A pseudomolecule assembly of the Rocky Mountain elk genome reveals putative immune system gene loss near chromosomal fissions

Rick E Masonbrink [1*], David Alt,[2] Darrel O. Bayles, [2] Paola Boggiatto, [2] William Edwards [3], Fred Tatum, [4] Jeffrey Williams, [2] Jenny Wilson-Welder, [2] Aleksey Zimin[5], Andrew Severin [1], Steven Olsen[2]

[1] Genome Informatics Facility, Department of Biotech, Iowa State University, Ames, IA, United States of America

[2] Infectious Bacterial Diseases Research Unit, National Animal Disease Center, Agricultural Research Service, U.S. Department of Agriculture, Ames, IA, United States of America

[3] Wildlife Health Laboratory, Wyoming Game and Fish Department, 1174 Snowy Range Road, Laramie WY, United States of America

[4] Respiratory Diseases Research Unit, National Animal Disease Center, Agricultural Research Service, U.S. Department of Agriculture, Ames, IA, United States of America

[5] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, United States of America

* Corresponding author
Email: remkv6@iastate.edu

Abstract

Rocky Mountain elk (*Cervus canadensis*) is a major reservoir for *Brucella abortus* in the Greater Yellowstone area, which has significant economic implications to the cattle industry. Vaccination attempts against intracellular bacterial diseases in elk populations have not been successful due to a negligible adaptive cellular immune response. A lack of genomic resources has impeded attempts to better understand why vaccination does not induce protective immunity. To overcome this limitation, PacBio, Illumina, and HiC sequencing with a total of 686-fold coverage was used to assemble the elk genome into 35 pseudomolecules. A robust gene annotation was generated resulting in 18,013 gene models and 33,422 mRNAs. The accuracy of the assembly was assessed using synteny to the red deer and cattle genomes identifying several chromosomal rearrangements, fusions and fissions. Because this genome assembly and annotation provide a foundation for genome-enabled exploration of Cervus species, we demonstrate its utility by exploring the conservation of immune system-related genes. We conclude by comparing cattle immune system-related genes to the elk genome, revealing nine putative gene losses in elk.

Author Summary

Brucellosis, also known as contagious abortion, is a bacterial disease that commonly affects livestock and remains prevalent in Rocky Mountain elk (*Cervus canadensis*). Since the 1920's the USDA has led a program to eradicate Brucellosis from cattle, yet wild Rocky Mountain elk continue to be a source of transmission. Attempts to vaccinate wild elk herds have been unsuccessful, due to a poor and short-lived immune response. To investigate the genetic basis for this inherent difference, we created the first genome and annotation for the Rocky Mountain elk. This genome assembly is of the highest quality and contains single linear sequences for all 35 chromosomes. In order to generate gene models, an array of RNA-Seq data and proteins from many different organ tissues and cells were used in gene prediction software. Specifically, we compare cattle immune system genes with the Rocky Mountain elk, revealing the putative loss of nine immune-system related genes in elk.

Introduction

Rocky Mountain elk (*Cervus canadensis*) were once distributed across much of North America but now inhabit remote areas. Rocky Mountain elk were nearly exterminated from the Rocky Mountains of Alberta and British Columbia in the early 1900s,(1) but were restocked between 1916-1920 with elk from the Greater Yellowstone Area (2-5). By 1940 elk populations expanded so greatly, that periodic culling was necessary (3, 6). While elk have been reintroduced to many areas, the densest populations are maintained in mountainous remote areas, like the Greater Yellowstone Area.

Elk typically avoid the presence of domesticated livestock, yet they will utilize the same grounds for grazing when livestock are absent (7). This can be problematic for ranchers occupying areas near elk populations like the Greater Yellowstone Area. Elk are known reservoirs for brucellosis, (*Brucella abortus*) a disease that is highly contagious and poses a risk to livestock and humans (8-10). Because of the potential for causing abortion in cattle, the USDA used vaccines and serologic testing to nearly eradicate *B. abortus* from domestic herds (11). Yet in the last 15 years, over 20 cases of transmission to cattle have been traced to wild elk populations in the Greater Yellowstone Area. Attempts to establish long-term immunity through vaccination have proven unfruitful, as elk have negligible adaptive cellular immune responses to existing Brucella vaccines (12). Because the eradication of *B. abortus* from cattle herds can cost hundreds of thousands of dollars and current tools make it unfeasible to control infection in wild elk, there is a need to dissect the genetic nature of limited immune responses in elk. With advances in sequencing technology (PacBio, Illumina and HiC), we are now able to investigate difference in adaptive immune response at the genomic level by examining the presence and absence of immune system-related genes. Here, we report a chromosomal level reference genome assembly and annotation of the Rocky Mountain elk and perform a preliminary investigation of immune gene loss between elk and cattle. Our results suggest a mechanism for gene loss of immune related genes through major chromosomal rearrangement and fusion.

