

## A high-quality genome assembly for the endangered golden snub-nosed monkey (*Rhinopithecus roxellana*) --Manuscript Draft--

<b>Manuscript Number:</b>	GIGA-D-19-00030R1	
<b>Full Title:</b>	A high-quality genome assembly for the endangered golden snub-nosed monkey ( <i>Rhinopithecus roxellana</i> )	
<b>Article Type:</b>	Data Note	
<b>Funding Information:</b>	National Natural Science Foundation of China (31622053)	Dr. Xiao-Guang Qi
<b>Abstract:</b>	<p><b>Background:</b> The golden snub-nosed monkey (<i>Rhinopithecus roxellana</i>) is an endangered colobine species endemic to China. This species has several distinctive traits and is an ideal model for analyses of the evolutionary development of social structures due to its unique social organization. Although a genome assembly for the subspecies <i>R. roxellana hubeiensis</i> is available, this assembly is incomplete and fragmented because it was constructed using short read sequencing technology. Thus, information important for the understanding of <i>R. roxellana</i>, such as genome structural variation and repeat sequences, may be absent from the available assembly. Therefore, a high-quality reference genome is needed.</p> <p><b>Findings:</b> To obtain a high-quality chromosomal assembly for <i>R. roxellana qinlingensis</i>, we used five different methods: Pacific Bioscience single-molecule real-time sequencing, Illumina paired-end sequencing, BioNano optical maps, 10X Genomics link-reads, and high-throughput chromosome conformation capture. The assembled genome was ~3.04 Gb, with a contig N50 of 5.72 Mb and a scaffold N50 of 144.56 Mb. This represented a 100-fold improvement over the previously published genome. In the new genome, 22,497 protein-coding genes were predicted, of which 22,053 were functionally annotated. Gene family analysis showed that 993 and 2,745 gene families were expanded and contracted, respectively, in the <i>R. r. qinlingensis</i> genome. The reconstructed phylogeny recovered a close relationship between <i>Rhinopithecus roxellana</i> and <i>Macaca mulatta</i>, and these two species diverged approximately 13.4 MYA.</p> <p><b>Conclusion:</b> We constructed a high-quality genome assembly of Qinling golden snub-nosed monkey; this genome had superior continuity and accuracy, which might be useful as reference for future genetic studies in this species. In addition, the updated genome assembly might improve our understanding of this species and might be particularly relevant to conservation efforts. Furthermore, this high-quality genome might serve as a new standard reference genome for colobine primates.</p>	
<b>Corresponding Author:</b>	Xiao-Guang Qi  CHINA	
<b>Corresponding Author Secondary Information:</b>		
<b>Corresponding Author's Institution:</b>		
<b>Corresponding Author's Secondary Institution:</b>		
<b>First Author:</b>	Xiao-Guang Qi	
<b>First Author Secondary Information:</b>		
<b>Order of Authors:</b>	Xiao-Guang Qi	
	Lu Wang	
	Jinwei Wu	
	Xiaomei Liu	

	Dandan Di
	Yuhong Liang
	Yifei Feng
	Suyun Zhang
	Baoguo Li
<b>Order of Authors Secondary Information:</b>	
<b>Response to Reviewers:</b>	<p>Editor Comments to Author:</p> <p>In particular major improvements are required in the writing and we would strongly recommend you use a native English speaker or professional company to improve the writing.</p> <p>Response to comment 1 Thanks for this comment, the manuscript has been revised and polished by an English language editing service of LetPub.</p> <p>We have strong policies regarding reproducibility and agree with the referees that significant additional methodological detail is required.</p> <p>Response to comment 2 Thanks for your comment, we added the methodological details substantially to be clear and straightforward. In addition, we added some key details about the generated data (sequencing, calibration times, N50 length et al.) and performed several additional analyses including CNVs identification, synteny analysis and SNP calling et al. as you and reviewers suggested.</p> <p>On top of including detail on the software versions and setting, we would strongly recommend you capture this detail using protocols.io. You can re-use and adapt the protocols we have stored in our group page or create your own (and if you provide these in stepwise manner we can even upload them for you):</p> <p>Response to comment 3 Following this comment, we have captured those methodological details using protocols.io. with our own account. Please check out on the website of "<a href="https://www.protocols.io/private/E AFC44C786ABCE2257FD0D5B9E0D7EF3">https://www.protocols.io/private/E AFC44C786ABCE2257FD0D5B9E0D7EF3</a>".</p> <p>Reviewer Comments to Author:</p> <p>Reviewer #1: This manuscript reports a new whole genome assembly for an interesting nonhuman primate species, <i>Rhinopithecus roxellana</i>. This is a colobine species that has a number of unusual characteristics, including but not limited to unusual pelage, highly derived facial morphology, and social organization that is not entirely unique but is rare among Old World monkeys or other anthropoids. There are five species in the genus, and all are threatened or endangered, so there is a conservation benefit to this genome sequencing as well as basic comparative primate evolutionary genomics. There is a previously published whole genome assembly for this species, but this new assembly is a significant improvement (see below). Consequently, there are several elements of this work that make it noteworthy.</p> <p>The new assembly is based on an effective and technically advanced combination of approaches. The authors began by sequencing this genome using PacBio Sequel long reads, and assembling them using FALCON and PBjelly. The authors generated Illumina short reads and polished the PacBio/FALCON assembly with those. The authors also make use of BioNano optical mapping and 10X linked-reads to increase completeness and contiguity. Finally, Hi-C mapping is used to produce near full chromosome length scaffolds. The result is a 3.04 gigabase assembly with contig N50 of 5.72 Mb and scaffold N50 of 144.56 Mb. These statistics make this one of the most complete and highly contiguous assemblies available for any nonhuman primate (confirmed using BUSCO and CEGMA analyses). The authors then annotated this genome using a series of annotation software tools, and identified 22,497 genes.</p>

This new genome assembly is a valuable resource for any investigator working on the genetics or genomics of *Rhinopithecus*. In addition, this is a high quality - high contiguity assembly, so it will be useful for laboratories working on other closely related colobines. Lastly, the authors report some initial analyses of repetitive sequences and gene family expansions and contractions using this new *Rhinopithecus* assembly.

While this genome sequence seems to be a valuable resource for the primate genomics community, this manuscript has a significant number of serious flaws and problems. One issue is that the quality of the grammar and text is not adequate. I realize that the authors may not be native speakers of English, and that this can be a challenge. But this manuscript needs major assistance in terms of editing before it is ready for serious consideration.

Response to this comment

Thanks for this comment, the manuscript has been revised and polished by an English language editing service of LetPub.

I have other specific concerns as well.

1) This is minor but having two different line numbering systems printed on the same pages causes confusion. I will use the numbers that are actually tied to specific lines in the text, rather than the more densely packed numbers that seem to just run down each page. The authors should delete the dense numbers.

Response to comment 1

Thanks for your kindly review. We deleted the dense numbers.

2) Page 4, lines 54-56. While the social organization of *Rhinopithecus roxellana* is interesting and deserves more study, it seems overly optimistic for the authors to argue that production of this genome assembly will ultimately support genetic studies that make contributions to our "...understanding the behavior patterns of human society in social-anthropology." Studies of comparative social relationships and social organization are important and primates can provide information about human evolution. But this statement seems to me to be overly ambitious in terms of research outcomes.

Response to comment 2

Thanks for this comment. We changed the statement as follows:

"Therefore, *R. roxellana* is an ideal model for the analysis of social structure evolution in primates and may also provide opportunities to investigate evolutionary and socio-anthropological patterns of human society."

3) There are mistakes in capitalization and spelling of words. For example, the Shennongjia Mountains are not capitalized in line 63, but "Gorillas" is incorrectly capitalized in line 75 and "Colobine" is regularly capitalized when it need not be. "Quiver" is misspelled in line 125.

Response to comment 3

Corrected. We also checked other mistakes in capitalization and spelling of words throughout the manuscript.

4) Line 75 states the gorillas and orangutans "...have the closest genetic relationship with humans" but of course that is chimpanzees and bonobos, not gorillas and/or orangutans.

Response to comment 4

Thanks for this comment. We are sorry that we made a mistake here and we changed the statement as follows:

"New sequencing technologies, including Pacific Bioscience's single-molecule real-time (SMRT) sequencing, BioNano optical mapping, and Hi-C-based chromatin interaction maps, have been used in several species closely related to humans, including gorillas (*Gorilla gorilla gorilla*) [17], chimpanzees (*Pan troglodytes*) [18], and Sumatran orangutans (*Pongo abelii*) [18], as well as in other species, including the domestic goat (*Capra hircus*) [19]."

5) I think the language in line 86 is a bit too optimistic and ambitious. The authors state that this assembly "...may allow us to comprehensively understand *R. roxellana*...". I do not know what it would mean to "comprehensively understand" a primate species, but I do not think we are yet close to that point.

Response to comment 5

Thanks for this comment. We changed the statement as follows:  
"This updated genome assembly may allow us to further investigate *R. roxellana*, providing new opportunities to analyze evolutionary history and to identify genetic changes associated with the development of specific traits in this species".

6) Page 6, line 87: It is not clear to me what the authors mean by "genetic-specific signatures of this species"?

Response to comment 6

Thanks for this valuable comment. In fact, we were intended to term those genetic changes associated with the development of species-specific traits as "genetic-specific signatures of this species". We realized that this sentence was confusing and not clear enough. We changed the statement as follows:

"genetic changes associated with the development of specific traits in this species".

7) Page 6, line 93-94. Was the animal used to produce the DNA for the sequencing wild-caught or captive bred at Louguantai? If captive bred, were the parents wild-caught?

Response to comment 7

Thanks for this comment. The animal used for the sequencing was an adult male *R. roxellana qinlingensis* in Qinling Mountain. The animal that died naturally in Qinling Mountain was immediately stored in ultra-cold storage freezer at Louguantai Breeding Centre. We reworded the statement as follows:

"The animal used for sequencing was an adult male *R. r. qinlingensis* from Qinling Mountain, who died naturally and the dead body was stored in ultra-cold storage freezer at Louguantai Breeding Centre, Xi'an, Shaanxi Province, China."

8) Page 7, lines 103-105. BioNano optical mapping is a technique for using restriction enzymes to nick and label DNA at short known target sequences. The map of nicked sites is used to scaffold a genome or confirm the organization of contigs. It is not clear what the authors mean when they state that they "...acquired 463.75 Gb clean reads" from the BioNano Genomics Irys platform. There are no sequence reads generated by the Irys platform. This section does not make sense to me. Instead, the authors should present the actual results of the optical mapping in terms of the number of sites examined and the concordance between the observed BioNano map and the predicted map based on the assembled contigs and scaffolds.

Response to comment 8

Thanks for your valuable comments. We are sorry that we used the wrong term here. Of course, there are no sequence reads generated by the Irys platform, the generation by which should be large DNA molecules. As for the number of sites examined in this study, the average label density for the BioNano map is 11.66 per 100 kb, while the average label density is 12.62 per 100 kb for the predicted map based on those assembled contigs and scaffolds. Thus, the observed BioNano map is consistent with the predicted map. We added several sentences to clarify this point.

"The average label density examined for the BioNano map is 11.66 per 100 kb, while the average label density is 12.62 per 100 kb for the predicted map based on the assembled contigs. Thus, the observed BioNano map is consistent with the predicted map. The BioNano map generated 463.75 Gb of large DNA molecules."

9) I do not think that Figure 2 adds much to this paper. The authors used Hi-C for scaffolding, and that does provide useful data. But simply inserting a figure showing Hi-C interaction frequencies without doing any further analysis of the details of DNA-DNA interaction or characterizing the topologically associating domains provides no significant new information or insight.

Response to comment 9

Thanks for your valuable comment. The fig. 2 was based on the interaction frequencies between pairs of 100-kb genomic regions. In principle, higher counts indicate increased frequency of chromatin interaction and closer spatial distance between the two sequences, darker red means stronger interaction strength. This strategy has significantly advanced the assembly quality with chromosome-length scaffolds. The fig. 2 presented here was used to indicate the reliability of our assembly.

As for the further analysis of the details of DNA-DNA interaction or characterizing the topologically associating domains, we agree that these analysis were useful. However, they may be beyond the scope of this report, which aims to report a high-quality genome for further studies. We also added several sentences to make the figure

legend of fig. 2 more clear.

"Hi-C interactions within and among chromosomes of *R. roxellana* chromosomes (Chr1–Chr22); interactions were drawn based on the chromatin interaction frequencies between pairs of 100-kb genomic regions (as determined by Hi-C). In principle, darker red cells indicate stronger and more frequent interactions, which in turn imply that the two sequences are spatially close."

10) Page 9, lines 151-152. I do not understand the sentence that begins "With a ratio number..."

Response to comment 10

Thanks for this comment. We reworded this sentence to clarify this point.

"Approximately 99.17% of the short reads were mapped to the genome assembly.

Further investigations indicated that these reads covered approximately 99.27% of the total assembly (Supplementary Table S6)."

11) Page 9-10, lines 152-159. Using BUSCO and CEGMA to assess the completeness of the genome assembly is a very good idea. But the authors should report not just how many BUSCO or CEGMA genes were identified, but how many were complete and unfragmented and how many were complete and fragmented.

Response to comment 11

Thanks for this comment. During the BUSCO analysis, the annotation results were classified as complete BUSCOs, fragmented BUSCOs and missing BUSCOs. We did not report those results in the manuscript, however, these details were shown in Supplementary Table S8. Simply, the complete BUSCOs occupied a proportion of 94.0%, while the fragmented BUSCOs occupied only 2.9%. In addition, we added the CEGMA results in Supplementary Table S9, which showed that the 220 genes were complete and unfragmented, while 13 was complete and fragmented. We also added these results in our manuscript.

"In addition, we estimated assembly completeness using Benchmarking Universal Single-copy Orthologs (BUSCO) v3.0.2 [27], with the parameters "-i -o -l -m genome -f -t." based on mammalia\_odb9 (creation date: 2016-02-13; number of species: 50; number of BUSCOs: 4,104). BUSCO analysis identified 4,104 mammalian BUSCOs in the newly assembled *R. roxellana* genome: 94.0% complete BUSCOs, 2.9% fragmented BUSCOs, and 3.1% missing BUSCOs (Supplementary Table S8). Assembly completeness was measured using the core eukaryotic gene (CEG)-mapping approach (CEGMA v2.5) [28]. Of the 248 CEGs known from six model species, 93.95% (233 of 248) were identified in our new genome assembly. Of these, 220 CEGs were complete and unfragmented, and the remaining 13 were complete but fragmented (Supplementary Table S9). Together, these analyses indicated that our new genome assembly was highly accurate and complete."

12) Page 12, lines 226-227. What fossil calibration times were used?

Response to comment 12

Thanks for this comment. The fossil calibration times were derived from Timetree (<http://www.timetree.org/>). The following calibration times were used: Homo sapiens VS Callithrix jacchus (40.6-45.7 MYA); Homo sapiens VS Pan troglodytes (6.2~7 MYA); Homo sapiens VS Mus musculus (85-94 MYA) and Homo sapiens VS Tarsius syrichta (71~77 MYA). We also added these fossil calibration times in our manuscript. "The following fossil calibrations were used: Homo sapiens vs. Callithrix jacchus (40.6–45.7 MYA, million years ago); Homo sapiens vs. Pan troglodytes (~6.2–7 MYA); Homo sapiens vs. Mus musculus (85–94 MYA); and Homo sapiens vs. Tarsius syrichta (~71–77 MYA)."

13) Page 14, lines 232-235. Rhinopithecus gene families were expanded or contracted compared to what taxa? Compared to human? Compared to the ancestral primate genome? Compared to an Old World monkey outgroup?

Response to comment 13

Thanks for this comment. The expansion and contraction of gene families of *Rhinopithecus roxellana* were estimated by comparing those of the most recent common ancestor between *Rhinopithecus roxellana* and *Macaca mulatta*. We added one sentence in the figure legend of fig. 4 to clarify this point.

"Numbers under each species indicate the number of gene families that have been expanded (green) and contracted (light yellow) since the split of species from the most recent common ancestor (MRCA)."

