genes

MDPI

Article

The Caribou (*Rangifer tarandus*) Genome

Rebecca S. Taylor [1,*], Rebekah L. Horn [1], Xi Zhang [2], G. Brian Golding [2], Micheline Manseau [1,3] and Paul J. Wilson [1]

1 Biology Department, Trent University, 1600 West Bank Drive, Peterborough, ON K9J 7B8, Canada
2 Department of Biology, McMaster University, 1280 Main St. West, Hamilton, ON L8S 4K1, Canada
3 Science and Technology Division, Environment and Climate Change Canada, 1125 Colonel By Drive, Ottawa, ON K1S 5R1, Canada
* Correspondence: becky.taylor3112@gmail.com

Received: 31 May 2019; Accepted: 16 July 2019; Published: 17 July 2019

Abstract: *Rangifer tarandus*, known as caribou or reindeer, is a widespread circumpolar species which presents significant variability in their morphology, ecology, and genetics. A genome was sequenced from a male boreal caribou (*R. t. caribou*) from Manitoba, Canada. Both paired end and Chicago libraries were constructed and sequenced on Illumina platforms. The final assembly consists of approximately 2.205 Gb, and has a scaffold N50 of 11.765 Mb. BUSCO (Benchmarking Universal Single-Copy Orthologs) reconstructed 3820 (93.1%) complete mammalian genes, and genome annotation identified the locations of 33,177 protein-coding genes. An alignment to the bovine genome was carried out, indicating sequence coverage on all bovine chromosomes. A high-quality reference genome will be invaluable for evolutionary research and for conservation efforts for the species. Further information about the genome, including a FASTA file of the assembly and the annotation files, is available on our caribou genome website. Raw sequence data is available at the National Centre for Biotechnology Information (NCBI), under the BioProject accession number PRJNA549927.

Keywords: caribou; reindeer; *Rangifer tarandus*; genome; genome assembly

1. Introduction

Rangifer tarandus, known as caribou in North America and reindeer in Europe and Asia, is the most widespread circumpolar ungulate species [1]. The species occurs in a variety of ecozones, including High Arctic, taiga, mountains, and boreal, and as such are hugely variable in their morphology, ecology, and genetics [1]. In Canada, caribou are declining due to a number of stressors, including anthropogenic disturbances and climate changes [2,3]. Even though many caribou populations fluctuate over time, the current declines appear to be surpassing the ability of many herds to recover [3]. Consequently, caribou are a conservation concern in most of Canada [3,4].

Although recognized as one species across its vast circumpolar range, caribou has a complex history and existed in multiple refugia during the Pleistocene leading to three main lineages—a Beringian, a North American and a High Arctic lineage [5,6]. Currently, caribou are divided into multiple subspecies, the number of which has been disputed but with nine listed in Banfield's often cited revision of their taxonomy [1]. Taxonomic clarification was provided by COSEWIC in 2011 [4] given the considerable variability even within some of the designated subspecies. In Canada, caribou are currently divided into 11 extant and 1 extinct Designatable Units (DUs) to ensure the conservation of all caribou diversity [4]; however, further research is essential to clarify the delineation and evolutionary history of caribou groups and to elucidate functional genomic regions underlying ecological adaptation. To help achieve this goal, we have sequenced a high-quality caribou reference genome from a male boreal caribou (*R. t. caribou*; COSEWIC DU6) from Snow Lake, Manitoba, Canada.

2. Materials and Methods

Neck muscle tissue was collected from an adult male boreal caribou (*R. t. caribou)*, which had been killed on a road in Manitoba in October 2009. The tissue was stored in RNA later ICE (Thermo Fisher Scientific, MA, USA). Phenol chloroform extraction [7] was performed using 0.2 g of tissue, and eluted in Tris-ethylenediaminetetraacetic acid (TE) buffer at 100 µL. The DNA was shipped to Dovetail Genomics for library preparation, sequencing and assembly.

Three Chicago libraries were prepared as described previously elsewhere [8]. Briefly, for each library, ~500 ng of high molecular weight genomic DNA (gDNA; mean fragment length = 85 kb) was reconstituted into chromatin in vitro and fixed with formaldehyde. Fixed chromatin was digested with DpnII, the 5′ overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters (New England BioLabs, Ipswich, MA, USA). Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeq X (Illumina, San Diego, CA, USA). The number and length of read pairs produced for each library was: 123 million, 2 × 101 bp for library 1; 66 million, 2 × 101 bp for library 2; and 125 million, 2 × 101 bp for library 3. Together, these Chicago library reads provided 50.8 × physical coverage of the genome (1–50 kb pairs).