Methods

Animal Selection

A male Rocky Mountain elk from a long term captive herd in Minnesota was selected for sequencing. The research protocol was approved by the National Animal Disease Center Animal Care and Use committee and all animals under the protocol were maintained in accordance with animal care regulations.

Sequencing

For the initial contig assembly we generated a hybrid data set with Illumina PCR-free 150bp paired end reads and PacBio RSII reads produced with P6-C4 chemistry. Chicago and HiC libraries were prepared as described previously(13, 14). Both Chicago and HiC libraries were prepared similarly, though HiC libraries were nuclear-fixed. Briefly, formaldehyde-fixed chromatin was digested with *Dpn*II, and 5' overhangs were sealed with biotinylated nucleotides. Blunt ends were ligated, followed by crosslink were reversed for DNA purification from protein. We then removed biotin that was not internal to ligated fragments. DNA was sheared to a mean length of ~350 bp for library construction with NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of the libraries. Both Chicago and HiC libraries were sequenced on an Illumina HiSeqX at 2x150bp, attaining totals of 470 million and 500 million reads, respectively.

Genome Assembly

An initial genome assembly was generated with Masurca version 3.2.3 (15), attaining a 2,559.8 Mbp genome size in 29,125 contigs with N50 size of 1,224,689bp. Dovetail Genomics scaffolded this assembly using an iterative HiRise analysis informed via alignments of Chicago and then HiC libraries with a modified SNAP aligner (http://snap.cs.berkeley.edu). This assembly contained 2,560.5 Mb, with an L90 of 31 scaffolds, and a N90 of 43.374 Mb. 1,004,453,472 Chicago and HI-C reads were used to scaffold this Dovetail assembly with a Juicer 1.5.6, 3D-DNA 180922, and JuiceBox 1.9.8 (16, 17). Reads were extracted from bam files with Picard 2.9.2 (18). The Dovetail assembly was masked using RepeatModeler 4.0.7 (19) and RepeatMasker 1.0.8 (20), prior to the alignment of HI-C reads with BWA mem 0.7.17(21). Alignments were processed using Juicer, 3D-DNA(22), and Juicebox (16, 17). The Juicebox assembly strategy consisted of: manually placing all contigs greater than 10kb,

incorporating scaffolds at the highest HI-C signal, placing scaffolds in non-repetitive regions when HI-C signal was equal between a repetitive and non-repetitive region, repeats were clustered whenever possible, and only obvious mis-joins were edited. The initial Juicebox scaffolding created 34 pseudomolecules, which was then compared to the *Cervus elaphus hippelaphus* genome assembly (GCA_002197005.1)(23)to reveal the merger of the X and Y chromosomes. A BLASTn (24) of the *C. elaphus hippelaphus* genome sequence was used to identify coordinates, allowing the correct separation the X and Y chromosome via the heatmap in Juicebox. The 3D-DNA assembly finished with 22,557 scaffolds.

The contigs that could not be integrated into the pseudomolecules were eliminated based on repetitiveness, duplicated heterozygous contigs, RNA-seq mapping potential, and contig size (>500 bp). BEDTools 2.25.0 (25) was used to merge coordinates from mapping these contigs to the pseudomolecules with BLAST+ 2.9 (score >300) and RepeatMasker 1.0.8 (20) masking coordinates. 22,065 contigs were eliminated that were less than 1kb, had at least 90% query coverage, and lacked a single unique mapping RNA-seq read, leaving 35 pseudomolecules, 457 contigs, and a mitochondrial genome.

