

# GigaScience

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

--Manuscript Draft--

<b>Manuscript Number:</b>	GIGA-D-17-00152R1	
<b>Full Title:</b>	Draft genome of the Reindeer ( <i>Rangifer tarandus</i> )	
<b>Article Type:</b>	Data Note	
<b>Funding Information:</b>	National Natural Science Foundation of China (31501984)	Dr Zhipeng Li
	Central Public-interest Scientific Institution Basal Research Fund (1610342016026)	Dr Zhipeng Li
	Talents Team Construction Fund of Northwestern Polytechnical University	Dr Qiang Qiu
	Talents Team Construction Fund of Northwestern Polytechnical University	Dr Wen Wang
<b>Abstract:</b>	<p>Background: Reindeer (<i>Rangifer tarandus</i>) is the only fully domesticated species in the Cervidae family, and is the only cervid with a circumpolar distribution. Unlike all other cervids, female reindeer regularly grow cranial appendages (antlers, the defining characteristics of cervids), as well as males. Moreover, reindeer milk contains more protein and less lactose than bovids' milk. A high quality reference genome of this species will assist efforts to elucidate these and other important features in the reindeer.</p> <p>Findings: We obtained 615 Gb (Gigabase) of usable sequences by filtering the low quality reads of the raw data generated from the Illumina Hiseq 4000 platform, and a 2.64 Gb final assembly, representing 95.7% of the estimated genome (2.76 Gb according to k-mer analysis), including 92.6% of expected genes according to BUSCO analysis. The contig N50 and scaffold N50 sizes were 89.7 kilo base (kb) and 0.94 mega base (Mb), respectively. We annotated 21,555 protein-coding genes and 1.07 Gb of repetitive sequences by de novo and homology-based prediction. Homology-based searches detected 159 rRNA, 547 miRNA, 1,339 snRNA and 863 tRNA sequences in the genome of <i>R. tarandus</i>. The divergence time between <i>R. tarandus</i>, and ancestors of <i>Bos taurus</i> and <i>Capra hircus</i>, is estimated to be about 29.5 million years ago (Mya).</p> <p>Conclusions: Our results provide the first high-quality reference genome for the reindeer, and a valuable resource for studying evolution, domestication and other unusual characteristics of the reindeer.</p>	
<b>Corresponding Author:</b>	Wen Wang, Ph.D CHINA	
<b>Corresponding Author Secondary Information:</b>		
<b>Corresponding Author's Institution:</b>		
<b>Corresponding Author's Secondary Institution:</b>		
<b>First Author:</b>	Zhipeng Li	
<b>First Author Secondary Information:</b>		
<b>Order of Authors:</b>	Zhipeng Li	
	Zeshan Lin	
	Lei Chen	
	Hengxing Ba	
	Yongzhi Yang	

	Kun Wang
	Wen Wang
	Qiang Qiu
	Guangyu Li
<b>Order of Authors Secondary Information:</b>	
<b>Response to Reviewers:</b>	<p>We have carefully revised the manuscript according to the reviewers' comments. The point by point responses to the reviewer's comments are listed below, and our responses are in bold and start with dashes "--".</p> <p>Reviewer reports: Reviewer #1</p> <p>In this Data Note the authors describe the first genome assembly for the reindeer. I am impressed by the amount and variety of analyses undertaken to demonstrate the quality of the assembly. That said, I think there are places in the manuscript where the methods could be more fully explained, and broader context given to the results. Specific comments are below</p> <p>--We greatly thank the reviewer to provide us these positive comments to improve the manuscript, especially in the method and result sections.</p> <p>Line 20: could be fair to mention that the amount of usable sequence was actually 615 Gb (line 66) --We change this sentence to "We obtained 615 Gb (Gigabase) usable sequences by filtering the low quality reads of the raw data generated from Illumina Hiseq 4000 platform" in lines 21-22.</p> <p>Lines 42-45: these two sentences should be re-worded for clarity. --Thanks for your suggestion. In these sentences, we want to address the importance of antlers and the interesting biology of reindeer, which is the only species that females grow antler in the Cervidae. We rewrite these sentences as "Antlers are the defining characteristic of male cervids, belonging to the secondary sexual appendage, which shed and regrow in each year throughout an animal's life. Interestingly, reindeer is the only species that females regularly grow antlers in cervids." in lines 43-46.</p> <p>Line 49: replace "special" with "this" -- Changed as suggested.</p> <p>Table S1: what is the difference between sequence and physical converge? -- Sequence coverage is the average number of times a base is read, physical coverage is the average number of times a base spanned by paired or mate paired reads. We have added the explanation in the note of the table.</p> <p>Lines 69-71: a fuller explanation of the k-mer analysis would be useful. Also, I noted that the distribution in Figure S1 is bimodal. Is this expected? Is it a problem for the analysis? Finally, why not use the traditional c-value estimate of genome size, or at least provide a comparison of the two estimates? -- We have now fully explained the k-mer analysis in lines 70-73 on page 5. The bimodal is common in this kind of analysis. A k-mer is related to an artificial sequence division of K nucleotides extracted iteratively from sequencing reads. We defined the k-mer length as 17 bp; thus, a L bp-long clean sequence would include (L-17 + 1) k-mers. The frequency of each k-mer can be calculated from the genome sequence reads. Typically, k-mer frequencies plotted against the sequence depth gradient follows a Poisson distribution in any given dataset. The k-mer method is regularly used to estimate genome size and heterozogosity in genome projects, and C value method is only used in some less studied taxa with unknown genome size range while the reindeer is one of mammals, whose genome sizes are relatively stable.</p> <p>Lines 87-89: it is stated that the accumulation curves in Figure S2 are similar, but to me it looks like the slope for the reindeer is much steeper and more linear than the other genomes. Are they statistically the same? If the reindeer one is different why might that be? -- The horizontal axis represents the error rate and the vertical axis represents the coverage. The error rate of the reindeer is the lowest at the same coverage, indicating</p>

