimated genome  
22 (2.76 Gb according to k-mer analysis), including 92.6% of expected genes according  
23 to BUSCO analysis. The contig N50 and scaffold N50 sizes were 89.7 kilo base (kb)  
24 and 0.94 mega base (Mb), respectively. We annotated 21,555 protein-coding genes  
25 and 1.07 Gb of repetitive sequences by *de novo* and homology-based prediction.  
26 Homology-based searches detected 159 rRNA, 547 miRNA, 1,339 snRNA and 863  
27 tRNA sequences in the genome of *R. tarandus*. The divergence time between *R.*  
28 *tarandus*, and ancestors of *Bos taurus* and *Capra hircus*, is estimated to be 29.55  
29 million years ago (Mya).

30 **Conclusions:** Our results provide the first high-quality reference genome for the  
31 reindeer, and a valuable resource for studying evolution, domestication and other  
32 unusual characteristics of the reindeer.

33 **Keywords:** *Rangier tarandus*, whole genome sequencing, assembly, annotation

## 34 **Background information**

35       The Cervidae is the second largest family in the suborder Ruminantia of the  
36 Artiodactyla, which are distributed across much of the globe in diverse habitats, from  
37 arctic tundra to tropical forests [1, 2]. Interestingly, reindeer (*Rangifer tarandus*) is  
38 the only species with a circumpolar distribution (present in boreal, tundra, subarctic,  
39 arctic and mountainous regions of northern Asia, North America and Europe). It is  
40 also the only cervid having been fully domesticated, although some other species,  
41 such as the sika deer (*Cervus nippon*), which has been semi-domesticated for more  
42 than 200 years and still has strong wild nature. Antlers, male secondary sexual  
43 appendage, are the defining characteristic of cervids, which shed and regrow each  
44 year throughout an animal's life. However, reindeer do not follow this rule, with the  
45 exception in which females also bear shedding antlers. Moreover, reindeer milk  
46 contains greater amount of proteins, and lower amount of lactose compared to that of  
47 bovids [3]. Here, we report a high-quality reindeer reference genome using material  
48 from a Chinese individual, which will be useful in elucidating special characteristics  
49 of special cervid.

## 50 **Data description**

### 51 **Animal and sample collecting**

52       Fresh blood was collected from a two-year-old, female reindeer of a  
53 domesticated herd maintained by Ewenki hunter-herders in the Greater Khingan

54 Mountains, Inner Mongolia Autonomous Region, China (50.77° N, 121.47° E). The  
55 sample was immediately placed in liquid nitrogen, and was then stored at -80°C for  
56 later analysis.

### 57 **Library construction, sequencing and filtering**

58 Genomic DNA was extracted from the fresh blood. The isolated genomic DNA  
59 was then used to construct five short-insert libraries (200, 250, 350, 400 and 450 base  
60 pair, bp) and four long-insert libraries (3, 6.5, 11.5 and 16 kb) following standard  
61 protocols provided by Illumina. Then, 150 bp paired-end sequencing was performed  
62 to generate 723.2 Gb raw data, using a whole genome shotgun sequencing strategy on  
63 an Illumina Hiseq 4000 platform (**Table S1**). To improve the quality of reads, we  
64 trimmed low-quality bases from both sides of reads and removed reads with more  
65 than 5% of uncalled (“N”) bases. Then reads of all libraries were corrected by  
66 SOAPec (version 2.03) [4]. Finally, clean reads amounting to 615 Gb were obtained  
67 for genome assembly.

### 68 **Evaluation of genome size**

69 The estimated genome size is 2.76 Gb according to k-mer analysis, based on the  
70 following formula:  $G = \text{k-mer\_number} / \text{k-mer\_depth}$  (**Figure S1**) [5]. All the clean  
71 reads provide approximately ~ 220-fold mean coverage.