The assembly was polished with Pilon 1.23 (26) using CCS PacBio reads and paired end Illumina DNA-seq. CCS PacBio reads were created from the PacBio subreads using bax2bam (27)and Bamtools 2.5.1 (28) and then aligned using Minimap 2.6 (29). Paired end reads were aligned using Hisat2 2.0.5 (30), followed by bam conversion and sorting with Samtools 1.9 (31). Due to uneven and excessive coverage in repetitive regions, paired end alignments were set at a max coverage of 30x using jvarkit (32). Due to the excessive repetitiveness of Chromosome_14, 50Mbp of this chromosome was not polished.

After polishing, another round of small contig elimination was performed by merging RepeatMasker (20) coordinates and coordinates from BLAST+ 2.9 (24) (score >300, width 1000bp) to the pseudomolecules with Bedtools 2.25.0 (25). If 90% of query length was repetitive and contained within the pseudomolecules, it was eliminated. BlobTools 1.11 (33) was run with PacBio subread alignments to the genome, and contigs annotated with BLAST (24) to the NT database(Supplemental Figure 1). All scaffolds passed contamination screening, resulting in a final assembly containing 35 pseudomolecules, 151 contigs, and the mitochondria.

Mitochondrial identification and annotation

BLAST+ 2.9 (24) was used to identify the mitochondrial genome by querying the mitochondrial scaffold of the *C. elaphus hippelaphus* GCA_002197005.1(23). Though the mitochondrial genome was identified, it contained three juxtaposed mitochondrial genome duplications. The scaffold was manually corrected with Samtools 1.9 (31). Genes were annotated in the mitochondrial genome using the Mitos2 webserver (34) with RefSeq 89 Metazoa, a genetic code of 2, and default settings.

Repeat prediction

A final version of predicted repeats was obtained using EDTA 1.7.9 (35) and RepeatModeler 1.0.8 (19) with RepeatMasker 4.1.0(20).

Gene prediction

A total of 753,228,475 RNA-seq reads aligned to the genome using Hisat2 2.0.5 (30) followed by bam conversion and sorting with Samtools 1.9 (31). RNA-seq read counts were obtained using Subread 1.5.2 (36). The alignments were assembled into genome-guided transcriptomes using Trinity 2.8.4 (37-39), Strawberry 1.1.1 (40), Stringtie 1.3.3b (41, 42), and Class2 2.1.7(43). The RNA-seq alignments were also used for a gene prediction via Braker2 2.1.4 (44)with Augustus 3.3.3 (45) on a genome soft-masked by RepeatMasker 1.0.8 (20)with a RepeatModeler 4.0.7 (19) library. High confidence exon splicing junctions were identified using Portcullis 1.1.2 (46). Each of these assemblies were then supplied to Mikado 2.0rc6 (47) to pick consensus transcripts, while utilizing Cervus-specific proteins from Uniprot (48) (downloaded 12-28-19). This mikado prediction was filtered for transposable elements using Bedtools 2.25.0 intersect (25) and filtered for pseudogenes via removing genes with five or fewer mapping RNA-seq reads. Using Bedtools 2.25.0 (25) intersect these filtered Mikado gene models were used to find corresponding Braker2 2.1.4 (44) gene models. Both of these predictions, together with a Genomethreader 1.7.1 (49) alignment of Uniprot proteins from the Pecora infraorder (downloaded 02-07-20) were used for a final round of Mikado gene prediction. The predicted transcripts and proteins were generated using Cufflinks (50) gffread (2.2.1), and subjected to functional annotation to: Interproscan 5.27-66.0 (51, 52) and BLAST (24) searches to NCBI

NT and NR databases downloaded on 10-23-19, as well as Swissprot/Uniprot databases downloaded on 12/09/2019.

BUSCO

Universal single copy orthologs were assessed using BUSCO 4.0 (53, 54), with the eukaryota_odb10 and cetartiodactyla_odb10 datasets in both genome and protein mode.

Synteny

With the predicted proteins from *B. taurus* (GCF_002263795.1_ARS-UCD1.2) (55), *C. elaphus* (GCA_002197005.1) (23) and *C. canadensis* genome assemblies, we inferred gene orthology using BLASTp (24) at cutoffs of an e-value of 1e-5, 50% query cover, and 70% identity. Gene-based synteny was predicted using iAdHoRe 3.0.01 [81] with prob_cutoff=0.001, level 2 multiplicons only, gap_size=5, cluster_gap=15, q_value=0.01, and a minimum of 3 anchor points. Synteny figures were produced using Circos (0.69.2) [82]. Dot plots were produced using MCScanX 20170403 (56).