that the high quality assembly of reindeer genome. We have added the explanation in the legend of Figure S2.

Lines 89-96: why was the goat genome chosen for syteny analyses? Is not the cow genome more complete?

-- The goat genome is generated by the third-generation sequencing technology recently with much longer contigs and higher accuracy compared to other ruminants.

Figure S3: please expand the figure legend so that it contains more information as to what is being shown.

-- Thanks for the suggestion. We have improved Figure S3 and added more explanation: The horizontal and vertical axis represents the chromosomes of goat (*Capra hircus*) and the scaffolds of reindeer (*Rangier tarandus*), respectively. Those red dots indicate the collinear regions of the two genomes.

Table S4: indicate where % corresponds to % of the genome versus % of elements found.

-- The % in Table S4 indicates that the percentage of repeat regions in reindeer genome. Moreover, we checked the results again and corrected some mistakes which are now marked in yellow.

I would suggest moving the reference to Table S6 from the end of Line 128 to the end of the sentence on Line 127. As it stands now when I went to look at the data I was expecting to see a summary of the functions annotated, not a comparison of how the different software's did. That said, a table summarizing the functions annotated would also be interesting.

-- Thanks for the suggestions. As suggested we have added a Gene ontology annotation to indicate distribution of gene functions in Figure S6.

Lines 130-131: state how many variants were found.

-- We added a sentence in lines 135-136 on page 8 "Finally, a total of 3,353,347 SNVs were found in the genome of reindeer (Table S7)."

Lines 151-153: is this divergence time in line with previous estimates? Please provide citations.

-- The estimated divergence time is consistent with the published results (Ref 1 and 2 listed below). We cited these papers in our manuscript in lines 157-158 of pages 9. "This is consistent with the previous findings from both fossil record and molecular phylogeny analysis (Ref 1 and 2)."

Ref 1.dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PCJ and Yang Z. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proceedings of the Royal Society B: Biological Sciences*. 2012;279 1742:3491-500.

Ref 2.Bibi F. A multi-calibrated mitochondrial phylogeny of extant Bovidae (Artiodactyla, Ruminantia) and the importance of the fossil record to systematics. *BMC Evol Biol*. 2013;13 1:166.

#### Reviewer #2

This is an extremely useful paper to those that are interested in farmed ruminants especially cervids. There are some minor typographical errors which are described below and several minor queries whose answers would improve the text. I checked through the ftp site and the annotation information looks very good and useful. I did not however, download and check each file.

--Thank you very much for your positive comments on this work.

#### Comments

##### Major

I am rather confused as nowhere in the text is it described how the assembly scaffolds were ordered and aligned into chromosomes and or genome order. This is important as a number of analyses depend on this aspect. For example the genome comparison with goat (Figure S3V5) and I note the axes of this figure are also cryptically not annotated with either bp or chromosome numbers. I suspect that this did not happen except via homology comparison with another species (sheep or cattle, maybe goat?).