A de novo assembly was constructed using a combination of paired end reads (mean insert sizes ~350 bp and 550 bp), which were sequenced on an Illumina HiSeq2500 (Illumina, San Diego, CA, USA) and an Illumina HiSeq X (Illumina, San Diego, CA, USA), respectively. De novo assembly was performed using Meraculous (version 2.2.2.5) [9] with a *kmer* (*k*) size of 43. The input data consisted 1.51 billion read pairs sequenced from paired end libraries (totaling 453 Gbp). Reads were trimmed for quality, sequencing adapters, and mate pair adapters using Trimmomatic [10].

The input de novo assembly, shotgun reads, and Chicago library reads were used as input data for HiRise, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies [8]. Shotgun and Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of Chicago read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold. After scaffolding, shotgun sequences were used to close gaps between contigs. Raw sequence data is available at the National Centre for Biotechnology Information (NCBI), under the BioProject accession number PRJNA549927. Mitochondrial DNA is not included in the assembly as a full mitogenome assessment is currently underway with additional samples and will be released in the future.

We used the gene prediction program AUGUSTUS 2.5.5 [11] to annotate the genome using predictions based on human genes. The genome was masked using RepeatMasker 3.2.6 [12] and run in Augustus using a partial gene model allowing the prediction of incomplete genes at the sequence boundaries.

We ran the final genome FASTA file through the stats.sh function of BBMap 38.42 [13] to calculate genome statistics such as the N50. We also used BUSCO (Benchmarking Universal Single-Copy Orthologs; [14]) to reconstruct 4104 conserved mammalian genes to assess genome completeness. We aligned the caribou genome to the bovine reference genome, as it is the most closely related (estimated to share a common ancestor around 25.8 million years ago [15]), highest quality reference genome. We downloaded the FASTA sequence from the bovine genome database (bovinegenome.org [16]), and aligned it to our genome using BWA-MEM [17]. Using the alignment, we created a Jupiter plot using the script written by J. Chu [18] and Circos 0.69-3 [19]. We plotted the largest scaffolds covering 75% of the bovine genome (as the figure becomes crowded and unclear when showing more) to assess synteny between the two genomes. We also downloaded the consensus FASTA sequence for a previously

published reindeer genome from Inner Mongolia [20] and did an alignment in the same way with the bovine genome. We used QUAST 5.0.2 [21] to assess the quality of both our caribou assembly and the reindeer assembly in comparison to the bovine genome.

3. Results and Discussion

The final *Rangifer tarandus* genome assembly consists of approximately 2.205 Gb, with a scaffold N50 of 11,765,000 base pairs (Table 1), and a GC content of 41.44% (Table 2). AUGUSTUS identified the locations of 33,177 protein-coding genes, and BUSCO indicated the presence of 3820 (93.1%) complete mammalian genes of the 4104 searched for. Our quality assessment statistics are similar to those of other recent non-model mammal species genome assemblies, including the American brown bear (*Ursos arctos* ssp. *horribilis*) [22]; the beluga whale (*Delphinapterus leucas*) [23]; and the northern sea otter (*Enhydra lutris kenyoni*) [24]. The previously published *Rangifer tarandus* genome, sequenced from a domesticated individual from Inner Mongolia [20], consists of 58,765 scaffolds, with a scaffold N50 of 986, 392 bp, and successfully reconstructed 92.6% of the BUSCO genes, indicating our genome to be a more contiguous assembly. We used the Chicago method which produces proximity ligation libraries that have a relationship between within-read pair distance and read count. This produces long-range sequence scaffolds during the assembly of genomes [8], and increased our scaffold contiguity compared to the reindeer.

Table 1. Assembly statistics of the caribou genome.

Statistic	*Rangifer tarandus* genome
Scaffold sequence total (bp)	22.052×10^8
Number scaffolds	4699
Scaffold N50 (bp)	11.765×10^6
Scaffold L50	52
Scaffold N90 (bp)	89.704×10^4
Scaffold L90	289
Contig sequence total (bp)	21.893×10^8
Number contigs	146,562
Contig N50 (bp)	32.819×10^3
Contig L50	19,701
Contig N90 (bp)	89.140×10^2
Contig L90	68,199

Table 2. Nucleotide base composition of the caribou genome assembly statistics of the caribou genome.