Identification of Immune System-related Genes

Immune system-related genes from *Bos taurus* were found in the GENE-DB database of the International ImMunoGeneTics website (www.imgt.org) (57). This database is comprised of immunoglobulins (IG), T cell receptors (TR) and major histocompatibility (MH) genes from vertebrate species. A tblastn (2.9.0+) was performed against the elk and cattle genome assembiles, with an e-value cutoff of 1e-3.

Results and Discussion

Here we presented the first pseudomolecule assembly of the *C. canadensis*, generated with 1.7 trillion base pairs of sequencing at a 686-fold coverage of the genome.

Genome Assembly

An initial assembly was created with MaSuRCA (15, 58) generating 23,302 contigs, an L90 of 2,500 contigs, and an N90 of 197,963bp. Through collaboration with Dovetail Genomics and then additional implementation of the Juicer/JuiceBox/3D-DNA pipeline(16, 17, 22), we generated an assembly of 33 autosomes, an X chromosome, a Y chromosome, a mitochondrial genome, and 151 unincorporated contigs. We utilized synteny to identify homologous chromosomes between elk and red deer, and found that nearly always, elk chromosome sizes fell within the estimated size of the red deer's assembled chromosomes (23) (Supplemental Table 1). The only exception is the Y chromosome, which was nearly twice (7.6 Mb) the largest predicted size (4 Mb) of the red deer chromosome. We investigated all putative contaminant contigs from Blobtools (33), and ruled out contamination (Supplemental Figure 1), but also took additional steps to ensure the completeness of the genome by mapping reads back to the assembly. We found that we captured the majority of genome, with 90.7% and 87.3% of PacBio CCS reads Illumina DNA-seq aligning to the genome (Supplemental Table 2). To evaluate the completeness of the genome we ran BUSCO 4.0.2 (54) (Benchmarking Universal Single Copy Orthologs) on genome. Of the possible 255 and 13335 genes in the eukaryota and certartiodactyla odb10 datasets, 62% and 88.1% were complete, 2.4% and 2.1% were duplicated, and 3.1% and 2.1% were fragmented, and 32.5% and 9.8% were missing, respectively.

Genome Annotation

To obtain a high-quality elk gene prediction, we pursued an extensive annotation of repeats in the genome using two repeat predictors. While EDTA (35) utilizes a comprehensive set of repeat prediction programs to create a repeat annotation, Repeatmodeler/Repeatmasker (19, 20) is a long-standing and comparable annotator of repeats that is more reliant on copy number. With EDTA, 25.8% of the genome was marked repetitive, with DNA transposons comprised the largest percentage of repeats in the genome, at 16% (Supplemental Table 3). In contrast, RepeatMasker assessed 36.5% of the genome as an interspersed repeat, with 28.8% of the genome being comprised LINE retrotransposons. We merged these repeat annotations with BEDTools (25) to reveal that 38% of the genome is repetitive. This is in contrast to the repetitive content in red deer, estimated at 22.7%. This difference could be due to technological improvements and could stem from the large proportion of gaps in the

198 red deer genome (1.5Gbp) (23). While together these differences could account for a large disparity in
199 chromosome sizes, only the elk Y chromosome was outside the gapped and sequence length range in red deer
200 chromosomes (23).

201 To annotate the genes in the genome we generated 1.5 billion paired end reads of sequencing from six tissues,
202 including kidney, lung, mesenteric lymph node, muscle, prescapular lymph node, and spleen. After masking repeat
203 sequences using Repeatmodeler (19) and Repeatmaker (20), we performed five de novo transcript/gene
204 predictions with a soft-masked genome and RNA-seq. The best transcripts were discerned using Mikado(47),
205 followed by clustering with Cufflinks (50) using *B. taurus* mRNAs to cluster transcripts into gene loci. Using this
206 approach 18,013 genes were predicted to encode 33,433 mRNAs (Supplemental Table 4). The functional
207 annotations of these genes were extremely high, with 17,938 of the 18,013 genes or 99.6% being annotated by at
208 least one of: Interproscan or BLAST to NR, NT, and Uniprot (Supplemental Table 5). The gene annotation was
209 evaluated for completeness with BUSCO in protein mode. A remarkable "Complete" score improvement is seen in
210 both eukaryota and cetartiodactyla at 97.7% and 92.1%, respectively. These results together suggest that both the
211 genome and the gene prediction are of high quality.