	<p>Why raise this point? Well to me a high quality assembly actually rests on the scaffolds being ordered and orientated based on data like Hi-C, optical mapping, linkage mapping, LD mapping, or radiation hybrids of which there is no mention. This aspect needs to be clarified and described and commented on.</p> <p>--We are sorry for any unclearness in the description of genome assembly. Indeed, this work didn't assemble the reindeer genome to the chromosomal level, but only to the draft (regular scaffold) level. We actually aligned the reindeer scaffolds to the goat genome which was assembled to the chromosomal level (Ref 3 listed below) to evaluate the quality of our draft assembly. So chromosome information in Figure S3 etc refers to those of goat rather than the reindeer. Usually a high quality draft genome assembly is enough for most biological analyses. If the reindeer chromosomal level information is needed in future studies, one indeed has to use Hi-C, optical mapping, or genetic maps generated with methods like radiation hybrids, and even Hi-C and optical mapping usually can only get longer scaffolds rather than complete chromosomes. We have explained more in the Figure S3 legend avoid ambiguity.</p> <p>Ref 3. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017;49 4:643-50.</p> <p>Minor * line 19 "species" rather than "specie"  -- Corrected as suggested.</p> <p>* line 58 The DNA extraction method (and reference) are not described. It is also impossible to call blood frozen and then presumably thawed "fresh blood". This makes me suspect this aspect is unclear to the authors.  -- Thank the reviewer to correct this statement. The DNA was extracted from the thawed blood. We rewrite this sentence to "Genomic DNA was extracted from the sample thawed from frozen blood using the DNeasy Blood &amp; Tissue Kit (QIAGEN, Valencia, CA, USA) according to the manufacturer's instructions." in lines 58-60 on page 5.</p> <p>* Table s1 needs web address for the deposit numbers.  -- We have added the deposit numbers in the Table S1 and their web link.</p> <p>* line 153 I suspect figure 1 = figure S7? Figure 1 the precision in the estimated divergence times is excessive and the legend should be altered to make clear it is a range.  -- Thank for pointing out this. We have now changed "Figure1" to "Figure S7" in line 157 of page 9. And we used a range rather than a concrete number for the time divergence. We have rewritten the legend of Figure 1 making it clearer.</p> <p>* line 168 "he" should be "the".  -- Sorry for the typo, we have corrected it.</p> <p>* line 172 "libraries" should be "library".  -- Corrected as suggested.</p>
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<b>Experimental design and statistics</b>	Yes
Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our <a href="#">Minimum Standards Reporting Checklist</a> . Information essential to interpreting the data presented should be made available in the figure legends.	

<p>Have you included all the information requested in your manuscript?</p>	
<p><b>Resources</b></p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite <a href="#">Research Resource Identifiers</a> (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	<p>Yes</p>
<p><b>Availability of data and materials</b></p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in <a href="#">publicly available repositories</a> (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	<p>Yes</p>





1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

46 **Keywords:** *Rangier tarandus*, Reindeer, Caribou, genomics, whole genome

47 sequencing, assembly, annotation



## Background information

The Cervidae is the second largest family in the suborder Ruminantia of the Artiodactyla, which are distributed across much of the globe in diverse habitats, from arctic tundra to tropical forests [1, 2]. Reindeer or Caribou (*Rangifer tarandus*, NCBI Taxon ID: 9870) is the only species with a circumpolar distribution (present in boreal, tundra, subarctic, arctic and mountainous regions of northern Asia, North America and Europe). It is also the only cervid having been fully domesticated, although some other species have been attempted, such as the sika deer (*Cervus nippon*), which has been semi-domesticated for more than 200 years and still has strong wild nature. Antlers are the defining characteristic of male cervids, belonging to the secondary sexual appendage, which shed and regrow in each year throughout an animal's life. Interestingly, reindeer is the only cervid species that females regularly grow antlers (**Figure 1**). Furthermore, reindeer milk contains greater amount of proteins, and lower amount of lactose compared to that of bovids [3]. Here, we report a high-quality reindeer reference genome using material from a Chinese individual, which will be useful in elucidating special characteristics of this cervid.

## Data description

### Animal and sample collecting

Fresh blood was collected from a two-year-old, female reindeer of a domesticated herd maintained by Ewenki (also know as Evenks) hunter-herders in the

1 67 Greater Khingan Mountains, Inner Mongolia Autonomous Region, China (50.77° N,  
2  
3 68 121.47° E). The sample was immediately placed in liquid nitrogen, and was then  
4  
5  
6 69 stored at -80°C for later analysis.  
7  
8  
9

## 10 70 **Library construction, sequencing and filtering**

11  
12  
13

14 71 Genomic DNA was extracted from the sample thawed from frozen blood using  
15  
16  
17 72 the DNeasy Blood & Tissue Kit (QIAGEN, Valencia, CA, USA) according to the  
18  
19  
20 73 manufacturer's instructions. Isolated genomic DNA was then used to construct five  
21  
22 74 short-insert libraries (200, 250, 350, 400 and 450 base pair, bp) and four long-insert  
23  
24  
25 75 libraries (3, 6.5, 11.5 and 16 kb) following standard protocols provided by Illumina.  
26  
27  
28 76 Then, 150 bp paired-end sequencing was performed to generate 723.2 Gb of raw data,  
29  
30  
31 77 using a whole genome shotgun sequencing strategy on the Illumina Hiseq 4000  
32  
33 78 platform (**Table S1**). To improve the read quality, we trimmed low-quality bases from  
34  
35  
36 79 both sides of the reads and removed reads with more than 5% of uncalled ("N") bases.  
37  
38  
39 80 Then reads of all libraries were corrected by SOAPec (version 2.03) [4]. Finally, clean  
40  
41  
42 81 reads amounting to 615 Gb were obtained for genome assembly.  
43  
44  
45