14) I think two column headings in Table 3 are switched. I doubt that the average intron length for the *Rhinopithecus Augustus* gene models is 196bp, while the average exon length for the same gene models is 5,112bp. Seems to me those two labels are probably switched.

Response to comment 14

Thanks for your valuable comment. We are sorry that we made a mistake here, we put them in right order now. Please see Table 3 for details.

Reviewer #2:

The authors present an assembly of golden snub-nosed monkey using a range of sequencing technologies, including long read sequencing. Overall the manuscript is mostly clear to follow and the assembly approaches are standard and appear to be well performed. A very large amount of data was generated, although the methods are very short and some details are lacking, it appears that standard and appropriate assembly approaches were used. Some key details about the generated data are missing, and there are some additional analyses that, if completed, would greatly improve the manuscript.

Response to comment 1

Thanks for your valuable comment. we added the methodological details substantially to be clear and straightforward. Please see the "De novo assembly" section for details. In addition, some key details about the generated data (N50 length, software parameters et al.) were present this time and several additional analyses including CNVs identification, synteny analysis and SNP calling et al. were also performed as suggested.

I could not find descriptions of the characteristics of the generated data, particularly average/n50 length of Pacbio reads, molecule size of the optical mapping and of 10X data. These are key parameters that should be reported.

Response to comment 2

Thanks for this comment. The average/N50 length of Pacbio reads and molecule size of the optical mapping was 16.69 kb and 338 kb, respectively. As for the 10X data, since paired-end of 350 bp sequencing was performed, N50 length was not applicable for this case. It was estimated that a total of 423.32 Gb clean reads were generated for 10X data. We added one sentence to describe the characteristics of the generated data.

"...., the average/N50 length of Pacbio reads was 16.69 kb."

"The average/N50 length of the molecules used for optical mapping was 338 kb."

Line 104 The description of the Bionano data should be clarified. I am not sure that "reads" is the right term for data from this optical mapping platform. Same for term 'sequence coverage' for optical mapping data in Table 1.

Response to comment 3

Thanks for this comment. We added several sentences to detail the BioNano data. We agree that no reads generated from optical mapping platform and we changed the term of "reads" as molecules. Also, the term "sequence coverage" was not an proper term for optical mapping data, we removed the sequence coverage value of optical mapping data in Table 1.

"The average/N50 length of the molecules used for optical mapping was 338 kb. The average BioNano optimal marker density was 11.66 per 100 kb, while the average marker density was 12.62 per 100 kb for the predicted map based on the assembled scaffolds. Thus, the observed BioNano map was consistent with the predicted map. The BioNano map generated 463.75 Gb of large DNA molecules."

The manuscript would benefit from some comparison of how much better the gene annotation is relative to previous assembly, but this and other biological/comparative analyses may be beyond the scope of this report.

Response to comment 4

Thanks for this comment. As for the gene annotation, our new assembly was better than previous assembly from at least two aspects. Firstly, we assessed genome assembly completeness by mapping transcriptome unigenes to the two assembly versions using BLAT v.36. Results showed that the completeness degree (percentage

of unigenes aligned to a single scaffold in genome) was higher in our assembly (95.35%) compared with that in previous assembly (89.28%) for unigenes larger than 1000 bp (Supplementary Table S15), demonstrating the contiguity of our new assembly. Secondly, the number of genes annotated to the public database to the total number of predicted genes was higher in our new assembly (98.03%) than that in previous version (94.52%).

From Supplementary Tables S2-3, it seems that the largest increase in n50 scaffold length came from 10X linked read data, not from the bionano optical map. I do not think this is expected, given that optical map data should provide very long range information. The manuscript would be clearer for the reader if some description for why such a gain was found from 10X data was described, and if such results are typical.

Response to comment 5

Thanks for this valuable comment. We checked our assembly description carefully and found some details were not shown. Actually, the first stage of assembly was conducted mainly from three procedures: (a). PacBio long reads assembly using the falcon pipeline, assembly was further polished by Quiver and Pilon-1.18 (contig N50: 4.7 Mb); (b). SSPACE-LongRead (version 1-1) was implemented for getting a longer scaffold (contig N50, 4.7 Mb; scaffold N50, 7.8 Mb); (c). PBJelly was used to close gaps (contig N50, 5.7 Mb; scaffold N50, 8.2 Mb). As you see, the increase of scaffold N50 in this stage mainly came from SSPACE-LongRead procedure (7.8 Mb VS 4.7 Mb). Then the assembled PacBio scaffolds were used as input for scaffolding by hybridScaffold software at the BioNano stage, which generated a hybrid assembly with scaffold N50 of 9.22 Mb. It seems that BioNano optical map did not increase N50 too much (9.22 Mb VS 8.2 Mb), we predicted that the main reason was the employment of SSPACE-LongRead procedure during the first stage assembly. This program dealt with the scaffold construction effectively and the efficiency may be overlap with the performance at the BioNano stage in our study. Therefore, it was reasonable the increase in scaffold N50 was not largely from the BioNano optical map stage. Following this, the 10X genomic linked reads were employed to construct larger scaffolds, fragScaff software was employed to finish the super-scaffold construction. This procedure has increased the genome assembly with a scaffold N50 of 24.09 Mb, suggesting the efficiency of 10X genomic linked reads in our work (24 Mb VS 9.2 Mb). The efficiency of 10X genomic linked reads was also seen in other publication (Mostovoy et al., 2016, Nature Methods), which shows that 10X linked read data contributes more to the increase in N50 length than the BioNano optical map. Despite this, we still did not know whether the largely increase from 10X data was typical or not, as only few publications were available using the combination of 10X reads and BioNano map. We added several sentences to expand our method section, particularly in the “De novo assembly of the R. roxellana genome” section.

Standard repeat masker, gene prediction, and other analysis is performed. The manuscript would be strengthened by also a consideration of duplicated sequences, which could be identified based on Illumina sequence data read depth. This may be beyond scope of this report, but could be considered.

Response to comment 6

Thanks for this comment. We added duplicated sequences/copynumbervariant (CNVs) analysis based on read depth estimated from illumine short reads to the assembled genome using BWA. Results showed that a total of 676 duplicated blocks were identified, whose total length was 9,198,900 bp. We added one paragraph to clarify this point.

“We also performed a CNV analysis. In brief, we first mapped the Illumina short reads to the assembled genome using BWA with default parameters. Then, the sorted mapping bam file was used as input for CNVnator v0.3.3 [38], with the parameters “-unique -his 100 -stat 100 -call 100.”. The obtained CNVs were filtered, retaining only those where q0 was <0.5 and e-val1 was <0.05. After filtering, 676 CNVs remained, with a total length of 9,198,900 bp (Supplementary Table S12).”.

Has the assembly itself been submitted to proper databases and repositories (such as Genbank)? I could not find this listed, only the raw data.

Response to comment 7

Thanks for this comment. The genome assembly and other supporting data have been submitted to GigaDB database and NCBI successfully. However, we did not release them now as interest competition exist and several research groups are also working

on this species. We appreciate the editor and reviewers understand the challenges in this case, and we will make related data available once this article is published.

In table 2 and others, what does the 'number' column mean? For example, are there 151 contigs  $\geq$  to the N50 length of 5.7mb? The meaning of the columns in the tables should be clearly explained.

Response to comment 8

Thanks for this comment. Yes, this example explains the exact mean of 'number' column. Following your comment, we revised Table 2 to be more clear. We added one sentence to explain the meaning of the 'number' column. In addition, we checked and revised other tables if not clearly explained (for example, Table 3 and Supplementary Table S1).

“The “Number” column represents the number of contigs/scaffolds longer than the value of the corresponding category.”.

The legend for figure 2 is not adequate. What does the color scale signify? What is the reader supposed to conclude from the figure?

Response to comment 9

Thanks for this comment. This plot shows the interactions between two 100-kb genomic regions (as determined by Hi-C), darker red means stronger interaction strength. We added two sentences in the figure legend to address this comment. “Hi-C interactions within and among chromosomes of *R. roxellana* chromosomes (Chr1–Chr22); interactions were drawn based on the chromatin interaction frequencies between pairs of 100-kb genomic regions (as determined by Hi-C). In principle, darker red cells indicate stronger and more frequent interactions, which in turn imply that the two sequences are spatially close.”

This figure tries to express the information of Hi-C interactions among 22 chromosomes with a 100 kb resolution. Stronger interactions are indicated in darker red and weak interactions are indicated in light yellow. The fig. 2 presented here was used to indicate the reliability of our assembly during the Hi-C stage.

Reviewer #3:

Wang, Wu et al. have produced a high-quality reference genome assembly for the emblematic golden snub-nosed monkey. The authors used a combination of long PacBio reads, 10-X linked reads, Hi-C contact maps, BioNano Optical maps, and Illumina paired end sequences, all of which were sequenced to a very high coverage. The resulting assembly has very high continuity and given the combination of different sequencing strategies essentially gives as good of an assembly as current methods can produce. The authors have used a state-of-the-art approach to produce their assembly, and the applied methodology is appropriate. The authors have also produced a gene annotation based on homology to other species, as well as expression data. The assembly provides a valuable genomic resource to study snub-nosed monkeys specifically, and Asian colobines in general.

General comments:

*R. roxellana* already has a genome assembly available, as the authors note in the manuscript. However, there is no comparison at all beyond a contig and scaffold N50. It would strengthen the manuscript if the authors could provide some comparisons to the previous assembly, e.g: A comparative, or what specific regions of the assembly were absent in the previous version, what do they contain, how many gaps were filled, how many of the gene family expansions/contractions are only detectable with the high quality assembly etc.

Response to comment 1

Thanks for this comment. We followed this comment and made some comparisons with previous assembly, including repeat analysis and synteny analysis. In comparison, our new assembly had a higher proportion of repeat sequences (50.82%) as compared to the previous version (46.15%); in particular, the number of LINE (long interspersed elements) transposable elements and tandem repeats was greatly increased (further details are given in the “Identification of repeat elements” section). Thus, the newly assembled genome was substantially more complete and continuous. Also, we aligned



our genome against the previous version using MUMMER (v4.0.0beta2) and identified a total of 2,217 insertions in our new assembly. These insertion regions were mainly located in the intergenic and repetitive regions. Further analysis showed that 6,452 gaps in the previous version that were predicted to be filled by >29.7 Mb of sequence in our new assembly. These filled gaps were mainly located in the intergenic and repetitive regions, with a small fraction of the sequence data annotated as gene regions.

We added several sentences to clarify this point.

“We evaluated our newly assembled *R. roxellana* genome against the previously published assembly. The contiguity of our *R. roxellana* genome was 100fold greater (contig N50: 5.72 Mb; scaffold N50: 144.56) than the previous version (contig N50: 25.5 kb; scaffold N50: 1.55 Mb) [11]. We also aligned our genome against the previous version using MUMMER (v4.0.0beta2) [37] and identified 6,452 gaps in the previous version that were predicted to be filled by >29.7 Mb of sequence in our new assembly. These filled gaps were mainly located in the intergenic and repetitive regions, with a small fraction of the sequence data annotated as gene regions. Our new assembly also had a higher proportion of repeat sequences (50.82%) as compared to the previous version (46.15%); in particular, the number of LINE (long interspersed elements) transposable elements and tandem repeats was greatly increased (further details are given below, in the “Identification of repeat elements” section). Thus, the newly assembled genome was substantially more complete and continuous. It was likely that the remarkable improvement in contiguity was due to the increased read length, deeper sequencing depth, improved gap assembly, and more sophisticated assembly algorithm.”

The authors use several different software packages for their analysis. The inclusions of version numbers for the software packages they used seems somewhat arbitrary. Furthermore, no parameter sets apart from "default parameters" are ever presented. Both package versions and parameter settings should absolutely be included, otherwise the methods of the study are not properly understandable. In its current state, I feel the methodological aspects of the manuscript need to be expanded.

Response to comment 2

Following this comment, we added the methodological details substantially to address this comment. Both package versions and parameter settings were included in this version. please see “De novo assembly of the *R. roxellana* genome” section and other sentences containing software names in our manuscript for details.

The manuscript will benefit from language editing, as at several points the phrasing is somewhat confusing.

Response to comment 3

Thanks for this comment, the manuscript has been revised and polished by an English-language editing service of LetPub.

Specific comments:

L19, L68, L80: The claim of "incompleteness" or "greatly improved" is not backed by a proper comparison to the previous assembly.

Response to comment 4

We followed this comment and made comparisons with previous assembly, including repeat analysis and synteny analysis. In comparison, our new assembly had a higher proportion of repeat sequences (50.82%) as compared to the previous version (46.15%); in particular, the number of LINE (long interspersed elements) transposable elements and tandem repeats was greatly increased. Also, We aligned our genome against the previous version using MUMMER (v4.0.0beta2) [37] and identified 6,452 gaps in the previous version that were predicted to be filled by >29.7 Mb of sequence in our new assembly. These filled gaps were mainly located in the intergenic and repetitive regions, with a small fraction of the sequence data annotated as gene regions. Most importantly, the newly assembled *R. roxellana* reference genome has 100fold higher contiguity than previous assembly (contig N50: 5.72 Mb versus 25.5 kb and scaffold N50: 144.56 Mb versus 1.55 Mb).

We added several sentences to address this comment in the “Assessment of the genome newly assembled” section. See also the response to your valuable comment 1.

L22: Genetic-specific signatures is awkwardly phrased.

Response to comment 5

Thanks for this valuable comment. In fact, we were intended to term those genetic changes associated with the development of species-specific traits as “genetic-specific signatures of this species”. We realized that this sentence was confusing and not straightforward. We changed the statement as follows:  
“genetic changes associated with the development of specific traits in this species”.

L25: Technology, not technique

Response to comment 6

Thank you for your kindly review. We did it.

L57: This sentence is vague, please be specific about what these studies have looked at. The term research-hotspot for this species might be a stretch.

Response to comment 7

Thanks for this comment. Specifically, Recent studies of *R. roxellana* have focused on behavioral dynamics, population history, and social systems. We removed the term research-hotspot in this sentence.

“Recent studies of *R. roxellana* have focused on behavioral dynamics, population history, and social systems [5-7],”

L58f: This sentence needs rephrasing. What are the groups?

Response to comment 8

Following this comment, we reworded this sentence and also specify species the groups included.

“Genomic analyses have helped to untangle the molecular evolution of several groups, including maize (*Zea mays*), bats (*Myotis brandtii*), and killifish (*Nothobranchius furzeri*) [8-10]”.

L60: differentiate -> be distinguished

Response to comment 9

Thank you for your kindly review. We did it.

L63: Was there more than one assembly before this study?

Response to comment 10

Thanks for this comment. Actually, there is only one assembly published in 2014 before our study. We reworded this sentence as follows to avoid confusing.

“To date, only a single genome assembly is available for *R. roxellana*. This assembly, published in 2014, was derived from short sequencing reads generated by the Illumina HiSeq 2000 platform.”

L71f: This sentence needs rephrasing; it is not clear to me what the authors want to say.

Response to comment 11

We followed this comment and reworded this sentence to make it clear enough.

“Indeed, many previously unreported transposable elements and specific genes in maize were identified using an improved reference genome [16].”.

L74,L78: Please be specific with respect to the sequencing technology. "High quality" is subjective and changes with sequencing technologies, so arguing that no "high quality assembly of *R. roxellana* has been reported" is debatable.

Response to comment 12

Thanks for this comment. These new sequencing technologies used here referred to PacBio SMRT sequencing, BioNano optical mapping, and Hi-C based chromatin interaction maps. Additionally, we agree that "High-quality" is subjective and changes with sequencing technologies. We reworded this sentence to clarify this point.

“However, the *R. roxellana* genome has not yet been updated using new sequencing approaches, slowing progress towards a better understanding of this endangered species.”.

L75: Ref 15. Also includes an assembly for the Chimpanzee, which is closer to Human than either Gorilla or the Orangutan. 'Widely' should be omitted in this sentence.

Response to comment 13

Following this comment, we added the assembly of the chimpanzee in our manuscript. In addition, we removed 'Widely' in this sentence.