### 72 **Genome assembly**

73 We used SOAPdenovo (version 2.04) with optimized parameters (pregraph -K  
74 79 -d 0; map -k 79; scaff -L 200) to construct contigs and original scaffolds [5]. All  
75 reads were aligned onto contigs for scaffold construction by utilizing the paired-end  
76 information. Gaps were filled using reads from three libraries (200, 250 and 350 bp)  
77 with GapCloser (version 1.12) [6]. The final reindeer genome assembly is 2.64 Gb  
78 long, including 95.7 Mb (3.6%) of unknown bases, smaller than that of the domestic  
79 goat (*Capra hircus*, 2.92 Gb) [7], and similar to that of sheep (*Ovis aries*, 2.61 Gb) [8].  
80 The contig N50 (> 200 bp) and scaffold N50 (> 500 bp) sizes are 89.7 kb and 0.94 Mb,  
81 respectively (**Table 1**).

## 82 **Quality assessments of the assembled genome**

83 We used BUSCO (benchmarking universal single-copy orthologs, version 2.0)  
84 software to assess the genome completeness [9]. Our assembly covered 92.6% of the  
85 core genes, with 3,803 genes being complete (**Table S2**). Feature-response curve  
86 (FRC, version 1.3.1) method [10] was then used to evaluate the trade-off between the  
87 assembly's contiguity and correctness. The results indicate that it has similar  
88 accumulated curve compare to published high quality assemblies for ruminant  
89 genomes including cattle, goat, and sheep (**Figure S2**). Subsequently, synteny  
90 analysis was applied to identify differences between the assembled genome and the  
91 domestic goat (*Capra hircus*) genome (**Figure S3**). 83.95% of two genome sequences  
92 could be 1:1 aligned, the average nuclear distance (percentage of different base pairs  
93 in the syntenic regions) was 7.18% (**Figure S4**). In addition, the density of different

94 types of break points (edges of structural variation) are about 69.88 per Mb (**Table**  
95 **S3**). These results suggest that the reindeer genome assembly is of good level of  
96 contiguity and correctness.

### 97 **Genome annotation**

98 To annotate the reindeer genome we initially used LTR\_FINDER [11] and  
99 RepeatModeller (version 1.0.4; <http://www.repeatmasker.org/RepeatModeler.html>) to  
100 find repeats. Next, RepeatMasker (version 4.0.5) [12] was used (with -nolow -no\_is  
101 -norna -parallel 1 parameters) to search for known and novel transposable elements  
102 (TE) by mapping sequences against the *de novo* repeat library and Repbase TE library  
103 (version 16.02) [13]. Subsequently, tandem repeats were annotated using Tandem  
104 Repeat Finder (version 4.07b; with 2 7 7 80 10 50 2000 -d -h parameters) [14]. In  
105 addition, we used RepeatProteinMask software [12] with -no LowSimple -p value  
106 0.0001 parameters to identify TE-relevant proteins. The combined results indicate that  
107 repeat sequences cover about 1.07 Gb, accounting for 40.4% of the reindeer genome  
108 assembly (**Table S4**).

109 The rest of the reindeer genome assembly was annotated using both *de novo* and  
110 homology-based gene prediction approaches. For *de novo* gene prediction, we utilized  
111 SNAP (version 2006-07-28), GenScan [15], glimmerHMM and Augustus (version  
112 2.5.5) [16] to analyze the repeat-masked genome. For homology-based predictions,  
113 sequences encoding homologous proteins of *Bos taurus* (Ensemble 87 release), *Ovis*  
114 *aries* (Ensemble 87 release) and *Homo sapiens* (Ensemble 87 release), were aligned to

115 the reindeer genome using TblastN (version 2.2.26) with an (E)-value cutoff of 1 e-5.  
116 Genewise (version wise2.2.0) [17] was then used to annotate structures of the genes.  
117 The *de novo* and homology gene sets were merged to form a comprehensive,  
118 non-redundant gene set using EVidenceModeler software (EVM, version 1.1.1),  
119 which resulted in 21,555 protein-coding genes (**Table S5**). We then compared the  
120 reindeer genome with species which used in homology prediction, and there is no  
121 significant difference among the four species in gene length and exon length  
122 distribution (**Figure S5**).