## 46 82 **Evaluation of genome size**

47  
48  
49

50 83 The estimated genome size is 2.76 Gb according to k-mer analysis, based on the  
51  
52 84 following formula:  $G = N * (L - 17 + 1) / K\_depth$  (**Figure S1**), where N is the total  
53  
54  
55 85 number of reads, and K\_depth is the frequency of reads occurring more often than  
56  
57  
58 86 others [5]. All the clean reads provide approximately ~ 220-fold mean coverage.  
59  
60  
61  
62  
63  
64  
65

1       87    **Genome assembly**

2  
3  
4       88        We used SOAPdenovo (SOAPdenovo2 , RRID:SCR\_014986)(version 2.04) with  
5  
6  
7       89        optimized parameters (pregraph -K 79 -d 0; map -k 79; scaff -L 200) to construct  
8  
9  
10      90        contigs and original scaffolds [5]. All reads were aligned onto contigs for scaffold  
11  
12  
13      91        construction by utilizing the paired-end information. Gaps were filled using reads  
14  
15  
16      92        from three libraries (200, 250 and 350 bp) with GapCloser (GapCloser,  
17  
18      93        RRID:SCR\_015026)(version 1.12) [6]. The final reindeer genome assembly is 2.64  
19  
20  
21      94        Gb long, including 95.7 Mb (3.6%) of unknown bases, smaller than that of the  
22  
23  
24      95        domestic goat (*Capra hircus*, 2.92 Gb) [7], and similar to that of sheep (*Ovis aries*,  
25  
26      96        2.61 Gb) [8]. The contig N50 (> 200 bp) and scaffold N50 (> 500 bp) sizes are 89.7  
27  
28  
29      97        kb and 0.94 Mb, respectively (**Table 1**).

30  
31  
32  
33      98        **Quality assessments**

34  
35  
36  
37      99        We used BUSCO (benchmarking universal single-copy orthologs, version 2.0)  
38  
39  
40     100        software to assess the genome completeness (BUSCO , RRID:SCR\_015008)[9]. Our  
41  
42  
43     101        assembly covered 92.6% of the core genes, with 3,803 genes being complete (**Table**  
44  
45     102        **S2**). Feature-response curve (FRC, version 1.3.1) method [10] was then used to  
46  
47  
48     103        evaluate the trade-off between the assembly's contiguity and correctness. The results  
49  
50  
51     104        indicate that it has a similar accumulated curve compared to published high quality  
52  
53  
54     105        assemblies for other ruminant genomes including cattle, goat, and sheep (**Figure S2**).  
55  
56  
57     106        Subsequently, synteny analysis was applied to identify differences between the  
58  
59  
60     107        assembled genome and the domestic goat (*Capra hircus*) genome (**Figure S3**). 83.95%

1 108 of two genome sequences could be 1:1 aligned, and the average nuclear distance  
2  
3  
4 109 (percentage of different base pairs in the syntenic regions) was 7.18% (**Figure S4**). In  
5  
6 110 addition, the density of different types of break points (edges of structural variation)  
7  
8  
9 111 are about 69.88 per Mb (**Table S3**). These results suggest that the reindeer genome  
10  
11  
12 112 assembly is of a good level of contiguity and correctness.

### 113 **Genome annotation**

114 To annotate the reindeer genome we initially used LTR\_FINDER (LTR\_Finder,  
115 RRID:SCR\_015247)[11] and RepeatModeller (RepeatModeler,  
116 RRID:SCR\_015027)(version 1.0.4;  
117 <http://www.repeatmasker.org/RepeatModeler.html>) to find repeats. Next,  
118 RepeatMasker (version 4.0.5) [12] was used (with -nolow -no\_is -norna -parallel 1  
119 parameters) to search for known and novel transposable elements (TE) by mapping  
120 sequences against the *de novo* repeat library and Repbase TE library (version 16.02)  
121 [13]. Subsequently, tandem repeats were annotated using Tandem Repeat Finder  
122 (version 4.07b; with 2 7 7 80 10 50 2000 -d -h parameters) [14]. In addition, we used  
123 RepeatProteinMask software [12] with -no LowSimple -p value 0.0001 parameters to  
124 identify TE-relevant proteins. The combined results indicate that repeat sequences  
125 cover about 1.03 Gb, accounting for 39.1% of the reindeer genome assembly (**Table**  
126 **S4**).