A	C	G	T	N
29.27%	20.72%	20.73%	29.28%	0.72%

The reindeer genome is larger, at 2.64 Gb, and so may cover more of the genome in total. However, QUAST results indicated that the missing data is much higher for the reindeer genome, with 3.6% of their bases as N's, whereas our assembly consists of 0.7% N's. Similarly, the reindeer annotation recovered fewer genes than we did [20], which may be because our annotation does not account for pseudogenes or incomplete proteins. However, it could also be because their assembly is fragmented with a higher percentage of missing data, which may have impacted the detection of genes. Using QUAST, our caribou assembly recovered 8402 genes from the Bovine annotation, whereas the reindeer recovered 5755. In addition, we recovered more conserved genes during the BUSCO analysis, suggesting the difference in the number of genes recovered during genome annotation is related to the differences in genome contiguity.

Our genome sequence is the first North American *Rangifer tarandus* (caribou) genome, and is from a wild animal which is important as genetic differences have been found between domesticated and wild reindeer [25]. Therefore, both genomes likely represent important but different genomic variation.

As one of our primary aims is to use the genome to aid with the conservation of wild populations, our assembly represents a valuable resource.

The Jupiter plot displays the largest 312 caribou scaffolds, out of a total of 4699, which cover 75% of the bovine genome (Figure 1). The coloured bands represent synteny between the caribou and bovine assemblies. The lines crossing the circle could be genomic rearrangements, but also likely represent break points in the assembly, particularly when appearing at the edges of scaffolds. The BWA results indicated that the sex chromosomes appear on smaller scaffolds within the caribou assembly, explaining why there are few alignments showing in the Jupiter plot. This reflects the difficulty in assembling large contigs and scaffolds for the sex chromosomes due to their highly repetitive nature [26]. Overall, the BWA alignment and Jupiter plot show good synteny between our caribou assembly and the bovine reference genome, and tells us on which bovine chromosomes the caribou scaffolds are syntetic to (see Supplementary File S1 for a list of which bovine chromosome the caribou scaffolds align to, and the caribou genome website for BAM and BED alignment files for our alignment to the bovine genome). The Jupiter alignment of the reindeer to the bovine genome also showed our assembly to be more contiguous in comparison (Supplementary File S2).

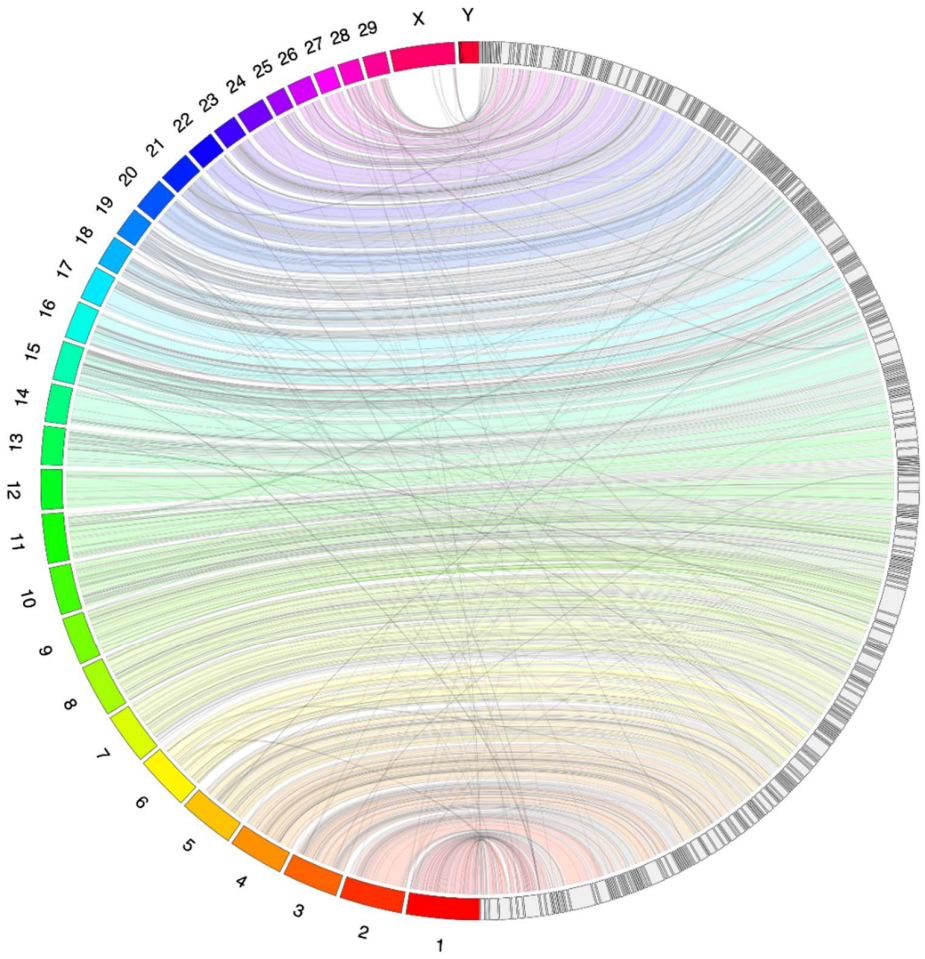

Figure 1. A Jupiter plot showing an alignment between the bovine chromosomes and the caribou genome assembly. The left of the circle shows the numbered bovine chromosomes, and the right of the circle has the largest 312 scaffolds from our assembly, which cover 75% of the bovine genome. Coloured bands represent synteny between the genomes, and lines crossing the circle indicate genomic rearrangements, or break points in the scaffolds.