“New sequencing technologies including Pacific Bioscience’s single-molecule real-time (SMRT) sequencing, BioNano optical mapping, and Hi-C-based chromatin interaction maps, have been used in several species closely related to humans, including gorillas (*Gorilla gorilla gorilla*) [17], chimpanzees (*Pan troglodytes*) [18], and Sumatran orangutans (*Pongo abelii*) [18], as well as in other species, including the domestic goat (*Capra hircus*) [19].”

L76: “A lot of new findings” is vague, please specify the specific advantages of the new assemblies.

Response to comment 14

Following this comment, we added several sentences to clarify the specific advantages of the new assemblies.

“Importantly, it was estimated that 87% of the missing reference exons and incomplete gene models were recovered using the new gorilla assembly [17]. In addition, several novel genes expressed in the brain were identified using the new orangutan assembly, and complete immune genes with longer repetitive structures were identified in the updated goat genome [19].”

L81: Through combined -> by combining

Response to comment 15

Thank you for your kindly review. We did it.

L110: Cutadapter -> Cutadapt

Response to comment 16

Thank you for your kindly review. We did it.

L115ff: The value for Kerror was omitted.

Response to comment 17

Thanks for this comment, we added the value for Kerror.

“Finally, a total number of 109,210,004,556 k-mers, 1,159,024,556 k-mers with sequencing errors were generated and the peak k-mer depth was 34.”

L125: Quier -> Quiver

Response to comment 18

Thank you for your kindly review. We did it.

L130: To the best of my knowledge, PBJelly doesn't know how to deal with phased assemblies. All previous assembly steps (Falcon, Quiver, Pilon, sspace) also do not talk about phasing information. Please clarify how phasing was dealt with or maintained at this point.

Response to comment 19

Thanks for your valuable comment. We agree that PBJelly and previous assembly steps (Falcon, Quiver, Pilon) could not deal with phased assemblies. The term “phased genome assembly” here was used to indicate the genome assembly finished at this period, but not the “phased haplotype-resolved genome assembly”. This sentence was confusing here, we now say: “Thus, at the end of the first stage, the genome assembly had a contig N50 of 5.72 Mb and a scaffold N50 of 8.20 Mb (Supplementary Table S3).”

L130: The authors only mention the scaffold N50 after gap-filling. I see the contig N50 is mentioned in the supplementary, but I cannot find the contig N50 of the base assembly before gap-filling anywhere. It would be worth to mention it to understand the relative contributions of additional steps.

Response to comment 20

Thanks for your comment. Following gap-filling with PBJelly software, contig N50 increased to 8.2 Mb from N50 of 7.8 Mb at previous step. We added details to clarify this point.

“Using the initial genome assembly, SSPACE-LongRead v1-1 [33] was implemented for getting a longer scaffold by processing PacBio long reads and the initial genome assembly with the command “perl SSPACE-LongRead.pl -c <contig-sequences> -p <pacbio-reads>.” This procedure generated a genome assembly with scaffold N50 of 7.81 Mb (Supplementary Table S2). The remaining gaps in the assembly were closed using the PBJelly module in the PBSuite (version 15.8.24) [34] with default settings.

Thus, at the end of the first stage, the genome assembly had a contig N50 of 5.72 Mb

and a scaffold N50 of 8.20 Mb (Supplementary Table S3).”

L136: due -> using

Response to comment 21

Thank you for your kindly review. We did it.

L144: Can the authors comment on the difference between the genome size based on k-mer estimates and the actual assembly size?

Response to comment 22

Thanks for your comment. This difference may be due to the large number of repeat sequences in the genome, which occupied more than 50% of the genome region. Despite the Pacbio reads were used, a lot of repeat sequences were still could not be assembled, for example in the centromeres regions. In addition, we checked the duplicated genes and found only 1.6% duplicated genes compared to 92.4% of complete BUSCO matches. This suggests major duplication did not account for this assembly.

L145: acquired -> assembled

Response to comment 23

Thank you for your kindly review. We did it.

L147ff: It would be great to actually show this, e.g. by checking what the filled gaps contain. What added value does the new assembly have.

Response to comment 24

Thanks for this comment. We made some comparisons between our new assembly and the previous assembly. we aligned our genome against the previous version using MUMMER (v4.0.0beta2) and identified a total of 2,217 insertions in our new assembly. These insertion regions were mainly located in the intergenic and repetitive regions. Further analysis showed that 6,452 gaps in the previous version that were predicted to be filled by >29.7 Mb of sequence in our new assembly. These filled gaps were mainly located in the intergenic and repetitive regions, with a small fraction of the sequence data annotated as gene regions. Also, our new assembly had a higher proportion of repeat sequences (50.82%) as compared to the previous version (46.15%); in particular, the number of LINE (long interspersed elements) transposable elements and tandem repeats was greatly increased (further details are given in the “Identification of repeat elements” section). Thus, the newly assembled genome was substantially more complete and continuous.

We added several sentences to address this comment. See also the response to your valuable comments 1 and 4.

L150: I feel that mapping ratios of Illumina data are not an adequate measure for assembly accuracy, especially given that BWA mem maps all reads very liberally. I understand the desire to include such a number, a better (albeit not perfect) solution might be to map the Illumina data, perform a standard variant calling and quantify the number of high confidence homozygous alternative variants as a proxy to the assemblies' error rate.

Response to comment 25

Thanks for this comment. We performed a standard variant calling by Samtools, results showed that the number of homozygous SNP was 7690, occupying a proportion of 0.0004% in all SNPs, suggesting a high assembly accuracy rate. We added two sentences and one table (supplementary table S7) to address this comment.

“Genome assembly accuracy was also measured using the standard variant calling method in samtools (<http://samtools.sourceforge.net/>), with the command “samtools mpileup -q 20 -Q 20 -C 50 -uDEF.” We found that the homozygous SNP (single nucleotide polymorphism) s comprised 0.0004% of all SNPs (7,690 of 559,048), suggesting that our genome assembly was highly accurate (Supplementary Table S7).”

L163: identified -> identify

Response to comment 26

Thank you for your kindly review. We did it.

L163: homolog -> homology

Response to comment 27

Thank you for your kindly review. We did it.

L165: I suppose the authors used all of RepBase, not only the TEs within it?

Response to comment 28

Thanks for this comment. Yes, we used all elements in the RepBase database, but not only the TEs within it. We corrected this sentence as follows.

“In the homology approach, we searched the genome for repetitive DNA elements (as listed in the Repbase database v16.02) using RepeatMasker v4.0.6 (<http://www.repeatmasker.org/>) [29] with the parameters “-a -nolow -no\_is -norna -parallel 1” and using RepeatProteinMask (implemented in RepeatMasker).”.

L168: The authors ran RepeatModeler in addition to RepeatMasker. It would be interesting to know if they detected repeat elements that are absent from RepBase and might be unknown/lineage specific.

Response to comment 29

We followed this comment and examine the repeat elements detected from RepeatModeler and RepeatMasker respectively. Results showed that several repeat elements including LINE and SINE absent from Repbase database were detected in the de novo approach (Supplementary Table S10). The total length of these repeat elements was 186,195,432bp, accounting for 6.13% of the genome, suggesting that these repeat elements may be specific for *R. roxellana*.

L178: Specify what database was used.

Response to comment 30

We followed this comment and added two sentences to clarify this point.

“Using BLASTN with an E-value of 1E-10, we identified four rRNAs in the *R. roxellana* genome homologous to human rRNAs: 28S, 18S, 5.8S, and 5S (GenBank accession numbers NR\_003287.2, NR\_003286.2, NR\_003285.2, and NR\_023363.1, respectively).”

L208ff: This sentence is very vague. Please be specific about what this comparison is about, and what "other mammals" were included and why.

Response to comment 31

Thanks for your valuable comment. Here, we want to compare the gene structure information including mRNA length, exon length, intron length and exon number between *R. roxellana qinlingensis* and other representative mammals. In this sentence, "other mammals" including *Homo sapiens*, *Gorilla gorilla*, *Macaca mulatta*, *Rhinopithecus bieti*, *Rhinopithecus roxellana hubeiensis*. We chose these mammals as human and gorilla are the most representative primates with high-quality genome, while *Macaca mulatta* could represent Cercopithecinae, the sister group of Colobinae consisting the sequencing species *Rhinopithecus roxellana qinlingensis*. As for *R. bieti* and *R. r. hubeiensis*, they were the congeneric species of *R. r. qinlingensis*, more importantly, the *R. r. hubeiensis* and *R. r. qinlingensis* are both the subspecies of *Rhinopithecus roxellana*.

We added several sentences to clarify this point.

“We also compared the gene structure, including mRNA length, exon length, intron length, and exon number, among *R. roxellana qinlingensis* and other representative primates (e.g., *H. sapiens*, *G. gorilla*, *M. mulatta*, *R. bieti*, and *R. r. hubeiensis*). We found that genome assembly patterns were similar among *R. roxellana qinlingensis* and the other primates (Supplementary Fig. S2).”.

L211: The authors need to specify what they mean by functional annotation, and how this annotation was performed. Assigning a biological function to 22053 seems a bit high.

Response to comment 32

Thanks for this comment. Functional annotation indicated those predicted genes were annotated with the known protein databases to better understand their biological function. We performed the annotation analysis by annotating the predicted genes to the known protein database (NR, SwissProt and KEGG et al.) with the blastp command, and the best match for each gene was identified with the blast E value of 1E-5. Nearly half (10,670 of 22,497) of these genes were annotated to the predicted proteins in NR database derived from the previous genome annotation for the *Rhinopithecus roxellana*. And it therefore was reasonable for the assignment of 22,053 genes with biological function. We added several sentences to clarify this point.

	<p>“To better understand the biological functions of the predicted genes, we used BLASTP (with an E-value of 1E-5) to identify the best match for each predicted gene across several databases, including the NCBI nonredundant protein database (NR v20180129), SwissProt (v20150821) [54], Kyoto Encyclopedia of Genes and Genomes (KEGG v20160503) [55], InterPro v29.0 [56], Pfam v31.0 [57], and GO (Gene Ontology)[58]. In this way, 22,053 predicted genes (98.42%) were functionally annotated (Supplementary Table S14). Nearly half (10,670 of 22,497) of these genes were annotated to the predicted proteins in NR database derived from the previous genome annotation for <i>Rhinopithecus roxellana</i>.”</p> <p>L235f: The authors present what looks like a GO-term enrichment analysis, but I can't find any mention as to how this analysis was performed. Response to comment 33 Thanks for this comment. It is true that we performed a GO-term enrichment analysis. This analysis was performed towards the significantly expanded gene families in <i>Rhinopithecus roxellana</i>. We added several sentences to address this comment. “ To explore the significantly expanded gene families, we performed a GO-term enrichment analysis with EnrichPipeline32 [66, 67], using the 1,370 genes belonging to the 314 significantly expanded gene families as input, and using all predicted genes as background. We considered GO term significant if adjusted the P-value was &lt;0.05. We found that the significantly expanded gene families were mainly associated with the hemoglobin complex, energy metabolism, and oxygen transport (Supplementary Table S16).”</p> <p>L250: I can't find this repository on SRA. Response to comment 34 Thanks for this comment. The genome assembly and other supporting data have been submitted to GigaDB database and NCBI successfully. However, we did not release them now as interest competition exist and several research groups are also working on this species. We appreciate the editor and reviewers understand the challenges in this situation and we will make related data available soon once this article is published.</p>
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<p><b>Experimental design and statistics</b></p> <p>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> <p>Have you included all the information requested in your manuscript?</p>	Yes
<p><b>Resources</b></p> <p>A description of all resources used, including antibodies, cell lines, animals</p>	Yes

<p>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><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>





12 **ABSTRACT**

13

14 **Background:** ~~The golden snub-nosed monkey (*Rhinopithecus roxellana*) is an endangered~~  
15 ~~colobine monkey species endemic to China. This species has several distinctive traits, and it is~~  
16 ~~an ideal model for analysing analyses of the evolutionary development of the social~~  
17 ~~structurestructures due to its unique social organization. Although there has been reported a~~  
18 ~~genome assembly for the subspecies *R. roxellana hubeiensis*, the is available, this~~ assembly  
19 ~~is incomplete and fragmented due to employingbecause it was constructed using~~ short  
20 ~~readsread~~ sequencing technology. ~~This drawback may lose~~ Thus, information important for the  
21 understanding of *R. roxellana*, such as genome structural variation and repeat sequences ~~which~~  
22 ~~are important for understanding this endangered species.~~, may be absent from the available  
23 assembly. Therefore, ~~to have a better understanding of evolutionary history and genetic~~  
24 ~~specific signatures,~~ a high-quality reference genome ~~of the taxon~~ is ~~need.~~ needed.

25 **Findings:** To obtain a high-quality ~~chromosome~~ chromosomal assembly ~~for~~ *R. roxellana*  
26 *qinlingensis*, we ~~combined a total of used~~ five ~~techniques including~~ different methods: Pacific  
27 Bioscience<sup>2</sup>s single-molecule real-time sequencing, Illumina<sup>2</sup>s paired-end sequencing,  
28 BioNano optical maps, 10X Genomics link-reads, and high-throughput chromosome  
29 conformation capture. The ~~results indicate the~~ assembled genome ~~is about was~~ ~3.04 Gb, with  
30 a contig N50 of 5.72 Mbp and a scaffold N50 of 144.56 Mbp, ~~which have made~~ Mb. This

31 ~~represented~~ a ~~40100~~-fold improvement ~~compared to past~~ over the previously published. ~~It is~~  
32 ~~shown that a total of 22497~~ genome. In the new genome, 22,497 protein-coding genes were  
33 predicted, of which ~~22053~~22,053 were functionally annotated. ~~Moreover, gene~~Gene family  
34 analysis ~~show~~showed that 993 and ~~27452~~7,745 gene families ~~are~~were expanded and contracted  
35 ~~in the *R. roxellana* genome~~, respectively, ~~in the *R. r. qinlingensis* genome~~. The reconstructed  
36 ~~phylogeny recovered a close relationship between *Rhinopithecus rollexana* and *Macaca*~~  
37 ~~*mulatta*, and these two species diverged approximately 13.4 MYA.~~

38 **Conclusion:** We ~~present the updated~~ constructed a high-quality genome assembly of ~~*R.*~~  
39 ~~*roxellana* with~~Qinling golden snub-nosed monkey; this genome had superior continuity and  
40 accuracy. ~~The assembled genome can, which might~~ be ~~used~~useful as reference for future  
41 genetic studies ~~of the~~in this species. In addition, the updated genome assembly might improve  
42 ~~our understanding of this~~ species. ~~Also, the updated genome assembly may contribute to our~~  
43 ~~comprehensive understanding of the species, which is and might be~~ particularly helpful in the  
44 ~~relevant to~~ conservation ~~of this endangered species efforts~~. Furthermore, ~~such genome with~~  
45 ~~superior continuity and accuracy can provide~~this high-quality genome might serve as a new  
46 standard reference genome for ~~Colobine~~colobine primates.

47

48 **Keywords:** high-quality; *Rhinopithecus roxellana*; genome assembly; annotation; BioNano  
49 optical maps

## Data Description

### Background information

~~Snub-nosed monkeys (*Rhinopithecus*) consist of five endangered species narrowly restricted to China and Vietnam [1]. Among those, the golden snub-nosed monkey (*Rhinopithecus roxellana*) is also referred to as the Sichuan snub-nosed monkey, with the northernmost distribution of all Asian colobinae species, found only in three isolated regions (Sichuan and Gansu, Shaanxi and Hubei provinces) in central and northwest China. The snub-nosed monkeys (genus *Rhinopithecus*) consist of five endangered species narrowly restricted to China and Vietnam [1]. Of those, the golden snub-nosed monkey (*Rhinopithecus roxellana*), also known as the Sichuan snub-nosed monkey, has the northernmost distribution of all Asian colobine species; this monkey is found only in three isolated regions in central and northwest China (the Sichuan, Gansu, Shaanxi, and Hubei Provinces) [2, 3]. This species. The golden snub-nosed monkey is characterized by several distinctive traits, such as including golden fur, a blue facial colour, color, an odd-shaped nose, more folivorous, most striking and folivory. In addition, the species has a unique multilevel social system with multilevel societies, a rare and; such complex system that is systems are found only in a few mammal species, including human beings [4]. mammals, including humans [4]. Therefore, *R. roxellana* Qinling golden snub-nosed monkey is an ideal model for analyzing the analysis of social structure evolution in primates~~

69 and may also provide chances/opportunities to investigate evolutionary and socio-  
70 anthropological patterns of human society ~~in social anthropology~~.