123 Next, we searched the KEGG, TrEMBL and SwissProt databases for best  
124 matches to the protein sequences yielded by EVM software, using BLASTP (version  
125 2.2.26) with an (E)-value cutoff of 1 e-5, and searched Pfam, PRINTS, ProDom and  
126 SMART databases for known motifs and domains in our sequences using  
127 InterProScan software (version 5.18-57.0). At least one function was assigned to  
128 19,004 (88.17%) of the detected reindeer genes through these procedures (**Table S6**).  
129 The reads from short-insert length libraries then were mapped to the reindeer genome  
130 with BWA (version 0.7.12-r1039) [18], then called single nucleotide variant (SNV)  
131 by SAMtools (version 1.3.1) [19]. Finally, we performed SnpEff (version 4.30) [20]  
132 to identify the distribution of SNV in the reindeer genome (**Table S7**).

133 In addition, we predicted rRNA-coding sequences based on homology with  
134 human rRNAs using BLASTN with default parameters. To annotate miRNA and  
135 snRNA genes we searched the Rfam database (release 9.1) with Infernal (version

136 0.81), and annotated tRNAs using tRNAscan-SE (version 1.3.1) software with default  
137 parameters. The final results identified 159 rRNAs, 547 miRNAs, 1,339 snRNAs and  
138 863 tRNAs (**Table S8**).

### 139 **Species-specific genes and phylogenetic relationship**

140 We clustered the detected reindeer genes in families by using OrthoMCL [21]  
141 with an (E)-value cutoff of  $1 \times 10^{-5}$ , and a Markov Chain Clustering with default  
142 inflation parameter in an all-to-all BLASTP analysis of entries for five species (*Homo*  
143 *sapiens*, *Equus caballus*, *Capra hircus*, *Bos taurus*, and *Rangifer tarandus*). The  
144 result showed that 335 gene families were specific to the reindeer (**Figure S6**).  
145 Moreover, we identified 7,505 single-copy gene families from these species and  
146 aligned coding sequences in the families using PRANK (version 3.8.31) [22].  
147 Subsequently, 4D-sites (four-fold degenerated sites) were extracted to construct a  
148 phylogenetic tree by RAxML (version 7.2.8) [23] with GTR+G+I model. Finally,  
149 phylogenetic analysis using PAML MCMCtree (version 4.5) [24], calibrated with  
150 published timings of the divergence of the reference species  
151 (<http://www.timetree.org/>), indicated that *Rangifer tarandus*, *Bos taurus* and *Capra*  
152 *hircus* diverged from a common ancestor approximately 29.6 (25.4-31.7) Mya  
153 (**Figure S7**).

### 154 **Conclusion**

155 In summary, we report the first sequencing, assembly and annotation of the

156 reindeer genome, which will be useful in analysis of the genetic basis of the unique  
157 characteristics of reindeer, and broader studies on ruminants.

#### 158 **Availability of supporting data**

159 The raw data have been deposited in Genome Sequence Archive (GSA), under  
160 BIG Data Center, Beijing Institute Genomics (BIG), Chinese Academy of Science,  
161 with the project accession PRJCA000451.

#### 162 **Abbreviations**

163 Gb: giga base; bp: base pair; kb: kilo base; Mb: mega base; TE: transposable  
164 element; EVM: EVidenceModeler; BUSCO: benchmarking universal single-copy  
165 orthologs; FRC: feature-response curves; SNV: single nucleotide variant; Mya:  
166 million years ago

#### 167 **Acknowledgements**

168 This work was supported by the he Natural Science Foundation of China (No.  
169 31501984) and Central Public-interest Scientific Institution Basal Research Fund (No.  
170 1610342016026) to ZPL, and Talents Team Construction Fund of Northwestern  
171 Polytechnical University (NWPU) to QQ and WW. Special thanks to Nowbio Biotech  
172 Inc., Kunming, China for its remarkable work on DNA libraries constructions and  
173 sequencing.