212 Comparison to Related Species

213 By utilizing these new gene predictions we evaluated the conservation of chromosome structure between *C.*
214 *canadensis*, *C. elaphus hippelaphus*, and *B. taurus* using gene-based synteny with i-ADHoRe (59). All elk
215 chromosomes were syntenic with all *C. elaphus* and *B. taurus* chromosomes, though chromosome Y lacked the
216 genes required for gene-based synteny (Figure 1, Table 1). As has been seen in previous Cervus assemblies (23),
217 multiple pairs of chromosomes are tandemly fused in *B. taurus* and vise-versa (Table 2). We confirmed previous
218 reports of chromosome fusions and fissions indicated that twelve cervus chromosomes fused into six in *B. taurus*,
219 as well as four chromosomes in *B. taurus* are fused into two cervus chromosomes (Table 2).

220 Two inter-chromosomal translocations were inferred between the two Cervus species, both having strong HiC
221 support in Elk (Figure 1, Table 3). Chromosome_15 and Chromosome_24 of elk, comprised large portions of
222 Ce_Chr_33 and a minor portion of *C. elaphus* Ce_Chr_8. With the majority of Chromosome_24 homologous to *C.*

elaphus hippelaphus Ce_Chr_8, a 17 MB region of Ce_Chr_33 may have been falsely attached to Ce_Chr_8 in *C. elaphus hippelaphus*. Another smaller chromosome translocation of 13.6 MB occurred between Ce_Chr_22 and Ce_Chr_3 of C. elaphus, attributed to chromosomes 21 and 25 in *C. canadensis*. A small region of Ce_Chr_22 was likely falsely attached to Ce_Chr_3 in *C. elaphus hippelaphus*. Interestingly, both of these translocations are between chromosomes in elk that are fused chromosomes in *B. taurus*, Bt_Chr_2 and Bt_Chr_5 (table). While it is possible that these translocations occurred since the divergence of these two species, because the *B. taurus* assembly was used to orient and join scaffolds in the *C. elaphus hippelaphus* genome assembly, it is likely that these translocations are misassemblies in the *C. elaphus hippelaphus* genome.

Immune Gene Loss

Nine *Bos taurus* immune genes were identified from the IMGT GENE-DB database (57) that did not align to the Elk genome: AY644518_TRGJ1, KT723008_IGHJ2, AC172685_IGHA, IMGT000049_TRDC, D16120_TRGJ2, AY149283_IGHJ1, AY2277782_TRAJ31, AY644517_TRG and NW_001494075_IGHJ1. These genes are all components of the T cell receptor: (gamma joining 1), (gamma joining 2), (alpha joining 31), and (delta constant) or of the Immunoglobulin complex: (heavy joining 1), (heavy joining 2) and (heavy constant alpha). Interestingly, seven of the nine genes are located at the very end of chromosomes 01 and 02 in *Bos taurus* (Table 4). *B. taurus* chromosomes 01 and 02 have split in *C. canadensis* into chromosomes 7 and 27, and chromosomes 24 and 15, respectively (Table 2) suggesting a possible mechanism for loss of these genes through large chromosomal rearrangements and fission. Future work will be required to investigate how the loss of these genes affects cellular immune response in elk, yet this may provide the foundation necessary to develop long-term immunity to Brucella.

Conclusions

This genome assembly and annotation of the Rocky Mountain elk is the most contiguous assembly of a Cervus species and will serve as an important tool for genomic exploration of all related Cervids. Elk's loss of immune system-related genes in relation to cattle, may provide a clue to establishing a successful vaccination strategy. This

chromosomal assembly of the elk genome will provide an excellent resource for investigating genes involved in elk's poor adaptive cellular immune response to Brucella vaccines.

Acknowledgements

The authors would like to thank Maryam Sayadi for fruitful discussion regarding the genome assembly paper, and Mary Wood regarding elk sample collection.

Author contributions

Conceptualization –SO; data curation – REM, AS; formal analysis – REM, AZ, AS, funding acquisition – SO; investigations – DA, DOB, PB, WE, FT, JWM, JWW; methodology – REM, AZ, AS; resources -- AS, software – REM, AZ, AS; supervision -- SO, validation – REM, AS; visualization -- REM, writing -- REM, AS; review and editing – REM, DA,WE, AS, SO

Data Availability: The Rocky Mountain elk genome has been deposited at GenBank accession and associated sequencing reads to the NCBI SRA database under BioProject PRJNA657053. All programs and scripts are available at https://github.com/ISUgenomics/elk_genomics.