71 ~~As a research hotspot, studies on *R. roxellana* have investigated various aspects [5-7].~~  
72 ~~Recently, genomic analysis offered a powerful tool and has successfully been employed to~~  
73 ~~underlie the molecular evolution of several groups [8-10]. According to the morphological~~  
74 ~~variation and distribution difference, *R. roxellana* can differentiate into three subspecies:~~  
75 ~~*Rhinopithecus roxellana roxellana* from Minshan mountains of Sichuan and Gansu province,~~  
76 ~~*R. r. Qinlingensis* from Qinling mountains of Shaanxi province, *R. r. hubeiensis* from~~  
77 ~~Shennongjia Mountains [3]. Up to now, the best genome assembly of *R. roxellana* was~~  
78 ~~published in 2014. Based on morphological variations and discontinuous distributions, *R.*~~  
79 ~~*roxellana* is distinguished into three subspecies: *R. r. roxellana* from the Minshan Mountain in~~  
80 ~~the Sichuan and Gansu Provinces, *R. r. qinlingensis* from the Qinling Mountain in Shaanxi~~  
81 ~~Province, and *R. r. hubeiensis* from Shennongjia Mountain in Hubei Province [3]. Recent~~  
82 ~~studies of *R. r.* have focused on behavioral dynamics, population history, and social systems~~  
83 ~~[5-7], which was derived from short reads sequencing on Illumina HiSeq 2000 platform. Based~~  
84 ~~on this achievement, studies on its folivorous dietary adaptations and the evolutionary history~~  
85 ~~of *R. roxellana* have been conducted. Genomic analyses have helped to untangle the molecular~~  
86 ~~evolution of several groups, including maize (*Zea mays*), bats (*Myotis brandtii*), and killifish~~  
87 ~~(*Nothobranchius furzeri*) [8-10]. Despite such progress, the information including structural~~

88 ~~variation and repeat sequences was largely absent or unreliable due to the incomplete and~~  
89 ~~fragmented genome assembly. To date, only a single genome assembly is available for golden~~  
90 ~~snub-nosed monkey. This assembly, published in 2014, was derived from short sequencing~~  
91 ~~reads generated by the Illumina HiSeq 2000 platform [11].~~

92 ~~Owing to the advances in sequencing technology, it is possible to obtain high quality~~  
93 ~~genome assembly that can provide new insights into the understanding of the organisms.~~  
94 ~~Indeed, many unreported transposable elements and specific genes were identified by using the~~  
95 ~~improved maize reference genome. Several studies have been published based on these data,~~  
96 ~~including analyses of the folivorous dietary adaptations of *R. r.* and its evolutionary history~~  
97 ~~[11-13]. Despite the utility of this previously published data, much relevant information,~~  
98 ~~including structural variations and repeat sequences, is largely absent or unreliable due to the~~  
99 ~~incomplete and fragmented genome assembly. ~~By combining new sequencing approaches, Seo~~~~  
100 ~~et al. [14, 15].~~

101 ~~Owing to advances in sequencing technology, it is now possible to obtain high-quality~~  
102 ~~genome assemblies that can provide new insights in organismal research. Indeed, many~~  
103 ~~previously unreported transposable elements and specific genes in maize were identified using~~  
104 ~~an improved reference genome [16]. ~~By combining new sequencing approaches, Seo et al. [14]~~~~  
105 ~~discovered clinically relevant structural variants and previously unreported genes in the~~  
106 ~~updated human genome. New sequencing technologies, including Pacific Bioscience's single-~~

107 molecule real-time (SMRT) sequencing, BioNano optical mapping, and Hi-C-based chromatin  
108 interaction maps, have been used in several species closely related to humans, including  
109 gorillas (*Gorilla gorilla gorilla*) [17], chimpanzees (*Pan troglodytes*) [18], and Sumatran  
110 orangutans (*Pongo abelii*) [18], as well as in other species, including the domestic goat (*Capra*  
111 *hircus*) [19]. Importantly, it was estimated that 87% of the missing reference exons and  
112 incomplete gene models were recovered using the new gorilla assembly [17]. In addition,  
113 several novel genes expressed in the brain were identified using the new orangutan assembly,  
114 and complete immune genes with longer repetitive structures were identified in the updated  
115 goat genome [19]. However, the *R. r.* genome has not yet been updated using new sequencing  
116 approaches, slowing progress towards a better understanding of this endangered species.

117 Here, we report a greatly improved assembly ~~and annotation~~ of the reference genome for  
118 *R. roxellana r.* ~~from through combined~~ generated by a combination of five technologies: ~~Pacific~~  
119 ~~Bioscience's single molecule real time~~ SMRT sequencing (SMRT), ~~Illumina's~~ from Pacific  
120 Biosciences (PacBio), HiSeq paired-end sequencing from Illumina (HiSeq), BioNano optical  
121 maps (BioNano), 10X Genomics link-reads (10X Genomics), and high-throughput  
122 chromosome conformation capture (Hi-C). ~~Also, this is~~ Our results represent the first colobine  
123 genome sequenced and assembled with both long reads and short reads. ~~The~~ This updated  
124 genome assembly may allow us to further investigate *R. roxellana r.*, ~~offering~~ providing new  
125 opportunities ~~in analyzing to analyze~~ evolutionary history and ~~searching those to identify~~

126 genetic changes associated with the development of specific traits in this species, ~~which. Such~~  
127 ~~analyses~~ may provide ~~new~~ insights ~~in~~ helpful for the conservation of this endangered primate.  
128 In addition, this genome ~~with, which has~~ superior continuity and accuracy, will ~~provide~~ act as  
129 a new ~~standard~~ reference genome for colobine primates.

130

## 131 Data Description

### 132 **Sample collection and sequencing**

133 The animal used for ~~the~~ sequencing was an adult ~~dead~~ male *R. ~~roxellana~~ qinlingensis* ~~in Qinlin~~  
134 ~~Mountains, from Qinling Mountain, who died naturally and~~ the dead body was stored in ultra  
135 ~~cold~~ storage freezer at Louguantai Breeding Centre, Xi'an, Shaanxi ~~province~~ Province, China.  
136 Total genomic DNA was extracted from ~~the~~ heart tissue. To acquire a high-quality genome  
137 assembly, we ~~applied a~~ combined five sequencing methods. Initially, PacBio's SMRT  
138 sequencing was ~~conducted~~ performed on the SEQUEL platform ~~according to manufactures,~~  
139 ~~after removing adaptors in polymerase reads, resulting a total of~~ following the manufacturer's  
140 instructions. After quality control, during which subreads shorter than 500 bp were removed,  
141 304.84 Gb clean long reads (95.86X coverage) ~~Different from~~ remained. The average/N50  
142 length of the PacBio sequencing reads was 16.69 kb. Simultaneously, paired-end sequencing  
143 was performed using an Illumina NovaSeq 6000 platform, with an insert size of 350 bp. Then  
144 those S short reads ~~derived from this step~~ were filtered by SOAPfilter v. 2.2 [20] ~~(using the~~

145 ~~SOAPdenovo2 software from SOAPdenovo2) with the following criteria: filtering those [20],~~  
146 ~~removing~~ reads with adapters, contaminations, ~~N>10% unknown~~ bases ~~more than 10% and (N),~~  
147 ~~or~~ low quality, ~~which generated~~. After filtering, 423.32 Gb ~~sequencing~~ clean reads ~~remained~~  
148 (133.12X coverage). ~~In addition, a~~ high-quality optical genome map was ~~also~~ constructed  
149 with ~~the~~ Irys platform (BioNano Genomics), ~~from which we acquired~~. The average/N50  
150 ~~length of the molecules used for optical mapping was 338 kb. The average BioNano optimal~~  
151 ~~marker density examined was 11.66 per 100 kb, while the average marker density was 12.62~~  
152 ~~per 100 kb for the predicted map based on the assembled contigs. Thus, the observed BioNano~~  
153 ~~map was consistent with the predicted map. The BioNano map generated~~ 463.75 Gb of large  
154 DNA molecules. ~~Besides~~Next, 10X genomic ~~linklinked~~-reads sequencing was ~~carried~~  
155 ~~outperformed~~ on ~~an~~ Illumina Hiseq Xten platform, ~~and~~generating 348.41 Gb clean reads  
156 (109.56X coverage) ~~were generated in total~~. Finally, ~~a~~an Hi-C library was prepared and  
157 sequenced with an Illumina NovaSeq 6000 platform ~~for to produce a~~ chromosome-scale  
158 scaffolding of ~~the~~ genome assembly. Adapter sequences and low-quality reads were discarded  
159 ~~by using CutadapterCutadapt v1.0 [21],-~~ [21] with the parameters “-e 0.1 -O 5 -m 100 -n 2 --  
160 ~~pair-filter=both,~~” yielding ~~a total of~~ 310.92 Gb clean data (97.77X coverage). ~~Statistics of the~~  
161 ~~Detailed~~ sequencing data ~~was detailed~~statistics are given in **Table 1**.

162

163



164 **De novo assembly of the *R. roxellana* genome**

165 ~~Estimation~~An estimation of genome size ~~is helpful to~~would increase our understanding of *R.*  
166 *roxellana*. ~~Generally~~Thus, we estimated the ~~size of the *R. roxellana* genome~~ size of *R.*  
167 ~~*roxellana* with the formula of~~  $G = (K_{total} - K_{error})/D$ , ~~in which~~where  $G$  ~~represents~~represented  
168 genome size, ~~while  $K_{total}$ ,  $K_{error}$  and  $D$  indicates~~ represented the total number of k-mers,  $K_{error}$   
169 represented the number of k-mers ~~which caused by~~with sequencing errors, and  $D$  indicated the  
170 k-mer depth ~~respectively~~. ~~Finally,~~ We generated 109,210,004,556 k-mers ~~were generated, and~~  
171 ~~the, 1,159,024,556 of which had sequencing errors.~~ The peak k-mer depth was 34. Thus, the  
172 genome size of *R. roxellana* was estimated to be about 3.18 Gb. The distribution of k-mer  
173 ~~frequency was shown~~frequencies is given in **Supplementary Fig. S1**.

174 ~~The *de novo* assembly of newly sequenced *R. roxellana* genome was performed in four~~  
175 ~~progressive steps. Firstly, the assembly was conducted with the FALCON assembler (default~~  
176 ~~parameters) [11] with the long reads obtained from the PacBio platform, which mainly includes~~  
177 ~~three steps: 1) detection of overlap and reads correction; 2) detection of overlap between~~  
178 ~~corrected reads; and 3) construction of string graph. Following FALCON step, the string graph~~  
179 ~~assembly was further polished by Quiver with long reads [22] and then corrected by Pilon with~~  
180 ~~Illumina short reads [23]. Based on this initial genome assembly, space longreads [20] with~~  
181 ~~default settings was implemented for getting a longer scaffold genome by using PacBio long~~  
182 ~~reads. Despite attempts have been made, scaffolding gaps were still found, those gaps were~~

183 further closed with the help of PBjelly software under default settings, which generated a Time-  
184 phased genome assembly with scaffold N50 of 8.20 Mbp (**Supplementary Table S1**).

185 Secondly, a hybrid assembly with scaffold N50 of 9.22 Mbp was constructed on the basis  
186 of Bionano optical map data using Bionano Solve3.1 ([www.bionanogenomics.com](http://www.bionanogenomics.com)) with  
187 default parameters (**Supplementary Table S2**). Thirdly, 10X genomic linked reads were  
188 employed to connect scaffolds from the second step by fragScaff software [24], which has  
189 updated the scaffold N50 of genome assembly to 24.09 Mbp (**Supplementary Table S3**).  
190 Subsequently, those short reads derived from Illumina were applied to correcting errors due  
191 to Burrows Wheeler Aligner (BWA) [25] and pilon 1.18 [23].

192 Finally, to build chromosome level assembly scaffolds, we mapped the Hi-C reads to the  
193 assembled scaffolds with BWA [25]. Then Hi-C data was subsequently applied to cluster, order,  
194 and orient scaffolds by Lachesis software [26]. The chromosome level scaffolds for *R.*  
195 *roxellana* allowed us to estimate the interaction frequency between chromosome loci, the  
196 interaction heatmap shown in **Fig. 2**.

197 — These processes together yielded a updated genome assembly of *R. roxellana* with its  
198 genome size of 3.04 Gb, contig N50 of 5.72 Mbp and scaffold N50 of 144.56 Mbp (**Table 2**).  
199 In comparison, the newly acquired *R. roxellana* reference genome has 100 fold higher  
200 contiguity than its previous (contig N50: 5.72 Mb versus 25.5 kb and scaffold N50: 144.56 Mb  
201 versus 1.55 Mb) [11]. We suppose that the remarkable improvement in contiguity can be

202 attributed to the longer read length, deeper sequencing depth, properly assembled gaps, and  
203 increased sophisticated assembly algorithm.

204 To assess the genome assembly accuracy, we aligned the Illumina short reads to the  
205 assembly by BWA program [25]. The mapping rate for the reads was about 99.17%, further  
206 investigations showed that those mapped reads covered approximately 99.27% of the assembly  
207 (Supplementary Table S4). In addition, we estimated the assembly completeness by  
208 conducting Benchmarking Universal Single-copy Orthologs (BUSCO) analysis with BUSCO  
209 V3.0 [27]. As for the BUSCO analysis, the annotation results were classified as complete  
210 BUSCOs, fragmented BUSCOs and missing BUSCOs. The results showed that among the  
211 4,104 mammalian BUSCOs, the complete BUSCOs, the fragmented BUSCOs and the missing  
212 BUSCOs occupied a proportions of 94.0%, 2.9% and 3.1% in the genome assembly of *R.*  
213 *roxellana qinlingensis*, respectively (Supplementary Table S5). The assembly completeness  
214 was also checked by core eukaryotic gene mapping approach (CEGMA) [28]. The results  
215 showed that 93.95% (233 of 248) conserved genes were found in our genome assembly  
216 (Supplementary Table S6). Together, these analyses indicated a high accuracy and  
217 completeness of our genome assembly.

218

219 **Identification of repeat elements**

220 ~~Repeat sequences occupy a large proportion of the genome sequences. Thus, it is~~  
221 ~~necessary for us to identified those repeat elements. In our study, we combined homolog based~~  
222 ~~and *de novo* based approach to predict and classify repeat elements. As for the homolog~~  
223 ~~approach, we searched transposable elements from the RepBase database [29] with~~  
224 ~~RepeatMasker v4.0.6 (<http://www.repeatmasker.org/>) and RepeatProteinMask (implemented~~  
225 ~~in RepeatMasker). The *de novo* method was employed with RepeatModeler V1.0.11 [30],~~  
226 ~~RepeatMasker v4.0.6 and Tandem Repeat Finder (TRF) (Version 4.07b) [31]. We merged the~~  
227 ~~findings from both methods. Results showed that 45.43% of the genome was predicted as~~  
228 ~~repeat elements (Supplementary Table S7). A closer investigation indicated that the largest~~  
229 ~~category of repeat elements in the species is the short (SINEs) and long (LINEs) interspersed~~  
230 ~~nuclear elements. The detailed categories of repeat elements are summarized in~~  
231 ~~Supplementary Table S8.~~

232 The *de novo* assembly of the newly sequenced *R. roxellana* genome was performed in  
233 four progressive stages. First, long reads obtained from the PacBio platform were assembled  
234 as follows: detection of overlap and read correction, detection of overlap between pairs of  
235 corrected reads, and string graph construction. Assembly of the PacBio long reads was  
236 performed using FALCON (version 0.4.0) [32] with the parameter set “length\_cutoff = 5000,  
237 length\_cutoff\_pr = 5000, pa\_HPCdaligner\_option = -v -B128 -e.70 -k14 -h128 -l2000 -w8 -T8  
238 -s700, ovlp\_HPCdaligner\_option = -v -B128 -e.96 -k16 -h480 -l1500 -w8 -T16 -s700” . Next,

239 the assembled PacBio contigs was polished using Quiver (SMRTLink version 5.1.0) with  
240 PacBio long reads [22], and also the contig assembly was corrected by Pilon-1.18 (java -  
241 Xmx500G -jar pilon-1.18.jar --diploid --threads 30) with Illumina short reads [23]. The contig  
242 N50 of the initial assembly was 4.74 Mb (Supplementary Table S1). Using the initial genome  
243 assembly, SSPACE-LongRead v1-1 [33] was implemented for getting a longer scaffold by  
244 processing PacBio long reads and the initial genome assembly with the command “perl  
245 SSPACE-LongRead.pl -c <contig-sequences> -p <pacbio-reads>.” This procedure generated a  
246 genome assembly with scaffold N50 of 7.81 Mb (Supplementary Table S2). The remaining  
247 gaps in the assembly were closed using the PBjelly module in the PBSuite (version 15.8.24)  
248 [34] with default settings. Thus, at the end of the first stage, the genome assembly had a contig  
249 N50 of 5.72 Mb and a scaffold N50 of 8.20 Mb (Supplementary Table S3).