#### 174 **Competing interests**

175           The authors declare that they have no competing interests.

176    **Authors' contributions**

177           ZPL collected the samples; ZSL, CL ZPL, YZ, KW and HB analyzed the data;

178    ZSL, QQ and ZPL wrote the manuscript; GL, ZL, QQ and WW conceived the study.

179

180 **References**

- 181 1. Fernández MH and Vrba ES. A complete estimate of the phylogenetic  
182 relationships in ruminantia: a dated species-level supertree of the extant  
183 ruminants. *Biological Reviews*. 2005;80 2:269-302.
- 184 2. Hassanin A, Delsuc F, Ropiquet A, Hammer C, Jansen van Vuuren B, Matthee  
185 C, et al. Pattern and timing of diversification of Cetartiodactyla (*Mammalia*,  
186 *Laurasiatheria*), as revealed by a comprehensive analysis of mitochondrial  
187 genomes. *C R Biol*. 2012;335 1:32-50.
- 188 3. Young W. Park GFWH. Handbook of milk of non-bovine mammals.  
189 Wiley-Blackwell; 2006.
- 190 4. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an  
191 empirically improved memory-efficient short-read de novo assembler.  
192 *GigaScience*. 2012;1 1:1-6.
- 193 5. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and de novo  
194 assembly of the giant panda genome. *Nature*. 2010;463 7279:311-7.
- 195 6. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of  
196 human genomes with massively parallel short read sequencing. *Genome Res*.  
197 2010;20 2:265-72.
- 198 7. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al.  
199 Single-molecule sequencing and chromatin conformation capture enable de  
200 novo reference assembly of the domestic goat genome. *Nat Genet*. 2017;49  
201 4:643-50.
- 202 8. Jiang Y, Xie M, Chen W, Talbot R, Maddox JF, Faraut T, et al. The sheep  
203 genome illuminates biology of the rumen and lipid metabolism. *Science*.  
204 2014;344 6188:1168-73.
- 205 9. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM.  
206 BUSCO: assessing genome assembly and annotation completeness with  
207 single-copy orthologs. *Bioinformatics*. 2015;31 19:3210-2.
- 208 10. Vezzi F, Narzisi G and Mishra B. Reevaluating assembly evaluations with  
209 feature response curves: GAGE and assemblathons. *PLoS ONE*. 2012;7  
210 12:e52210.
- 211 11. Xu Z and Wang H. LTR\_FINDER: an efficient tool for the prediction of  
212 full-length LTR retrotransposons. *Nucleic Acids Res*. 2007;35  
213 suppl\_2:W265-W8.
- 214 12. Tarailo-Graovac M and Chen N. Using RepeatMasker to identify repetitive  
215 elements in genomic sequences. *Current Protocols in Bioinformatics*. John  
216 Wiley & Sons, Inc.; 2009.
- 217 13. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O and  
218 Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements.  
219 *Cytogenet Genome Res*. 2005;110 1-4:462-7.
- 220 14. Benson G. Tandem repeats finder: a program to analyze DNA sequences.  
221 *Nucleic Acids Res*. 1999;27 2:573-80.
- 222 15. Burge C and Karlin S. Prediction of complete gene structures in human  
223 genomic DNA1. *J Mol Biol*. 1997;268 1:78-94.

- 224 16. Stanke M, Keller O, Gunduz I, Hayes A, Waack S and Morgenstern B.  
225 AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.*  
226 2006;34 suppl\_2:W435-W9.
- 227 17. Birney E, Clamp M and Durbin R. GeneWise and Genomewise. *Genome Res.*  
228 2004;14 5:988-95.
- 229 18. Heng L. Aligning sequence reads, clone sequences and assembly contigs with  
230 BWA-MEM. arXiv. 2013;1303.3997
- 231 19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The  
232 sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25  
233 16:2078-9.
- 234 20. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A  
235 program for annotating and predicting the effects of single nucleotide  
236 polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster*  
237 strain w(1118); iso-2; iso-3. *Fly (Austin).* 2012;6 2:80-92.
- 238 21. Li L, Stoeckert CJ and Roos DS. OrthoMCL: Identification of Ortholog  
239 Groups for Eukaryotic Genomes. *Genome Res.* 2003;13 9:2178-89.
- 240 22. Löytynoja A and Goldman N. An algorithm for progressive multiple  
241 alignment of sequences with insertions. *Proc Natl Acad Sci U S A.* 2005;102  
242 30:10557-62.
- 243 23. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and  
244 post-analysis of large phylogenies. *Bioinformatics.* 2014;30 9:1312-3.
- 245 24. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol*  
246 *Evol.* 2007;24 8:1586-91.