127 The rest of the reindeer genome assembly was annotated using both *de novo* and  
128 homology-based gene prediction approaches. For *de novo* gene prediction, we utilized

1 129 SNAP (version 2006-07-28), GenScan (GENSCAN , RRID:SCR\_012902)[15],  
2  
3 130 glimmerHMM (GlimmerHMM, RRID:SCR\_002654) and Augustus (Augustus: Gene  
4  
5  
6 131 Prediction, RRID:SCR\_008417) (version 2.5.5) [16] to analyze the repeat-masked  
7  
8  
9 132 genome. For homology-based predictions, sequences encoding homologous proteins  
10  
11 133 of *Bos taurus* (Ensemble 87 release), *Ovis aries* (Ensemble 87 release) and *Homo*  
12  
13 134 *sapiens* (Ensemble 87 release), were aligned to the reindeer genome using TblastN  
14  
15 135 (TBLASTN, RRID:SCR\_011822)(version 2.2.26) with an (E)-value cutoff of 1 e-5.  
16  
17 136 Genwise (version wise2.2.0) [17] was then used to annotate structures of the genes.  
18  
19  
20 137 The *de novo* and homology gene sets were merged to form a comprehensive,  
21  
22 138 non-redundant gene set using EVidenceModeler software (EVM, version 1.1.1),  
23  
24 139 which resulted in 21,555 protein-coding genes (**Table S5**). We then compared the  
25  
26 140 reindeer genome with species which were used in homology prediction, and there was  
27  
28 141 no significant difference among the four species in gene length and exon length  
29  
30 142 distribution (**Figure S5**).

31  
32  
33  
34  
35  
36  
37  
38  
39  
40 143 Next, we searched the KEGG, TrEMBL and SwissProt databases for best  
41  
42 144 matches to the protein sequences yielded by EVM software, using BLASTP (version  
43  
44 145 2.2.26) with an (E)-value cutoff of 1 e-5, and searched Pfam, PRINTS, ProDom and  
45  
46 146 SMART databases for known motifs and domains in our sequences using  
47  
48 147 InterProScan software (InterProScan , RRID:SCR\_005829)(version 5.18-57.0)[18].  
49  
50  
51 148 At least one function was assigned to 19,004 (88.17%) of the detected reindeer genes  
52  
53 149 through these procedures (**Table S6**). Of them, 14,138 genes were used to do the gene  
54  
55 150 ontology annotation (**Figure S6**). The reads from short-insert length libraries then

1 151 were mapped to the reindeer genome with BWA (BWA, RRID:SCR\_010910)(version  
2  
3 152 0.7.12-r1039) [19], then single nucleotide variants (SNVs) were called by SAMtools  
4  
5  
6 153 (SAMTOOLS, RRID:SCR\_002105)(version 1.3.1) [20]. Finally, we performed  
7  
8  
9 154 SnpEff (version 4.30) [21] to identify the distribution of SNV in the reindeer genome.  
10  
11  
12 155 Finally, a total of 3,353,347 SNVs were found in the genome of reindeer (**Table S7**).

13  
14  
15  
16 156 In addition, we predicted rRNA-coding sequences based on homology with  
17  
18 157 human rRNAs using BLASTN with default parameters (BLASTN,  
19  
20  
21 158 RRID:SCR\_001598). To annotate miRNA and snRNA genes we searched the Rfam  
22  
23  
24 159 database (release 9.1) with Infernal (Infernal, RRID:SCR\_011809)(version 0.81)[22],  
25  
26  
27 160 and annotated tRNAs using tRNAscan-SE (version 1.3.1) software with default  
28  
29  
30 161 parameters (tRNAscan-SE, RRID:SCR\_010835)[23]. The final results identified 159  
31  
32 162 rRNAs, 547 miRNAs, 1,339 snRNAs and 863 tRNAs (**Table S8**).

### 33 34 35 36 163 **Species-specific genes and phylogenetic relationship**

37  
38  
39  
40 164 We clustered the detected reindeer genes in families by using OrthoMCL  
41  
42  
43 165 (OrthoMCL DB: Ortholog Groups of Protein Sequences, RRID:SCR\_007839) [24]  
44  
45  
46 166 with an (E)-value cutoff of  $1 \times 10^{-5}$ , and a Markov Chain Clustering with default  
47  
48  
49 167 inflation parameter in an all-to-all BLASTP analysis of entries for five species (*Homo*  
50  
51 168 *sapiens*, *Equus caballus*, *Capra hircus*, *Bos taurus*, and *Rangifer tarandus*). The  
52  
53  
54 169 result showed that 335 gene families were specific to the reindeer (**Figure S7**).  
55  
56  
57 170 Moreover, we identified 7,505 single-copy gene families from these species and  
58  
59  
60 171 aligned coding sequences in the families using PRANK (version 3.8.31) [25].