250 In the second stage, the BioNano molecules were filtered, requiring a minimum length of  
251 150 kb and minimum of nine labels per molecule. Then, a genome map was assembled *de novo*  
252 with IrysView (version 2.3; BioNano Genomics), based on the optically mapped molecules.  
253 The assembled PacBio scaffolds were input into hybridScaffold [35]. In brief, the hybrid  
254 scaffolding process included the alignment of the PacBio scaffolds against the BioNano  
255 genome maps, followed by the identification and resolution of conflicting alignments. At the  
256 end of stage two, the hybrid genome assembly had a scaffold N50 of 9.22 Mb (Supplementary  
257 Table S4).

258 In the third stage, the 10X genomic linked reads were connected with the scaffolds  
259 generated in stage two to construct super-scaffolds. In brief, we used the long ranger basic  
260 pipeline (<https://support.10xgenomics.com/genome-exome/software/downloads/>) to handle  
261 the basic read in and barcode processing of the 10X genomic linked reads. The processed 10X  
262 linked reads were then mapped to the hybrid genome assembly from stage two with bowtie2  
263 [36], using the command “bowtie2 genome.fa -1 reads1.fq.gz -2 reads2.fq.gz -p 12 -D 1 -R 1 -  
264 N 0 -L 28 -i S,0.2,50 --n-ceil L,0,0.02 --rdg 5,10 --rfg 5,10).” We also used a self-against-self  
265 (genome.fa-against-genome.fa) blastn to generate two bed files, and merged these files using  
266 fragScaff (version 140324.1) [24], with the parameters “-fs1 '-m 3000 -q 20 -E 30000 -o 60000'  
267 -fs2 '-C 2', -fs3 '-j 1.5 -u 2'.” These procedures generated an updated genome assembly with a  
268 scaffold N50 of 24.09 Mb (Supplementary Table S5). Subsequently, we corrected errors in  
269 the assembly, based on the Illumina short reads, using the Burrows-Wheeler Aligner (BWA)  
270 [25] and Pilon-1.18 [23].

271 In the fourth stage, the Hi-C data were used to build chromosome-level assembly scaffolds.  
272 In brief, Hi-C sequencing data were first aligned to the assembled genome using BWA [25].  
273 Scaffolds were then clustered, ordered, and oriented using Lachesis [26], with the parameter  
274 set “CLUSTER MIN RE SITES = 1800, CLUSTER MAX LINK DENSITY = 4, and  
275 CLUSTER NONINFORMATIVE RATIO = 0.” This procedure generated 22 accurately  
276 clustered and ordered pseudo-chromosomes, with a genome size of 3.04 Gb, a contig N50 of

277 5.72 Mb, and a scaffold N50 of 144.56 Mb (Table 2). The pseudo-chromosomes were divided  
278 into 100-kb bins and the interaction frequencies between pairs of 100-kb genomic regions were  
279 determined (Fig. 2).

### 280 **Assessment of the genome newly assembled**

281 We evaluated our newly assembled *R. roxellana* genome against the previously published  
282 assembly. The contiguity of our *R. roxellana* genome was 100-fold greater (contig N50: 5.72  
283 Mb; scaffold N50: 144.56) than the previous version (contig N50: 25.5 kb; scaffold N50: 1.55  
284 Mb) [11]. We also aligned our genome against the previous version using MUMMER  
285 (v4.0.0beta2) [37] and identified 6,452 gaps in the previous version that were predicted to be  
286 filled by >29.7 Mb of sequence in our new assembly. These filled gaps were mainly located in  
287 the intergenic and repetitive regions, with a small fraction of the sequence data annotated as  
288 gene regions. Our new assembly also had a higher proportion of repeat sequences (50.82%) as  
289 compared to the previous version (46.15%); in particular, the number of LINE (long  
290 interspersed elements) transposable elements and tandem repeats was greatly increased (further  
291 details are given below, in the “Identification of repeat elements” section). Thus, the newly  
292 assembled genome was substantially more complete and continuous. It was likely that the  
293 remarkable improvement in contiguity was due to the increased read length, deeper sequencing  
294 depth, improved gap assembly, and more sophisticated assembly algorithm.

295 To assess the accuracy of our genome assembly, we aligned the Illumina short reads to  
296 the assembly using BWA [25], with the parameters “-o 1 -i 15”. Approximately 99.17% of the  
297 short reads were mapped to the genome assembly. Further investigations indicated that these  
298 reads covered approximately 99.27% of the total assembly (Supplementary Table S6).  
299 Genome assembly accuracy was also measured using the standard variant calling method in  
300 samtools (<http://samtools.sourceforge.net/>), with the command “samtools mpileup -q 20 -Q 20  
301 -C 50 -uDef.” We found that the homozygous SNP (single nucleotide polymorphism) s  
302 comprised 0.0004% of all SNPs (7,690 of 559,048), suggesting that our genome assembly was  
303 highly accurate (Supplementary Table S7). In addition, we estimated assembly completeness  
304 using Benchmarking Universal Single-copy Orthologs (BUSCO) v3.0.2 [27], with the  
305 parameters “-i -o -l -m genome -f -t.” based on mammalia odb9 (creation date: 2016-02-13;  
306 number of species: 50; number of BUSCOs: 4,104). BUSCO analysis identified 4,104  
307 mammalian BUSCOs in the newly assembled *R. roxellana* genome: 94.0% complete BUSCOs,  
308 2.9% fragmented BUSCOs, and 3.1% missing BUSCOs (Supplementary Table S8).  
309 Assembly completeness was measured using the core eukaryotic gene (CEG)-mapping  
310 approach (CEGMA v2.5) [28]. Of the 248 CEGs known from six model species, 93.95% (233  
311 of 248) were identified in our new genome assembly. Of these, 220 CEGs were complete and  
312 unfragmented, and the remaining 13 were complete but fragmented (Supplementary Table



313 S9). Together, these analyses indicated that our new genome assembly was highly accurate and  
314 complete.

315

### 316 **Identification of repeat elements**

317 Repeat sequences account for a large proportion of the total genome is thus important  
318 to identify repeat elements. Here, we predicted and classified repeat elements both based on  
319 homology and *de novo*. In the homology approach, we searched the genome for repetitive DNA  
320 elements (as listed in the Repbase database v16.02) using RepeatMasker v4.0.6  
321 (<http://www.repeatmasker.org/>) [29] with the parameters “-a -nolow -no is -norna -parallel 1”  
322 and using RepeatProteinMask (implemented in RepeatMasker). To identify repetitive elements  
323 *de novo*, we used RepeatModeler v1.0.11 [30], with the parameters “-database genome -engine  
324 ncbi -pa 15.” Tandem repeats in the genome were detected using Tandem Repeat Finder (TRF)  
325 v4.07b [31], with parameters “2 7 7 80 10 50 2000 -d -h”). We merged the results of the two  
326 methods. In total, the new genome assembly comprised 50.81% repetitive sequences  
327 (**Supplementary Table S10**). Closer investigation indicated that the largest categories of  
328 repeat elements in the *R. roxellana* genome were the short and long interspersed nuclear  
329 elements (SINEs and LINEs, respectively). In addition, several repeat elements absent from  
330 Repbase database were detected in the *de novo* approach (**Supplementary Table S10**). The  
331 total length of these repeat elements was 186,195,432bp, accounting for 6.13% of the genome.

332 suggesting that these repeat elements may be specific for *R. roxellana*. Compared with the  
333 repeat sequences in the previous assembly, our genome included relatively more LINE  
334 transposable elements (28.23% vs. 6.21%) and tandem repeats (6.20% vs. 2.82%). The detailed  
335 categories of repeat elements are summarized in **Supplementary Table S11**.

### 336 **Copy number variation (CNV)**

337 We also performed a CNV analysis. In brief, we first mapped the Illumina short reads to  
338 the assembled genome using BWA with default parameters. Then, the sorted mapping bam file  
339 was used as input for CNVnator v0.3.3 [38], with the parameters “-unique -his 100 -stat 100 -  
340 call 100.”. The obtained CNVs were filtered, retaining only those where q0 was <0.5 and e-  
341 val1 was <0.05. After filtering, 676 CNVs remained, with a total length of 9,198,900 bp  
342 (**Supplementary Table S12**).

343

### 344 **Non-coding RNA prediction**

345 ~~Non-coding RNA consists of several RNAs, as such ribosomal RNA (rRNA), transfer~~  
346 ~~RNA (tRNA), microRNAs (miRNA) and small nuclear RNA (snRNA). This RNA group~~  
347 ~~mainly plays a regulation role in biological processes. In our study, we detected rRNA from a~~  
348 ~~Human rRNA database with BLASTN command, and the E-value was set as 1E-10. Similarly,~~  
349 ~~miRNAs and snRNAs were searched against the Rfam database [39] with INFERNAL 1.1rc4~~  
350 ~~[40]. The tRNAs were predicted by tRNAscan-SE 1.3.1 software [41]. The numbers of rRNA,~~

351 ~~miRNA, snRNA and tRNA were 608, 17,813, 3,656 and 460, respectively in the genome of~~  
352 ~~the species (Supplementary Table S9).~~

353 Non-coding RNAs included ribosomal RNAs (rRNAs), transfer RNAs (tRNAs),  
354 microRNAs (miRNAs), and small nuclear RNAs (snRNAs). Non-coding RNAs primarily  
355 regulate biological processes. Using BLASTN with an E-value of 1E-10, we identified four  
356 rRNAs in the *R. roxellana* genome homologous to human rRNAs: 28S, 18S, 5.8S, and 5S  
357 (GenBank accession numbers NR\_003287.2, NR\_003286.2, NR\_003285.2, and NR\_023363.1,  
358 respectively). We also searched for miRNAs and snRNAs in the new genome using  
359 INFERNAL v1.1rc4 [40] against the Rfam database release 13.0 [39]. The tRNAs were  
360 predicted by tRNAscan-SE 1.3.1 [41]. We identified 608 rRNAs, 17,813 miRNAs, 3,656  
361 snRNAs, and 460 tRNAs in the *R. roxellana* genome (Supplementary Table S13).

362

### 363 **Gene prediction and functional annotation**

364 ~~We combined prediction methods based on *de novo*, homolog prediction and~~  
365 ~~transcriptome data to estimate genes. As for *ab initio* based prediction, a total of five programs,~~  
366 ~~namely Augustus v. 3.2.2 [42], GlimmeHMM v. 3.0.1 [43], GENSCAN [44], GENEID [45]~~  
367 ~~and SNAP V2013-11-29 [46] were employed to predict protein-coding genes. Subsequently,~~  
368 ~~we used the homolog-based prediction approach. Protein sequences from five homolog species~~  
369 ~~(*Homo sapiens*, *Gorilla gorilla*, *Macaca mulatta*, *Rhinopithecus bieti*, *Rhinopithecus roxellana*~~

370 *hubeiensis*) were downloaded from Ensemble Release 75  
371 (<http://www.ensembl.org/info/data/ftp/index.html>), and used to perform TBLASTN blast  
372 against the repeat-masked genome sequences [47]. The related homologous genome sequences  
373 were then annotated to the matching proteins by GeneWise 2.4.1 [48]. Finally, we estimated  
374 genes based on transcriptome data. During this process, high quality RNAs from heart and skin  
375 tissue were sequenced by an Illumina Novaseq 6000 platform. RNA-seq reads were assembled  
376 with trinityrnaseq 2.1.1. We predicted genes using a combination of approaches: *de novo*,  
377 homology prediction, and transcriptome. For *ab initio* predictions of protein-coding genes, we  
378 used Augustus v3.2.2 [42], with parameters "--uniqueGeneId = true --noInFrameStop = true --  
379 gff3 = on --genemodel = complete --strand = both"; GlimmeHMM v3.0.1 [43], with parameters  
380 "-g -l"; GENSCAN [44], GENEID [45], and SNAP v2013-11-29 [46].

381 Next, we predicted genes using homology-based approach. Protein sequences from five  
382 homologous species (*Homo sapiens*, *Gorilla gorilla*, *Macaca mulatta*, *Rhinopithecus bieti*, and  
383 *Rhinopithecus roxellana hubeiensis*) were downloaded from Ensemble Release 75  
384 (<http://www.ensembl.org/info/data/ftp/index.html>). We compared these sequences to the  
385 repeat-masked *R. roxellana* genome using TBLASTN (-p tblastn -e 1e-05 -F T -m 8 -d) against  
386 the repeat-masked genome sequences [47], with parameters "-p tblastn -e 1e-05 -F T -m 8 -d."  
387 The identified homologous genome sequences were annotated using GeneWise (Version 2.4.1)  
388 [48], with the parameters "-tfor -genesf -gff."

389 Finally, we estimated genes based on transcriptome data. High-quality RNAs from the  
390 heart and skin tissue of the *R. roxellana qinlingensis* specimen were sequenced on an Illumina  
391 Novaseq 6000 platform. RNA-seq reads were assembled using trinityrnaseq-2.1.1 [49]. The  
392 with the parameters “--seqType fq --CPU 20 --max\_memory 200G --normalize reads --  
393 full\_cleanup --min\_glue 2 --min\_kmer\_cov 2 --KMER\_SIZE 25.” To identify validate  
394 transcripts, the assembled transcript sequences were aligned to the *R. roxellana* genome  
395 by using Assemble Spliced Alignment (PASA) [50] with default parameters. In addition, we  
396 estimated the expression levels of transcripts by Tophat 2.0.13 [51] and Cufflinks [52].

397 The genes predicted from those three approaches were merged with EVidenceModeler  
398 [53]. Furthermore, untranslated regions and alternative splicing of those predicted gene sets  
399 were further checked by PASA with the help of transcriptome data, with default parameters.  
400 We estimated transcript expression levels using Tophat 2.0.13 [51] (with the parameters “-p 6  
401 --max-intron-length 500000 -m 2 --library-type fr-unstranded”) and Cufflinks [52].

402 The genes predicted by each of the three approaches were merged using  
403 EVidenceModeler [53] with the parameters “--segmentSize 200000 --overlapSize 20000.” We  
404 weighted transcript predictions most highly, followed by homology-based predictions and *ab*  
405 *initio* predictions. Untranslated regions and alternative splicing of the predicted gene were  
406 explored using PASA, in conjunction with the transcriptome data [50]. Finally, a total of  
407 22497, 22,497 genes were predicted for in the assembly genome of *R. roxellana* genome (Table

408 3), and each of them consisted containing an average of 7.71 exons on average. The detailed  
409 results generated during of the gene prediction process were shown are given in Table 3. And,  
410 and Fig. 3.

411 We also compared the gene prediction evidence based on different methods were shown  
412 in Fig. 3. In addition, we made a comparison between the structure, including mRNA length,  
413 exon length, intron length, and exon number, among *R. roxellana qinlingensis* and other  
414 mammals, suggesting a comparable pattern of the representative primates (e.g., *Homo sapiens*,  
415 *Gorilla gorilla*, *Macaca mulatta*, *Rhinopithecus bieti*, and *Rhinopithecus roxellana hubeiensis*).  
416 We found that genome assembly for patterns were similar among *R. roxellana qinlingensis* and  
417 the other primates (Supplementary Fig. S2).