## Tables

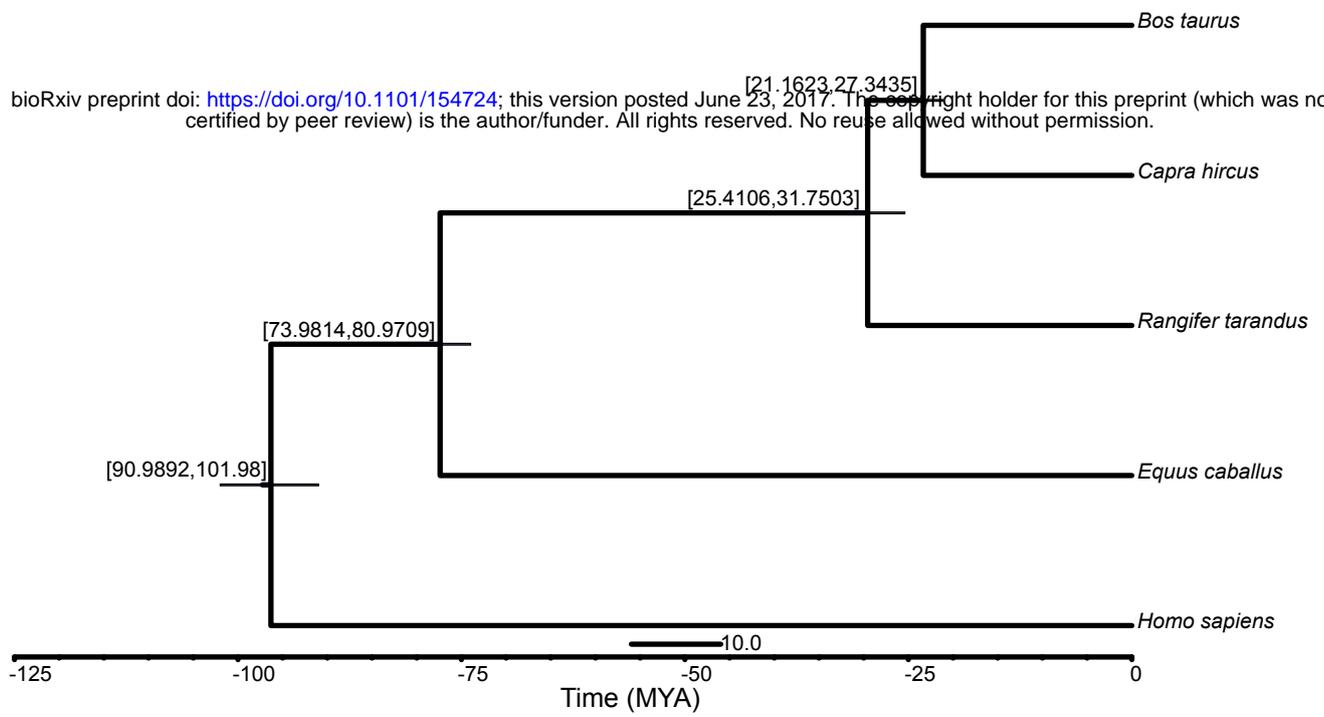
247 **Table 1 Summary of the genome assembly of *Rangier tarandus***

Type	Scaffold (bp)	Contig (bp)
Total number	58,765	117,102
Total length	2,832,785,815	2,732,476,387
N50 length	986,392	91,805
N90 length	151,297	17,480
Max length	4,664,725	770,474
GC content(%)	41.24	40.98

## Figure legends

248 **Figure 1. Phylogenetic relationships of *Rangier tarandus* and four species based**  
249 **on four-fold degenerated sites.** Estimated divergence times are shown above the  
250 nodes. MYA, million years ago.





**Figure 1.** Phylogenetic relationships of *Rangifer tarandus* and four species based on four-fold degenerated sites. Estimated divergence times are shown above the nodes. MYA, million years ago.