1 172 Subsequently, 4D-sites (four-fold degenerated sites) were extracted to construct a  
2  
3  
4 173 phylogenetic tree by RAxML (RAxML, RRID:SCR\_006086)(version 7.2.8) [26] with  
5  
6 174 GTR+G+I model. Finally, phylogenetic analysis using PAML MCMCtree (version  
7  
8  
9 175 4.5) (PAML, RRID:SCR\_014932)[27], calibrated with published timings of the  
10  
11  
12 176 divergence of the reference species (<http://www.timetree.org/>)[28], indicated that  
13  
14 177 *Rangifer tarandus*, *Bos taurus* and *Capra hircus* diverged from a common ancestor  
15  
16  
17 178 approximately 29.5 (25.41-31.75) Mya (**Figure 2**). This is consistent with the  
18  
19  
20 179 previous findings from both fossil record and molecular phylogeny analysis [29, 30].  
21  
22  
23

## 24 180 **Conclusion**

25  
26  
27  
28 181 In summary, we report the first sequencing, assembly and annotation of the  
29  
30  
31 182 reindeer genome, which will be useful in analysis of the genetic basis of the unique  
32  
33  
34 183 characteristics of reindeer, and broader studies on ruminants.  
35  
36

## 37 184 **Availability of supporting data**

38  
39  
40  
41 185 The raw sequence data have been deposited in the Short Read Archive (SRA)  
42  
43  
44 186 under accession numbers SRR5763125-SRR5763133. Assemblies, annotations and  
45  
46  
47 187 other supporting data are also available from the *GigaScience* database[31].  
48  
49  
50

## 51 188 **Abbreviations**

52  
53  
54  
55 189 Gb: giga base; bp: base pair; kb: kilo base; Mb: mega base; TE: transposable  
56  
57  
58 190 element; EVM: EVIDENCEModeler; BUSCO: benchmarking universal single-copy  
59  
60  
61  
62  
63  
64  
65

1 191 orthologs; FRC: feature-response curves; SNV: single nucleotide variant; MYA:  
2  
3 192 million years ago  
4  
5  
6

7 193 **Acknowledgements**  
8  
9

10  
11 194 This work was supported by the Natural Science Foundation of China (No.  
12  
13 195 31501984) and Central Public-interest Scientific Institution Basal Research Fund (No.  
14  
15 196 1610342016026) to ZPL, and Talents Team Construction Fund of Northwestern  
16  
17 197 Polytechnical University (NWPU) to QQ and WW. Special thanks to Nowbio Biotech  
18  
19 198 Inc., Kunming, China for its remarkable work on DNA library constructions and  
20  
21 199 sequencing.  
22  
23  
24  
25  
26

27  
28  
29 200 **Competing interests**  
30  
31

32  
33 201 The authors declare that they have no competing interests.  
34  
35  
36

37 202 **Authors' contributions**  
38  
39  
40

41 203 ZPL collected the samples; ZSL, CL ZPL, YZY, KW and HXB analyzed the data;  
42  
43 204 ZSL, QQ and ZPL wrote the manuscript; WW and GYL conceived the study.  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65



1 205 **References**

- 2  
3 206 1. Fernández MH and Vrba ES. A complete estimate of the phylogenetic  
4 207 relationships in ruminantia: a dated species-level supertree of the extant  
5 208 ruminants. *Biological Reviews*. 2005;80 2:269-302.  
6  
7 209 2. Hassanin A, Delsuc F, Ropiquet A, Hammer C, Jansen van Vuuren B, Matthee  
8 210 C, et al. Pattern and timing of diversification of Cetartiodactyla (Mammalia,  
9 211 Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial  
10 212 genomes. *C R Biol*. 2012;335 1:32-50.  
11 213 3. Young W. Park GF. *Handbook of milk of non-bovine mammals*.  
12 214 Wiley-Blackwell; 2006. ISBN: 978-0-8138-2051-4  
13  
14 215 4. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an  
15 216 empirically improved memory-efficient short-read de novo assembler.  
16 217 *GigaScience*. 2012;1 1:1-6.  
17  
18 218 5. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and de novo  
19 219 assembly of the giant panda genome. *Nature*. 2010;463 7279:311-7.  
20  
21 220 6. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of  
22 221 human genomes with massively parallel short read sequencing. *Genome Res*.  
23 222 2010;20 2:265-72.  
24  
25 223 7. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al.  
26 224 Single-molecule sequencing and chromatin conformation capture enable de  
27 225 novo reference assembly of the domestic goat genome. *Nat Genet*. 2017;49  
28 226 4:643-50.  
29  
30  
31 227 8. Jiang Y, Xie M, Chen W, Talbot R, Maddox JF, Faraut T, et al. The sheep  
32 228 genome illuminates biology of the rumen and lipid metabolism. *Science*.  
33 229 2014;344 6188:1168-73.  
34  
35 230 9. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM.  
36 231 BUSCO: assessing genome assembly and annotation completeness with  
37 232 single-copy orthologs. *Bioinformatics*. 2015;31 19:3210-2.  
38  
39 233 10. Vezzi F, Narzisi G and Mishra B. Reevaluating assembly evaluations with  
40 234 feature response curves: GAGE and assemblathons. *PLoS ONE*. 2012;7  
41 235 12:e52210.  
42  
43 236 11. Xu Z and Wang H. LTR\_FINDER: an efficient tool for the prediction of  
44 237 full-length LTR retrotransposons. *Nucleic Acids Res*. 2007;35  
45 238 suppl\_2:W265-W8.  
46  
47 239 12. Tarailo-Graovac M and Chen N. Using RepeatMasker to identify repetitive  
48 240 elements in genomic sequences. *Current Protocols in Bioinformatics*. John  
49 241 Wiley & Sons, Inc.; 2009.  
50  
51 242 13. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O and  
52 243 Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements.  
53 244 *Cytogenet Genome Res*. 2005;110 1-4:462-7.  
54  
55 245 14. Benson G. Tandem repeats finder: a program to analyze DNA sequences.  
56 246 *Nucleic Acids Res*. 1999;27 2:573-80.  
57  
58  
59  
60  
61  
62  
63  
64  
65