Information about the genome is available and continuously updated at www.caribougenome.ca. The website includes a BLAST function, as well as the ability to download the genome in FASTA

format, the annotation (gff3 and bed) files, and a RepeatMasked version of the genome. The availability of a high-quality genome will be invaluable for answering evolutionary questions relating to this wide ranging and variable species, but also for conservation efforts. For example, a larger number of molecular markers can be developed using the whole genome data, as well as investigation into variation of potentially functional importance [27].

Supplementary Materials: The following is available online at http://www.mdpi.com/2073-4425/10/7/540/s1, Supplementary File S1: A list of which bovine chromosome the caribou scaffolds align to, Supplementary File S2: The Jupiter plot showing the alignment of the reindeer genome to the bovine genome.

Author Contributions: R.S.T. carried out analyses and wrote the manuscript. R.L.H. and X.Z. performed analyses and edited the manuscript; G.B.G., M.M. and P.J.W. conceived the project, acquired funding, and edited the manuscript.

Funding: Funding was provided through an NSERC Collaborative Research & Development (CRD) grant, NSERC grant RGPIN-2015-04477, Manitoba Hydro, Saskatchewan Power, and Weyerhaeuser Inc.

Acknowledgments: We are thankful to Jill Lalor and Bridget Redquest for extracting DNA, and Dovetail Genomics for their work sequencing and assembling the genome, to the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute Canada/ Calcul Canada for high-performance computing services. We are also thankful to Sonesinh Keobouasone for his help with data analysis, bioinformatics, and the construction of the website. The tissue sample was provided by Dale Cross and Kent Whaley, Manitoba Conservation.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Banfield, A.W.F. *A Revision of the Reindeer and Caribou, Genus Rangifer*; National Museum of Canada, Bulletin No. 177, Queen's Printer: Ottawa, ON, Canada, 1967.
2. Vors, L.S.; Boyce, M.S. Global declines of caribou and reindeer. *Glob. Change Biol.* **2009**, *15*, 2626–2633. [CrossRef]
3. Festa-Bianchet, M.; Ray, J.C.; Boutin, S.; Côté, S.D.; Gunn, A. Conservation of caribou (*Rangifer tarandus*) in Canada: An uncertain future. *Can. J. Zool.* **2011**, *89*, 419–434. [CrossRef]
4. COSEWIC. *Designatable Units for Caribou (Rangifer tarandus) in Canada*; Committee on the Status of Endangered Wildlife in Canada: Ottawa, ON, Canada, 2011.
5. Yannic, G.; Pellissier, L.; Ortego, J.; Lecomte, N.; Couturier, S.; Cuyler, C.; Dussault, C.; Hundertmark, K.J.; Irvine, R.J.; Jenkins, D.A.; et al. Genetic diversity in caribou linked to past and future climate change. *Nat. Clim. Chang.* **2014**, *4*, 132–137. [CrossRef]
6. Klütsch, C.F.C.; Manseau, M.; Anderson, M.; Sinkins, P.; Wilson, P.J. Evolutionary reconstruction supports the presence of a Pleistocene Arctic refugium for a large mammal species. *J. Biogeogr.* **2017**, *44*, 2729–2739. [CrossRef]
7. Maniatis, T.; Fritch, E.F.; Sambrook, J. *Molecular Cloning: A Laboratory Manual*; Cold Spring Harbor Laboratory: New York, NY, USA, 1982.
8. Putnam, N.H.; O'Connell, B.L.; Stites, J.C.; Rice, B.J.; Blanchette, M.; Calef, R.; Troll, C.J.; Fields, A.; Hartley, P.D.; Sugnet, C.W.; et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **2016**, *26*, 342–350. [CrossRef] [PubMed]
9. Chapman, J.A.; Ho, I.; Sunkara, S.; Luo, S.; Schroth, G.P.; Rokhsar, D.S. Meraculous: De novo genome assembly with short paired-end reads. *PLoS ONE* **2011**, *6*, e23501. [CrossRef] [PubMed]
10. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
11. Stanke, M.; Diekhans, M.; Baertsch, R.; Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **2008**, *24*, 637–644. [CrossRef]
12. Smit, A.F.A.; Hubley, R.; Green, P. Repeat Masker. Available online: http://repeatmasker.org (accessed on 30 June 2018).
13. Bushnell, B.; Rood, J.; Singer, E. BBMerge—Accurate paired shotgun read merging via overlap. *PLoS ONE* **2017**, *12*, e0185056. [CrossRef]