418 — To have a better understanding the biological functions of those predicted genes, they were  
419 annotated with several databases including NCBI nonredundant protein database (NR),  
420 SwissProt [54], Kyoto Encyclopedia of Genes and Genomes (KEGG) [55], InterPro. To better  
421 understand the biological functions of the predicted genes, we used BLASTP (with an E-value  
422 of 1E-5) to identify the best match for each predicted gene across several databases, including  
423 the NCBI nonredundant protein database (NR v20180129), SwissProt (v20150821) [54],  
424 Kyoto Encyclopedia of Genes and Genomes (KEGG v20160503) [55], InterPro v29.0 [56],  
425 Pfam [57] and GO database [58]. In total, 22053 genes (98.42%) were functionally annotated  
426 (Supplementary Table S10), Pfam v31.0 [57], and GO (Gene Ontology) [58]. In this way,

427 22,053 predicted genes (98.42%) were functionally annotated (Supplementary Table S14).

428 Nearly half (10,670 of 22,497) of these genes were annotated to the predicted proteins in NR

429 database derived from the previous genome annotation for *Rhinopithecus roxellana*.

430 In addition, we estimated the genome assembly completeness using transcriptome data. The

431 transcripts were derived from the *de novo* assembly with trinityrnaseq-2.1.1 mentioned above.

432 Those transcripts were clustered into unigenes with the help of using TGICL (TIGR gene

433 indices clustering program, v2.1) [59] with 95% identity similarity cut-off. The generated

434 unigenes were aligned to our assembly version and previous version using BLAT v. 36. Results

435 showed that the completeness degree (percentage of unigenes aligned to a single scaffold in

436 genome) was higher in our assembly (95.35%) compared with that in previous assembly

437 (89.28%) for unigenes larger than 1000 bp (Supplementary Table S15), demonstrating the

438 contiguity of our new assembly.

#### 439 **Phylogenetic ~~relationship~~ analysis and gene family estimation**

440 ~~— Coding~~The coding regions and protein sequences of 11 representative mammals were

441 downloaded from ~~Ensemble (Ensemble~~Ensembl (Ensembl Release 75). ~~The longest transcript~~

442 ~~was chosen if~~For genes possess many with multiple transcript isoforms, the longest was chosen.

443 Treefam [60] approach was adopted to estimate gene families. Following all-to-all blast, a total

444 ~~of 17,560 gene families were identified. We reconstructed the phylogenetic relationship~~

445 ~~between *R. roxellana* and other mammals based on four fold degenerate sites extracted from~~

446 the 5,418 single-copy gene families. Phyml (version 3.2) [61] was employed to construct a  
447 maximum likelihood tree under the GTR + gamma model that was inferred from  
448 JMODELTEST (version 2.1.10) [62]. Furthermore, we estimated the divergence time with  
449 MCMCTREE in PAML [63]. MCMCTREE was performed on the basis of bayesian method  
450 and the fossil calibration times from timetree were used as input. Generally, the following  
451 calibration times were used: *Homo sapiens* VS *Callithrix jacchus* (40.6–45.7MYA); *Homo*  
452 *sapiens* VS *Pan troglodytes* (6.2–7MYA); *Homo sapiens* VS *Mus musculus* (85–94MYA) and  
453 *Homo sapiens* VS *Tarsius syrichta* (71–77MYA). The reconstructed phylogeny confirmed the  
454 close relationship between *R. rollexana* and *M. mulatta*. Moreover, we estimated that *R.*  
455 *rollexana* and *M. mulatta* diverged approximately 13.4 million years ago (Mya) (Fig. 4).

456 — To have a better understanding the evolutionary history of *R. roxellana*, we estimated the  
457 expansion and contraction of gene family in *R. roxellana* by using CAFE 3.0 [64]. A gene  
458 family with *p* value less than 0.05 was considered for further analysis. As a result, 993 and  
459 2,745 gene families were expanded and contracted in *R. roxellana* genome, respectively (Fig.  
460 4). Its genome showed substantial expansion of gene families which are mainly related to  
461 hemoglobin complex, energy metabolisms and oxygen transport (Supplementary Table S11).

462 — was used to estimate gene families. Using an all-to-all blast, we identified 17,560 gene  
463 families. We reconstructed the phylogenetic relationships among *R. roxellana* and other  
464 mammals based on four-fold degenerate sites extracted from the 5,418 single-copy gene



465 families. Phym1 v3.2 [61] was used to construct a maximum-likelihood tree using the GTR +  
466 gamma model, as inferred by JMODELTEST v2.1.10 [62]. We estimated divergence times  
467 with MCMCTREE in PAML v4.8 [63], using the Bayesian method and the fossil calibration  
468 times from timetree (<http://www.timetree.org/>) [65]. The following fossil calibrations were  
469 used: *H. sapiens* vs. *Callithrix jacchus* (40.6–45.7 MYA, million years ago); *Homo sapiens* vs.  
470 *Pan troglodytes* (~6.2–7 MYA); *Homo sapiens* vs. *Mus musculus* (85–94 MYA); and *Homo*  
471 *sapiens* vs. *Tarsius syrichta* (~71–77 MYA). The reconstructed phylogeny recovered a close  
472 relationship between *R. rollexana* and *M. mulatta*. We estimated that *R. rollexana* and *M.*  
473 *mulatta* diverged approximately 13.4 MYA (Fig. 4).

474 To investigate the evolutionary history of *R. r.*, we estimated the expansion and  
475 contraction of gene family in this species with CAFE 3.0 [64]. A random birth and death model  
476 was used to study gene family variations along each lineage in the phylogenetic tree. This  
477 analysis identified 993 expanded gene families and 2,745 contracted gene families in the *R.*  
478 *roxellana* genome (Fig. 4). To determine the significance of each gene family, *P*-values in each  
479 lineage were estimated by comparing conditional likelihoods derived from a probabilistic  
480 graphical model (PGM). All gene family with *P*-values < 0.05 were further analyzed. To  
481 explore the significantly expanded gene families, we performed a GO-term enrichment analysis  
482 with EnrichPipeline32 [66, 67], using the 1,370 genes belonging to the 314 significantly  
483 expanded gene families as input, and using all predicted genes as background. We considered

484 GO term significant if adjusted the  $P$ -value was  $<0.05$ . We found that the significantly  
485 expanded gene families were mainly associated with the hemoglobin complex, energy  
486 metabolism, and oxygen transport (Supplementary Table S16).  
487

## 488 **Conclusion**

489 — In this study, we generated a high-quality genome assembly ~~of~~for the golden snub-  
490 nosed monkey (*R. roxellana*) ~~by~~using a combination of five advanced technologies. ~~This~~Our  
491 results will ~~be helpful to investigate~~inform studies of the origin and evolutionary history of the  
492 snub-nosed monkey. In addition, ~~the~~this genome may ~~lay~~provide a ~~foundation~~framework  
493 within which to survey the mechanisms ~~about~~underlying the formation of the distinct  
494 morphological and sociological characters ~~and understand the unique multilevel societies in~~  
495 ~~R of R. roxellana. Also, such~~This genome may ~~provide~~also stimulate new insights ~~for~~  
496 amending into the ~~conservation~~improvement of strategies to conserve and ~~management~~  
497 ~~of~~manage this endangered species. ~~Furthermore~~Finally, this genome ~~with, which has~~ superior  
498 continuity and accuracy ~~can provide, may serve as~~ a new standard reference genome for  
499 colobine primates.

500

501

502

503 **Declarations**

504 **Availability of supporting data**

505 The raw data discussed in this publication have been deposited in NCBI's short read archive  
506 under the accession number PRJNA524949. Supporting data are available in the GigaDB  
507 database.

508 **Competing interests**

509 The authors declare that they have no competing interests.

510 **Funding**

511 This work was financially supported by Strategic Priority Research Program of the Chinese  
512 Academy of Sciences (XDB31020302), the National Natural Science Foundation of China  
513 (31622053), the Promotional project for Innovation team, the Department of Science and  
514 Technology of Shaanxi Prov. China (2018TD-017), and the National Key Programme of  
515 Research and Development, the Ministry of Science and Technology of China  
516 (2016YFC0503200).

517 **Abbreviations**

518 Gb: gigabase; kb: kilobase; Mb: megabase; PE: paired-end; PacBio: Pacific Biosciences;  
519 SMRT: single molecule real-time sequencing; Hi-C: high-throughput chromosome  
520 conformation capture; BUSCO: Benchmarking Universal Single-copy Orthologs; GEGMA:  
521 core eukaryotic gene-mapping approach; GO: gene ontology; TFS: transposable element;

522 TRF: Tandem Repeat Finder; SINEs: Short interspersed nuclear elements; LINEs: long  
523 interspersed nuclear elements; PASA: genome by Assemble Spliced Alignment; NR: NCBI  
524 nonredundant protein database; KEGG: Kyoto Encyclopedia of Genes and Genomes. Mya:  
525 million years ago.

#### 526 **Author contributions**

527 X.G.Q. conceived and designed the project, L.W., J.W.W. contributed to the work on genomic  
528 sequencing and performing data analyses. B.G.L. helped with sample collection. L.W., J.W.  
529 W. and X.G.Q. wrote the manuscript. All authors provided input for the paper and approved  
530 the final version.

531

#### 532 **Acknowledgements**

533 We thank Mr. Yiliang Xu, Mr. Qiqi Liang, Mrs. Yue Xie from Novogene for their technical  
534 support. Mr. Xuanmin Guang and Mr. Chi Zhang from BGI for their assistance in data analysis.

535 We thank to Mr. Ruliang Pan for his gracious help polishing the language. We are also grateful  
536 to Mr. Yinghu Lei from Louguantai Breeding Center, Dr. Zhipang Huang, and Dr. Pei Zhang  
537 from Northwest University for their helping with the sampling collection. We specially  
538 appreciate Prof. Zhengbing Wang, and Prof. Jiang Chang from Discipline Development  
539 Department of Northwest University for their support. This study was fundamentally supported  
540 by Discipline Construction Project of Northwest University.

541 **Figures and tables**

542

543 **Figure legends:**

544

545 **Fig. 1. ~~The photoImage~~ of *R. roxellana*, taken in the Qinling ~~mountains-~~Mountain, China.**

546 **Fig. 2. Hi-C heatmap of interactions between ~~chromosomepairs of chromosomal~~ loci**

547 **throughout the genome. Hi-C ~~interactomeinteractions~~ within and among ~~chromosomes of~~**

548 *R. roxellana* chromosomes (Chr1–Chr22); interactions were drawn based on the chromatin

549 interaction frequencies between pairs of 100-kb genomic regions (as determined by Hi-C). In

550 principle, darker red cells indicate stronger and more frequent interactions, which in turn imply

551 that the two sequences are spatially close.

552 ~~Fig. 3. The gene prediction evidence based on different methods.~~ (a). Number of the

553 Fig. 3. Gene predictions. (a) Number of genes estimated by ~~the various~~ prediction approaches

554 ~~based on:~~ de novo (blue-color), homolog prediction, homologys (pink-color), and RNA-seq

555 data (green-color). The labels rna\_0.5, denove\_0.5, and homology\_0.5 ~~indicates those indicate~~

556 the genes predicted by each method with an overlap are larger than ≥50% in each method; %.

557 (b) Number of ~~the genes shown in combination with the prediction~~ genes predicted based on *de*

558 *novo*, homology, and RNA-seq approaches ~~detailed~~, in fig-2a and the addition to expression

559 level ~~standard~~ (in rpkm). The labels rna\_0.5, denovedenove\_0.5, and homology\_0.5 ~~indicates~~

560 ~~that those indicate the~~ genes predicted by each method with an overlap ~~are larger than~~  $\geq 50\%$   
561 ~~in each method.~~%, while  $\text{rpkm} > 1$  indicates those genes with ~~ana relative~~ expression level ~~larger~~  
562 ~~than~~  $\geq 1$ .

563 **Fig. 4. ~~The~~ *R. roxellana* phylogenetic relationships ~~of *R. roxellana* and other mammals and~~**  
564 **~~Gene family analysis in *R. roxellana* genome, and gene families.~~** Phylogenetic ~~relationship~~  
565 ~~was~~ relationships were inferred from 54185,418 single-copy gene families ~~in *R. roxellana* and~~  
566 ~~other mammals.~~ All nodes ~~received 100%~~ had support values ~~The estimated of 100%.~~  
567 Estimated divergence times are ~~indicated~~ given near ~~the nodes.~~ ~~The images in the figure are~~  
568 ~~credited as “Illustrations copyright 2013 Stephen D. Nash / IUCN SSC Primate Specialist~~  
569 ~~Group. Used with permission”.~~ MYA: ~~million years ago each node.~~ Numbers under each  
570 species indicate the number of gene families that have been expanded (green) and contracted  
571 (light yellow) since the split of species from the most recent common ancestor (MRCA). The  
572 numbers on each branch correspond to the numbers of gene families that have been expanded  
573 (red) and contracted (green) in the mammalian genome. ~~MRCA: most recent common ancestor.~~  
574 Those monkey images are copyright 2013 Stephen D. Nash of the IUCN SSC Primate Specialist  
575 Group and are used with permission. MYA: million years ago.

576  
577

578 **Table 1. Reads generated ~~from~~by the five ~~different~~-sequencing methods.**

<del>Pair</del> Paired-end libraries	Insert size (bp)	Total clean data (Gb)	Read length (bp)	Sequence coverage (X)
Illumina <del>reads</del>	350	423.32	150	133.12
Pacbio <del>reads</del>	20 k	304.84	<del>\n/a</del>	95.86
10X Genomics	500 <del>—</del> 700	348.41	150	109.56
<del>Bionano</del> BioNano	<del>\n/a</del>	463.75	<del>\n/a</del>	<del>\n/a</del>
Hi-C	350	310.92	<del>\n/a</del>	97.77
Total	<del>\n/a</del>	1,851.24	<del>\n/a</del>	582.15

579 Note: The sequence coverage was calculated ~~with~~based on an estimated genome size of 3.18

580 Gb. ~~The sign of backslash indicates that the insert size was absent.~~~~\n/a: not applicable.~~

581

582 Table 2. ~~The Summary of the final *R. roxellana* genome assembly statistics of *R. roxellana*.~~

Category	Contig		Scaffold	
	<del>length</del> Length (bp)	<del>number</del> Number	Length (bp)	<del>number</del> Number
Total	3,038,184,325	6,099	3,038,467,325	3,269
Max	30,757,641	<del>\n/a</del>	206,558,726	<del>\n/a</del>
<del>&gt;=</del> ≥2000 bp	<del>\n/a</del>	5,708	<del>\n/a</del>	2,879
N50	5,723,610	151	144,559,847	9
N60	4,241,389	211	141,075,955	11
N70	3,173,235	292	135,203,321	14
N80	2,063,823	408	118,350,466	16
N90	896,517	622	83,045,532	19

583 Note: The “Number” column represents the number ~~indicated those of~~ contigs/scaffolds  
584 ~~larger longer~~ than the ~~length value~~ of ~~it~~the corresponding category. ~~The sign of backslash~~  
585 ~~indicates that the length/number was absent.~~ n/a: not applicable.