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65
- 247 15. Burge C and Karlin S. Prediction of complete gene structures in human  
248 genomic DNA1. J Mol Biol. 1997;268 1:78-94.
  - 249 16. Stanke M, Keller O, Gunduz I, Hayes A, Waack S and Morgenstern B.  
250 AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res.  
251 2006;34 suppl\_2:W435-W9.
  - 252 17. Birney E, Clamp M and Durbin R. GeneWise and Genomewise. Genome Res.  
253 2004;14 5:988-95.
  - 254 18. Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li,  
255 Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka,  
256 Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim  
257 Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter (2014).  
258 InterProScan 5: genome-scale protein function classification. Bioinformatics,  
259 Jan 2014; doi:10.1093/bioinformatics/btu031
  - 260 19. Heng L. Aligning sequence reads, clone sequences and assembly contigs with  
261 BWA-MEM. arXiv. 2013;1303.3997
  - 262 20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The  
263 sequence alignment/map format and SAMtools. Bioinformatics. 2009;25  
264 16:2078-9.
  - 265 21. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A  
266 program for annotating and predicting the effects of single nucleotide  
267 polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster*  
268 strain w (1118); iso-2; iso-3. Fly (Austin). 2012;6 2:80-92.
  - 269 22. E. P. Nawrocki and S. R. Eddy, Infernal 1.1: 100-fold faster RNA homology  
270 searches, Bioinformatics 29:2933-2935 (2013).
  - 271 23. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for  
272 analysis of transfer RNA genes. Nucleic Acids Res. 2016 Jul 8;44(W1):W54-7.  
273 doi: 10.1093/nar/gkw413.
  - 274 24. Li L, Stoeckert CJ and Roos DS. OrthoMCL: Identification of ortholog groups  
275 for eukaryotic genomes. Genome Res. 2003;13 9:2178-89.
  - 276 25. Löytynoja A and Goldman N. An algorithm for progressive multiple  
277 alignment of sequences with insertions. Proc Natl Acad Sci U S A. 2005;102  
278 30:10557-62.
  - 279 26. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and  
280 post-analysis of large phylogenies. Bioinformatics. 2014;30 9:1312-3.
  - 281 27. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol  
282 Evol. 2007;24 8:1586-91.
  - 283 28. Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. Tree of life reveals  
284 clock-like speciation and diversification. Mol Biol Evol. 2015  
285 Apr;32(4):835-45. doi: 10.1093/molbev/msv037.
  - 286 29. dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PCJ and Yang Z.  
287 Phylogenomic datasets provide both precision and accuracy in estimating the  
288 timescale of placental mammal phylogeny. Proceedings of the Royal Society  
289 B: Biological Sciences. 2012;279 1742:3491-500.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

290 30. Bibi F. A multi-calibrated mitochondrial phylogeny of extant Bovidae  
291 (Artiodactyla, Ruminantia) and the importance of the fossil record to  
292 systematics. *BMC Evol Biol.* 2013;13 1:166.29.  
293 31. Li, Z; Lin, Z; Chen, L; Ba, H; Yang, Y; Wang, K; Wang, W; Qiang, Q; Li G.  
294 (2017): Draft genomic data of the Reindeer (*Rangifer tarandus*). GigaScience  
295 Database. <http://dx.doi.org/100370>  
296

1 297 **Figure legends**

2  
3  
4 298 **Figure 1. Male (above) and female (below) *Rangier tarandus* individuals, the only**  
5  
6  
7 299 **cervid species that both sexes are able to produce velvet antlers.** Pictures courtesy  
8  
9  
10 300 of Yifeng Yang from the Institute of Special Animal and Plant Sciences, Chinese  
11  
12 301 Academy of Agricultural Sciences.

13  
14  
15  
16 302 **Figure 2. Phylogenetic relationships of *Rangier tarandus* and four species based**  
17  
18 303 **on four-fold degenerated sites.** The blue numbers in the square brackets above the  
19  
20  
21 304 nodes are the 90% confidence interval of divergence time from the present. MYA,  
22  
23  
24 305 million years ago.