14. Waterhouse, R.M.; Seppey, M.; Simao, F.A.; Manni, M.; Ioannidis, P.; Klioutchnikov, G.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **2018**, *35*, 543–554. [CrossRef]
15. Lorenzini, R.; Garofalo, L. Insights into the evolutionary history of *Cervus* (Cervidae, tribe Cervini) based on Bayesian analysis of mitochondrial marker sequences, with first indications for a new species. *J. Zoolog. Syst. Evol. Res.* **2015**, *53*, 340–349. [CrossRef]
16. Elsik, C.G.; Unni, D.R.; Diesh, C.M.; Tayal, A.; Emery, M.L.; Nguyen, H.N.; Hagen, D.E. Bovine Genome Database: New tools for gleaning function from the *Bos taurus* genome. *Nucleic Acids Res.* **2016**, *44*, D834–D839. [CrossRef] [PubMed]
17. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]
18. Chu, J. Jupiter Plot: A Circos-Based Tool to Visualize Genome Assembly Consistency (Version 1.0). Zenodo. Available online: https://zenodo.org/record/1241235#.XA92q2hKiUk (accessed on 15 March 2019).
19. Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; Jones, S.J.; Marra, M.A. Circos: An information aesthetic for comparative genomics. *Genome Res.* **2009**, *19*, 1639–1645. [CrossRef] [PubMed]
20. Li, Z.; Lin, Z.; Ba, H.; Chen, L.; Yang, Y.; Wang, K.; Qiu, Q.; Wang, W.; Li, G. Draft genome of the reindeer (*Rangifer tarandus*). *GigaScience* **2017**, *6*, 1–5. [CrossRef]
21. Mikheenko, A.; Prjibelski, A.; Saveliev, V.; Antipov, D.; Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **2018**, *34*, i142–i150. [CrossRef] [PubMed]
22. Taylor, G.A.; Kirk, K.; Coombe, L.; Jackman, S.D.; Chu, J.; Tse, K.; Cheng, D.; Chuah, E.; Pandoh, P.; Carlson, R.; et al. The genome of the North American brown bear or grizzly: *Ursos arctos* ssp. *horribilis*. *Genes* **2018**, *9*, 598. [CrossRef] [PubMed]
23. Jones, S.J.M.; Taylor, G.A.; Chan, S.; Warren, R.L.; Austin Hammond, S.; Bilobram, S.; Mordecai, G.; Suttle, C.A.; Miller, K.M.; Schulze, A.; et al. The genome of the Beluga whale (*Delphinapterus leucas*). *Genes* **2017**, *8*, 378. [CrossRef] [PubMed]
24. Jones, S.J.; Haulena, M.; Taylor, G.A.; Chan, S.; Bilbobram, S.; Warren, R.L.; Austin Hammond, S.; Mungall, K.L.; Choo, C.; Kirk, H.; et al. The genome of the Northern Sea Otter (*Enhydra lutris kenyoni*). *Genes* **2017**, *8*, 379. [CrossRef] [PubMed]
25. Kharzinova, V.R.; Dotsev, A.V.; Deniskova, T.E.; Solovieva, A.D.; Fedorov, V.I.; Layshev, K.A.; Romanenko, T.M.; Okhlopkov, I.M.; Wimmers, K.; Reyer, H.; et al. Genetic diversity and population structure of domestic and wild reindeer (*Rangifer tarandus* L. 1758): A novel approach using BovineHD BeadChip. *PLoS ONE* **2018**, *13*, e0207944. [CrossRef] [PubMed]
26. Li, S.; Ajimura, M.; Chen, Z.; Liu, J.; Chen, E.; Guo, H.; Tadapatri, V.; Reddy, C.G.; Zhang, J.; Kishino, H.; et al. A new approach for comprehensively describing heterogametic sex chromosomes. *DNA Res.* **2018**, *25*, 375–382. [CrossRef] [PubMed]
27. Primmer, C.R. From conservation genetics to conservation genomics. *Ann. N. Y. Acad. Sci.* **2009**, *1162*, 357–368. [CrossRef] [PubMed]