586

587



588 **Table 3. Summary of ~~and characteristics of the~~ predicted protein-coding genes ~~and their~~**  
 589 **~~characteristics.~~**

Gene set	Number	Average transcript length (bp)	Average CDS length (bp)	Average <del>intron</del> exon length (bp)	Average <del>intron</del> exon length (bp)	Average exons per gene
Augustus	32,928	23,441	1,052	196	5,112	5.38
GlimmerHMM	618,957	4,204	404	166	2,654	2.43
<i>De novo</i> SNAP	97,298	49,851	755	144	1,1597	5.23
Geneid	36,863	35,242	1,035	188	7,615	5.49
Genscan	50,419	40,635	1,137	167	6,800	6.81
Ggo	25,281	19,893	1,055	184	3,971	5.74
Hsa	38,444	14,763	826	182	3,942	4.54
Homology Mmu	21,959	29,709	1,470	187	4,123	7.85
Rbi	25,320	25,685	1,387	196	3,991	7.09
Rro	24,121	28,439	1,420	185	4,043	7.68
RNASeq PASA	66,620	28,449	1,219	164	4,247	7.41
Cufflinks	73,199	31,497	2,737	409	5,052	6.69
EVM	30,102	22,298	1,098	182	4,199	6.05
Pasa-update*	29,403	27,638	1,180	181	4,782	6.53
Final set*	22,497	34,153	1,369	178	<del>48854.885</del>	7.71

590 Note: Pasa-update\* ~~indicates~~includes only the UTRs (~~Untranslated~~untranslated regions) ~~and~~  
 591 ~~were considered during the filter process, and;~~ other regions were not included. Final set\*  
 592 ~~indicates~~represents the results ~~were acquired following~~after the Pasa-update filtering process,  
 593 ~~with the criteria of~~where the longest isoform was chosen if ~~there were~~the case of multiple  
 594 splicing isoforms, ~~and the;~~ redundant single exons were also discarded. ~~Number indicates~~The  
 595 ~~“Number” column gives~~ the ~~specific values~~number of protein-coding genes predicted by each  
 596 method.

597

598

599 **Supplementary files:**

600 **Supplementary Fig. S1.** Genome size estimation using the k-mer method.

601 **Supplementary Fig. S2.** ~~The comparison~~Comparisons of each element ~~in the genome~~among  
602 genomes of homologous species.

603 **Supplementary Table S1.** The ~~results of~~contig assembly ~~with~~based on PacBio ~~long reads~~  
604 ~~and gap filling~~subreads.

605 **Supplementary Table S2.** The ~~results of~~scaffold assembly ~~with Bionano optical map~~  
606 ~~data~~based on sspace-longreads results.

607 **Supplementary Table S3.** The ~~results of~~ assembly ~~with 10X Genomics link reads~~after gap-  
608 filling.

609 **Supplementary Table S4.** The ~~mapping rate of reads and coverage of assembled genome~~  
610 ~~with BWA~~assembly based on BioNano optical map data.

611 **Supplementary Table S5.** ~~Assessment results by using BUSCO annotation~~The assembly  
612 based on 10X Genomics linked reads.

613 **Supplementary Table S6.** The ~~completeness test results~~read mapping rate and the coverage  
614 of the assembled genome determined with ~~CEGMA software~~BWA.

615 **Supplementary Table S7.** ~~Results of repeats elements predictions from~~The SNPs identified  
616 in the genome assembly of *R. roxellana*.

617 **Supplementary Table S8.** ~~The results of TEs elements predicted from the genome~~  
618 assemblyGenome assessment based on BUSCO annotations.

619 **Supplementary Table S9.** ~~Summary of predicted RNAs and their characteristics~~Genome  
620 assessment based on CEGMA annotations.

621 **Supplementary Table S10.** ~~The functional annotation~~Prediction of repeat elements prediction  
622 in the genes predicted from *R. roxellana* genome assembly.

623 **Supplementary Table S11.** ~~The GO annotation results~~Prediction of expansion gene  
624 familiesrepetitive sequences in the genome assembly.

625 **Supplementary Table S12.** The CNVs identified in the genome assembly.

626 **Supplementary Table S13.** Summary and characteristics of the predicted RNAs.

627 **Supplementary Table S14.** The functional annotations of the genes predicted in the *R.*  
628 *roxellana* genome.

629 **Supplementary Table S15.** Assessment of the new genome assembly using unigenes.  
630 sequences

631 **Supplementary Table S16.** The GO annotations of the expanded gene families in the *R.*  
632 *roxellana* genome (adjusted *P*-value < 0.05)

633 **References**

- 634 1. Li BG, Pan RL, Oxnard CE. Extinction of snub-nosed monkeys in China during the  
635 past 400 years. *Int J Primatol*, 2002; **23**(6):1227-1244.
- 636 2. Luo MF, Liu ZJ, Pan HJ, Zhao L, Li M. Historical geographic dispersal of the golden  
637 snub-nosed monkey (*Rhinopithecus roxellana*) and the influence of climatic  
638 oscillations. *Am J Primatol*, 2012; **74**(2):91-101.
- 639 3. Fang G, Li M, Liu X-J, Guo W-J, Jiang Y-T, Huang Z-P, Tang S-Y, Li D-Y, Yu J, Jin  
640 T *et al.* Preliminary report on Sichuan golden snub-nosed monkeys (*Rhinopithecus*  
641 *roxellana roxellana*) at Laohegou Nature Reserve, Sichuan, China. *Sci Rep*, 2018;  
642 **8**(1):16183.
- 643 4. Grueter CC, Qi X-G, Li B, Li M. Multilevel societies. *Curr Biol*, 2017; **27**(18):R984-  
644 R986.
- 645 5. Qi X-G, Li BG, Garber PA, Ji WH, Watanabe K. Social dynamics of the golden snub-  
646 nosed monkey (*Rhinopithecus roxellana*): female transfer and one-male unit  
647 succession. *Am J Primatol*, 2009; **71**(8):670-679.
- 648 6. Li H, Meng S-J, Men Z-M, Fu Y-X, Zhang Y-P. Genetic diversity and population  
649 history of golden monkeys (*Rhinopithecus roxellana*). *Genetics*, 2003; **164**(1):269-  
650 275.
- 651 7. Qi X-G, Garber PA, Ji W, Huang Z-P, Huang K, Zhang P, Guo S-T, Wang X-W, He  
652 G, Zhang P *et al.* Satellite telemetry and social modeling offer new insights into the  
653 origin of primate multilevel societies. *Nat Commun*, 2014; **5**:5296.
- 654 8. Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang  
655 JW, Fulton L, Graves TA *et al.* The B73 Maize Genome: Complexity, Diversity, and  
656 Dynamics. *Science*, 2009; **326**(5956):1112-1115.
- 657 9. Seim I, Fang X, Xiong Z, Lobanov AV, Huang Z, Ma S, Feng Y, Turanov AA, Zhu  
658 Y, Lenz TL *et al.* Genome analysis reveals insights into physiology and longevity of  
659 the Brandt's bat *Myotis brandtii*. *Nat Commun*, 2013; **4**:2212.
- 660 10. Valenzano DR, Benayoun BA, Singh PP, Zhang E, Etter PD, Hu C-K, Clément-Ziza  
661 M, Willemsen D, Cui R, Harel I *et al.* The African turquoise killifish genome  
662 provides insights into evolution and genetic architecture of lifespan. *Cell*, 2015;  
663 **163**(6):1539-1554.
- 664 11. Zhou X, Wang B, Pan Q, Zhang J, Kumar S, Sun X, Liu Z, Pan H, Lin Y, Liu G *et al.*  
665 Whole-genome sequencing of the snub-nosed monkey provides insights into folivory  
666 and evolutionary history. *Nat Genet*, 2014; **46**:1303-1310.
- 667 12. Kuang W-M, Ming C, Li H-P, Wu H, Frantz L, Roos C, Zhang Y-P, Zhang C-L, Jia  
668 T, Yang J-Y *et al.* The origin and population history of the endangered golden snub-  
669 nosed monkey (*Rhinopithecus roxellana*). *Mol Biol Evol*, 2018:msy220-msy220.

- 670 13. Hong YY, Duo HR, Hong JY, Yang JY, Liu SM, Yu LH, Yi TY. Resequencing and  
671 comparison of whole mitochondrial genome to gain insight into the evolutionary  
672 status of the Shennongjia golden snub-nosed monkey (*SNJ R-roxellana*). *Ecol Evol*,  
673 2017; **7**(12):4456-4464.
- 674 14. Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, Hastie A, Cao H, Yun JY, Kim J  
675 *et al.* De novo assembly and phasing of a Korean human genome. *Nature*, 2016;  
676 **538**(7624):243-247.
- 677 15. Chaisson MJ, Wilson RK, Eichler EE. Genetic variation and the de novo assembly of  
678 human genomes. *Nat Rev Genet*, 2015; **16**(11):627-640.
- 679 16. Jiao YP, Peluso P, Shi JH, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC,  
680 Wei XH, Chin CS *et al.* Improved maize reference genome with single-molecule  
681 technologies. *Nature*, 2017; **546**(7659):524-527.
- 682 17. Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM,  
683 Malig M, Raja A, Fiddes I, Hillier LW *et al.* Long-read sequence assembly of the  
684 gorilla genome. *Science*, 2016; **352**(6281):aae0344.
- 685 18. Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS,  
686 Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML *et al.* High-resolution  
687 comparative analysis of great ape genomes. *Science*, 2018; **360**(6393):eaar6343.
- 688 19. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET,  
689 Liachko I, Sullivan ST *et al.* Single-molecule sequencing and chromatin conformation  
690 capture enable de novo reference assembly of the domestic goat genome. *Nat Genet*,  
691 2017; **49**:643-650.
- 692 20. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y *et al.*  
693 SOAPdenovo2: an empirically improved memory-efficient short-read de novo  
694 assembler. *Gigascience*, 2012; **1**(1):18.
- 695 21. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing  
696 reads. *EMBnetjournal*, 2011; **17**:10-12.
- 697 22. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A,  
698 Copeland A, Huddleston J, Eichler EE *et al.* Nonhybrid, finished microbial genome  
699 assemblies from long-read SMRT sequencing data. *Nat Methods*, 2013; **10**(6):563-+.
- 700 23. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA,  
701 Zeng QD, Wortman J, Young SK *et al.* Pilon: An Integrated Tool for Comprehensive  
702 Microbial Variant Detection and Genome Assembly Improvement. *Plos One*, 2014;  
703 **9**(11).
- 704 24. Adey A, Kitzman JO, Burton JN, Daza R, Kumar A, Christiansen L, Ronaghi M,  
705 Amini S, Gunderson KL, Steemers FJ *et al.* In vitro, long-range sequence information  
706 for de novo genome assembly via transposase contiguity. *Genome Res*, 2014;  
707 **24**(12):2041-2049.

- 708 25. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler  
709 transform. *Bioinformatics*, 2009; **25**(14):1754-1760.
- 710 26. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-  
711 scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat*  
712 *Biotechnol*, 2013; **31**(12):1119-1125.
- 713 27. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO:  
714 assessing genome assembly and annotation completeness with single-copy orthologs.  
715 *Bioinformatics*, 2015; **31**(19):3210-3212.
- 716 28. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in  
717 eukaryotic genomes. *Bioinformatics*, 2007; **23**(9):1061-1067.
- 718 29. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichewicz J.  
719 Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome*  
720 *Res*, 2005; **110**(1-4):462-467.
- 721 30. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large  
722 genomes. *Bioinformatics*, 2005; **21 Suppl 1**:i351-358.
- 723 31. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic*  
724 *Acids Res*, 1999; **27**(2):573-580.
- 725 32. Chin CS, Peluso P, Sedlazeck FJ. Phased diploid genome assembly with single-  
726 molecule real-time sequencing. *Nat Methods*, 2016; **13**(12):1050-1054.
- 727 33. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes  
728 using long read sequence information. *BMC Bioinformatics*, 2014; **15**:211.
- 729 34. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG,  
730 Worley KC *et al*. Mind the gap: upgrading genomes with Pacific Biosciences RS  
731 long-read sequencing technology. *Plos One*, 2012; **7**(11):e47768.
- 732 35. Shelton JM, Coleman MC, Herndon N, Lu N, Lam ET, Anantharaman T, Sheth P,  
733 Brown SJ. Tools and pipelines for BioNano data: molecule assembly pipeline and  
734 FASTA super scaffolding tool. *BMC Genomics*, 2015; **16**:734-734.
- 735 36. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient  
736 alignment of short DNA sequences to the human genome. *Genome Biol*, 2009;  
737 **10**(3):R25.
- 738 37. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg  
739 SL. Versatile and open software for comparing large genomes. *Genome Biol*, 2004;  
740 **5**(2):R12.
- 741 38. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover,  
742 genotype, and characterize typical and atypical CNVs from family and population  
743 genome sequencing. *Genome Res*, 2011; **21**(6):974-984.

- 744 39. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam:  
745 annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, 2005;  
746 **33**:D121-D124.
- 747 40. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments (vol  
748 25, pg 1335, 2009). *Bioinformatics*, 2009; **25**(13):1713-1713.
- 749 41. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer  
750 RNA genes in genomic sequence. *Nucleic Acids Res*, 1997; **25**(5):955-964.
- 751 42. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab  
752 initio prediction of alternative transcripts. *Nucleic Acids Res*, 2006; **34**:W435-W439.
- 753 43. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source  
754 ab initio eukaryotic gene-finders. *Bioinformatics*, 2004; **20**(16):2878-2879.
- 755 44. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J*  
756 *Mol Biol*, 1997; **268**(1):78-94.
- 757 45. Guigo R. Assembling genes from predicted exons in linear time with dynamic  
758 programming. *J Comput Biol*, 1998; **5**(4):681-702.
- 759 46. Korf I. Gene finding in novel genomes. *BMC Bioinformatics*, 2004; **5**:59.
- 760 47. Kent WJ. BLAT - The BLAST-like alignment tool. *Genome Res*, 2002; **12**(4):656-  
761 664.
- 762 48. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. *Genome Res*, 2004;  
763 **14**(5):988-995.
- 764 49. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X,  
765 Fan L, Raychowdhury R, Zeng QD *et al*. Full-length transcriptome assembly from  
766 RNA-Seq data without a reference genome. *Nat Biotechnol*, 2011; **29**(7):644-U130.
- 767 50. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R,  
768 Ronning CM, Rusch DB, Town CD *et al*. Improving the Arabidopsis genome  
769 annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*, 2003;  
770 **31**(19):5654-5666.
- 771 51. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate  
772 alignment of transcriptomes in the presence of insertions, deletions and gene fusions.  
773 *Genome Biol*, 2013; **14**(4):R36.
- 774 52. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg  
775 SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-  
776 seq experiments with TopHat and Cufflinks. *Nat Protoc*, 2012; **7**(3):562-578.
- 777 53. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR,  
778 Wortman JR. Automated eukaryotic gene structure annotation using  
779 EVIDENCEModeler and the program to assemble spliced alignments. *Genome Biol*,  
780 2008; **9**(1):R7.