25  
26  
27 306  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

307 **Tables**

308 **Table 1 Summary of 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

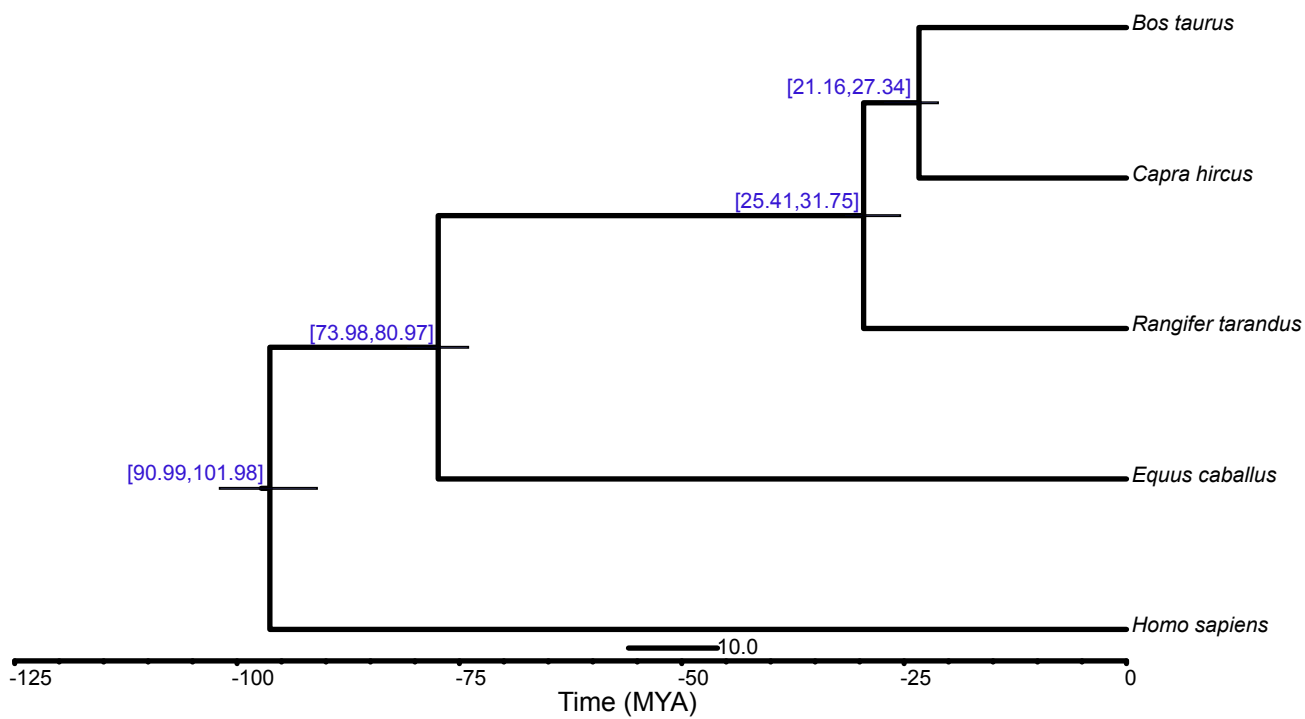
---

309

**Figure 1. Male (above) and female (below) *Rangier tarandus* individuals, the only cervid species that both sexes are able to produce velvet antlers. Pictures courtesy of Yifeng Yang from the Institute of Special Animal and Plant Sciences, Chinese Academy of Agricultural Sciences.**



Figure 2

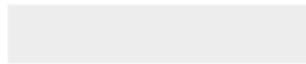





[Click here to access/download](#)

**Supplementary Material**


Supplementary tables\_REVISED-1017.doc







Click here to access/download  
**Supplementary Material**  
Figure S1.pdf




Click here to access/download  
**Supplementary Material**  
Figure S2.pdf




Click here to access/download  
**Supplementary Material**  
Fig.S3V7.pdf

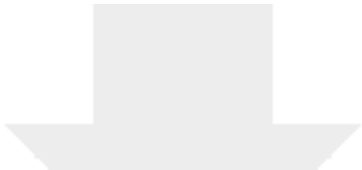





Click here to access/download  
**Supplementary Material**  
Figure S4.pdf

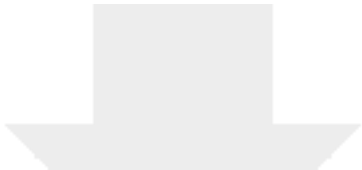


Click here to access/download  
**Supplementary Material**  
Figure S5.pdf




Click here to access/download  
**Supplementary Material**  
Figure S6.pdf





Click here to access/download  
**Supplementary Material**  
Figure S7.pdf



Dear editor of *GigaScience*,

Thank you very much for your consideration on our manuscript (ID **GIGA-D-17-00152**) entitled “Draft genome of the Reindeer (*Rangifer tarandus*)” as a Data Note in *GigaScience*.

We have carefully revised the manuscript according to the reviewers’ comments and your instructions. The revised parts are marked with yellow. The point by point responses to the reviewer’s comments are listed below, and **our responses are in bold and start with dashes “--”**.

We hope that the revised manuscript has addressed all the concerns of reviewers and is now publishable

Yours sincerely,  
Wen Wang, Ph.D