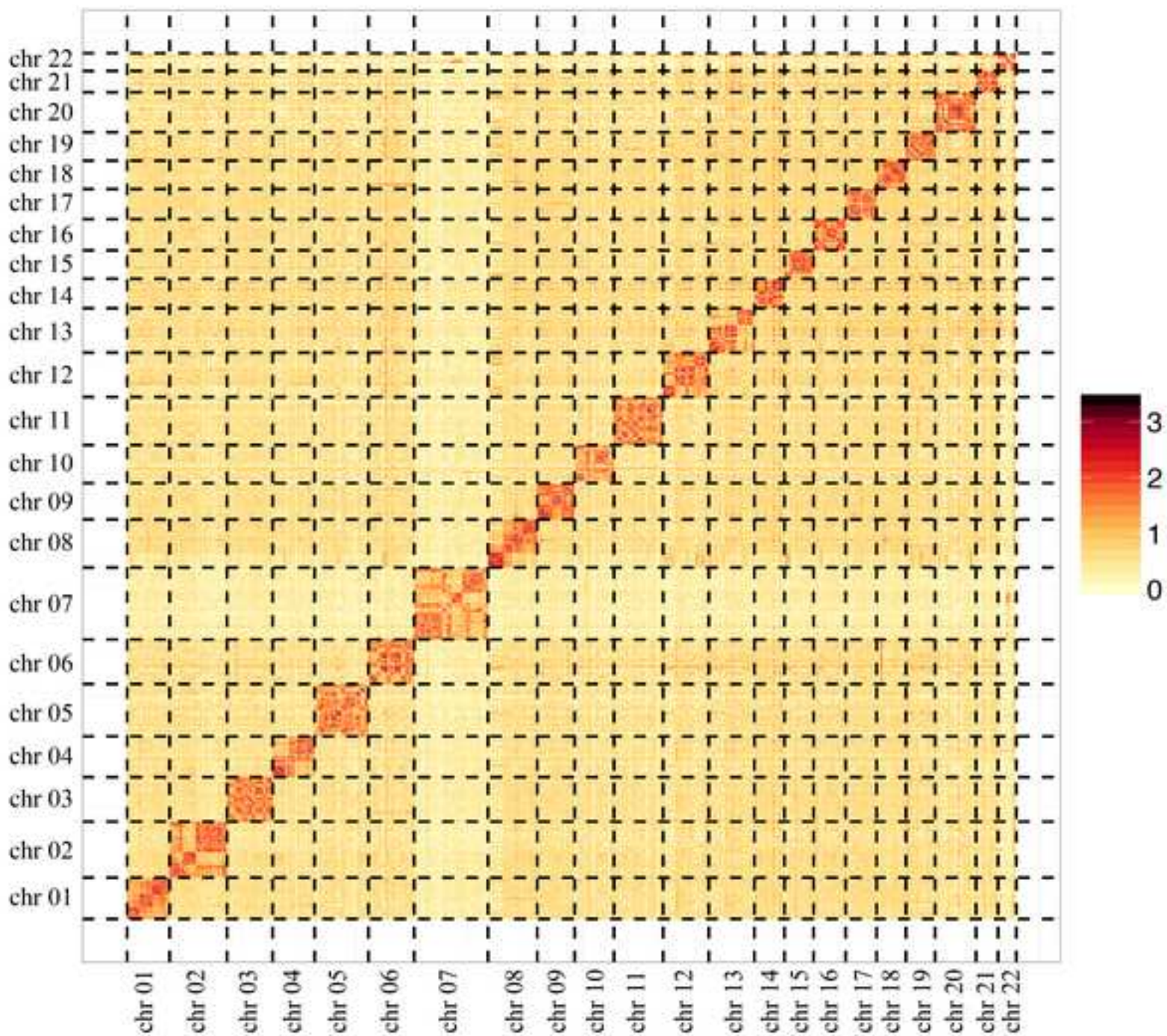
- 781 54. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E,  
782 Huang H, Lopez R, Magrane M *et al.* The Universal Protein Resource (UniProt).  
783 *Nucleic Acids Res*, 2005; **33**(Database issue):D154-D159.
- 784 55. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic*  
785 *Acids Res*, 2000; **28**(1):27-30.
- 786 56. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD,  
787 Chang HY, El-Gebali S, Fraser MI *et al.* InterPro in 2019: improving coverage,  
788 classification and access to protein sequence annotations. *Nucleic Acids Res*, 2019;  
789 **47**(D1):D351-D360.
- 790 57. Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta  
791 M, Qureshi M, Sangrador-Vegas A *et al.* The Pfam protein families database: towards  
792 a more sustainable future. *Nucleic Acids Res*, 2016; **44**(D1):D279-D285.
- 793 58. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP,  
794 Dolinski K, Dwight SS, Eppig JT *et al.* Gene Ontology: tool for the unification of  
795 biology. *Nat Genet*, 2000; **25**(1):25-29.
- 796 59. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White  
797 J, Cheung F, Parvizi B *et al.* TIGR Gene Indices clustering tools (TGICL): a software  
798 system for fast clustering of large EST datasets. *Bioinformatics*, 2003; **19**(5):651-652.
- 799 60. Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, Li RQ, Liu T, Zhang  
800 Z, Bolund L *et al.* TreeFam: a curated database of phylogenetic trees of animal gene  
801 families. *Nucleic Acids Res*, 2006; **34**:D572-D580.
- 802 61. Guindon S, Delsuc F, Dufayard J-F, Gascuel O: **Estimating maximum likelihood**  
803 **phylogenies with PhyML**. In: *Bioinformatics for DNA sequence analysis*. Springer; 2009:  
804 113-137.
- 805 62. Posada D. jModelTest: phylogenetic model averaging. *Mol Biol Evol*, 2008;  
806 **25**(7):1253-1256.
- 807 63. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol*,  
808 2007; **24**(8):1586-1591.
- 809 64. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the  
810 study of gene family evolution. *Bioinformatics*, 2006; **22**(10):1269-1271.
- 811 65. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A resource for timelines,  
812 timetrees, and divergence times. *Mol Biol Evol*, 2017; **34**(7):1812-1819.
- 813 66. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths  
814 toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*,  
815 2009; **37**(1):1-13.
- 816 67. Beissbarth T, Speed TP. GOstat: find statistically overrepresented Gene Ontologies  
817 within a group of genes. *Bioinformatics*, 2004; **20**(9):1464-1465.



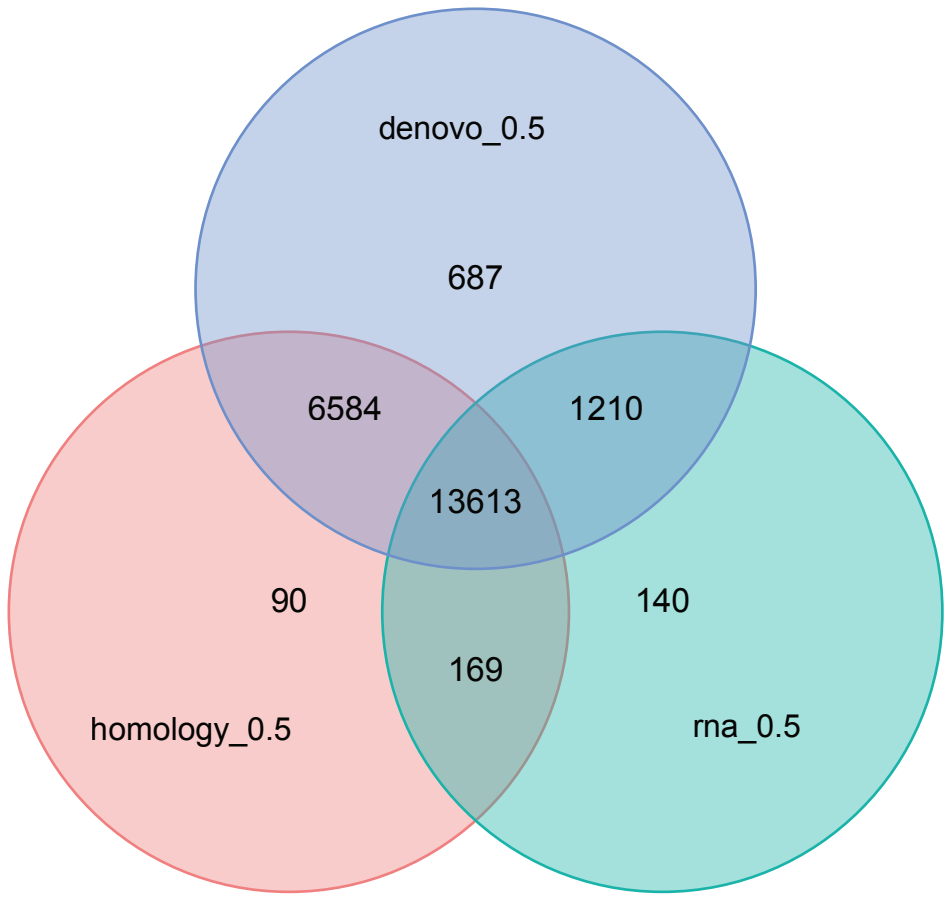


Figure 2

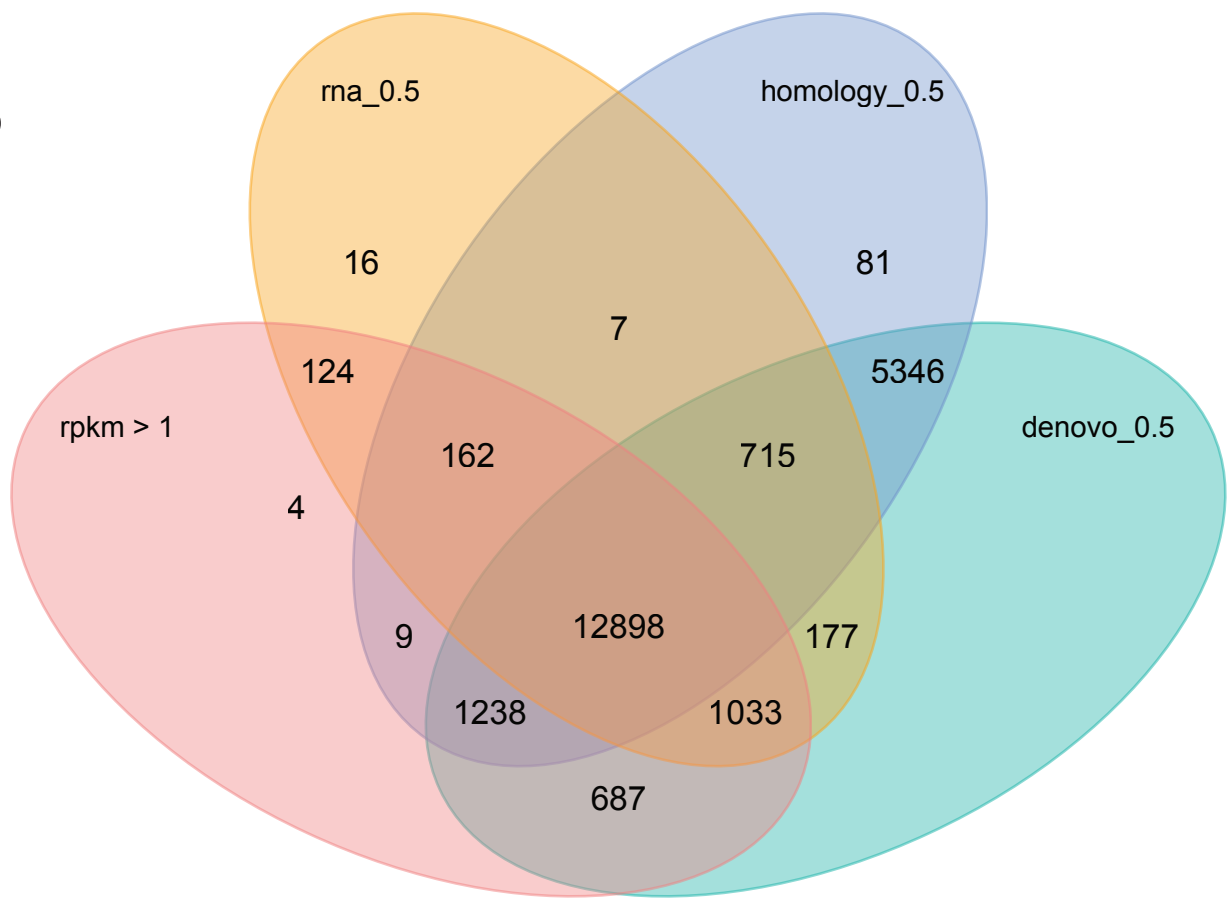
[Click here to access/download;Figure;fig. 2\\_HIC.png](#)

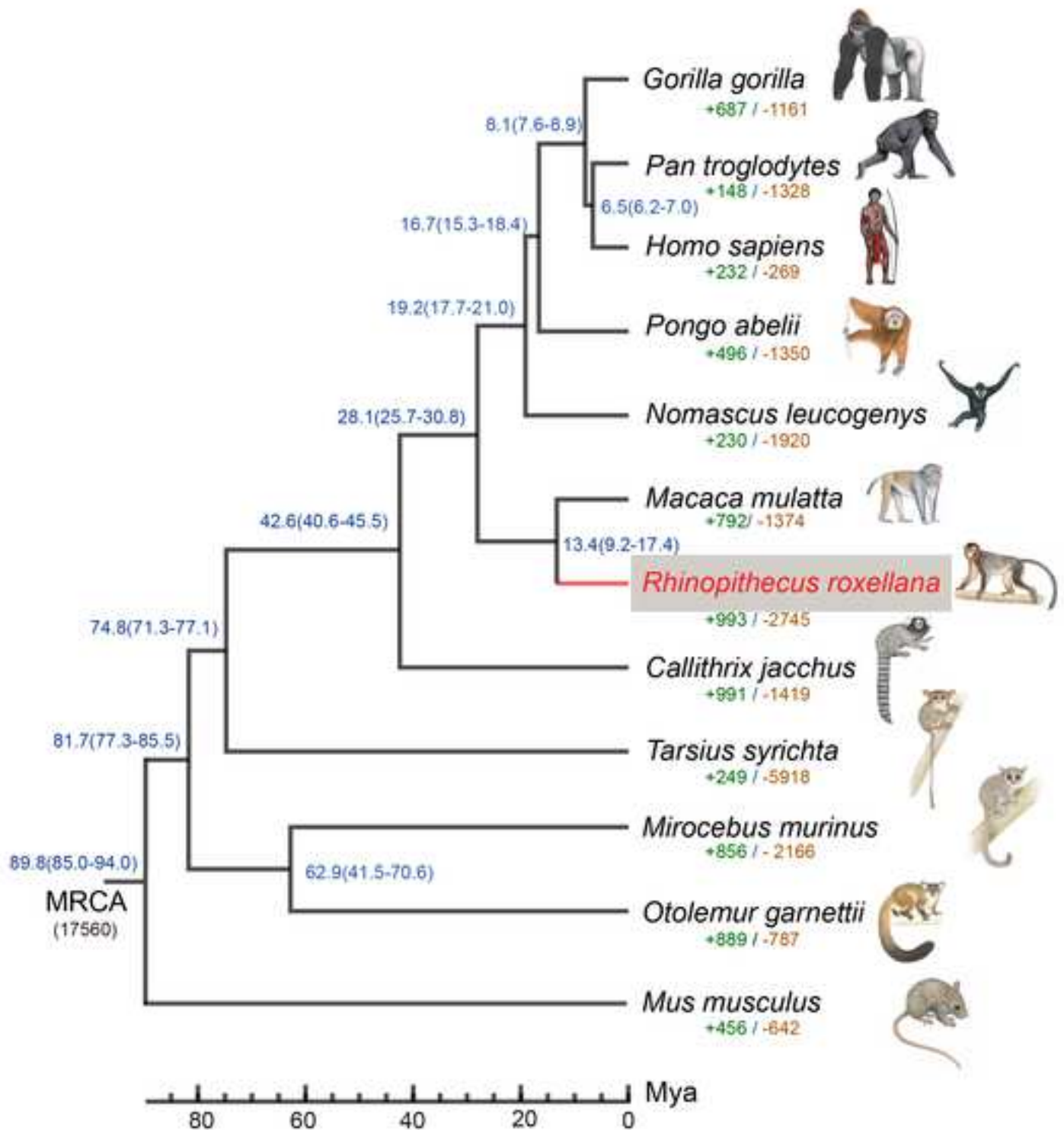



a



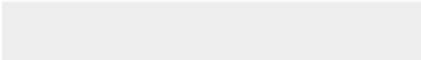

b







Click here to access/download  
**Supplementary Material**  
SI\_giga\_0505.docx



Dear Scott and Handling Editor,

Thanks for handling our manuscript, and we appreciate the valuable comments from you and three referees.

After digesting these comments, we have substantially revised our manuscript. Firstly, we employed an English-language editing service, LetPub, to polish our wording. Secondly, we expanded the methodological details substantially including analysis procedures, software versions and settings, we also capture these details using protocols.io.. as recommended. Thirdly, we added some key details about the generated data, including sequencing data, calibration times and N50 length to be more clear. Forth, we performed several additional analysis including CNVs identification, synteny analysis and SNP calling et al. as reviewers suggested. In addition, other comments were also addressed following the instructions from you and three referees.

In this revised version, corrections were made in a document with “Track Changes” mode. Point-by-point responses to the reviewers are also submitted. After addressing the issues raised, we feel the quality of the paper is much improved and hope that our revised manuscript is acceptable for publication in *GigaScience*.

Thanks for your consideration, we look forward to your advice.

Yours sincerely,

Xiao-Guang Qi

Shaanxi Key Laboratory for Animal Conservation

College of Life Sciences

Northwest University

Email: qixg@nwu.edu.cn