Funding: This work was supported by the USDA National Institute of Food and Agriculture under grant 2018-67015-28199 to AZ.

Competing interests: The authors have declared that no competing interests exist.

1. Stelfox J. Elk in north-west Alberta. Land-Forest-Wildlife. 1964;6(5):14-23.
2. Pybus MJ, Butterworth EW, Woods JG. An expanding population of the giant liver fluke (Fascioloides magna) in elk (Cervus canadensis) and other ungulates in Canada. Journal of Wildlife Diseases. 2015;51(2):431-45.
3. Green H. The elk of Banff National Park. Unpubl. 1946:32.
4. Lloyd H. Transfers f elk for re-stocking. *Can Field Nat*. 1927;41:126-7.
5. Lothian W. A history of Canada's National Parks. 1981;4:155.

6. Flook DR. A Study of the Apparent Unequal Sex Ration of Wapiti: University of Alberta (Ph. D.); 1967.
7. Stewart KM, Bowyer RT, Kie JG, Cimon NJ, Johnson BK. Temporospatial Distributions of Elk, Mule Deer, and Cattle: Resource Partitioning and Competitive Displacement. Journal of Mammalogy. 2002;83(1):229-44.
8. Cotterill GG, Cross PC, Merkle JA, Rogerson JD, Scurlock BM, Du Toit JT. Parsing the effects of demography, climate and management on recurrent brucellosis outbreaks in elk. Journal of Applied Ecology. 2020;57(2):379-89.
9. Godfroid J. Brucellosis in wildlife. Revue Scientifique et Technique-Office international des épizooties. 2002;21(1):277-86.
10. Lowry J, Goodridge L, Vernati G, Fluegel A, Edwards W, Andrews G. Identification of Brucella abortus genes in elk (Cervus elaphus) using in vivo-induced antigen technology (IVIAT) reveals novel markers of infection. Veterinary microbiology. 2010;142(3-4):367-72.
11. Yingst S, Hoover D. T cell immunity to brucellosis. Critical reviews in microbiology. 2003;29(4):313-31.
12. Nol P, Olsen SC, Rhyan JC, Sriranganathan N, McCollum MP, Hennager SG, et al. Vaccination of elk (Cervus canadensis) with Brucella abortus strain RB51 overexpressing superoxide dismutase and glycosyltransferase genes does not induce adequate protection against experimental Brucella abortus challenge. Frontiers in cellular and infection microbiology. 2016;6:10.
13. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome research. 2016;26(3):342-50.
14. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. science. 2009;326(5950):289-93.
15. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29(21):2669-77.
16. Dudchenko O, Shamim MS, Batra SS, Durand NC, Musial NT, Mostofa R, et al. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. Biorxiv. 2018:254797.
17. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell systems. 2016;3(1):99-101.
18. Intitute B. Picard Tools. 2019.
19. Smit A, Hubley R, Green P. RepeatModeler Open-1.0. 2008-2010. Access date Dec. 2014.
20. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013–2015. Institute for Systems Biology http://repeatmasker org. 2015.
21. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
22. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356(6333):92-5.
23. Bana NÁ, Nyiri A, Nagy J, Frank K, Nagy T, Stéger V, et al. The red deer Cervus elaphus genome CerEla1. 0: sequencing, annotating, genes, and chromosomes. Molecular Genetics and Genomics. 2018;293(3):665-84.
24. Madden T. The BLAST sequence analysis tool. The NCBI Handbook [Internet] 2nd edition: National Center for Biotechnology Information (US); 2013.
25. Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. Current protocols in bioinformatics. 2014:11.2. 1-.2. 34.

26. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS one. 2014;9(11):e112963.
27. Biosciences P. SMRT Link. 2017.
28. Barnett D, Garrison E, Marth G, Stromberg M. BamTools. 2013.
29. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016;32(14):2103-10.
30. Kim D, Langmead B, Salzberg S. HISAT2: graph-based alignment of next-generation sequencing reads to a population of genomes. 2017.
31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
32. Lindenbaum P. JVarkit: java-based utilities for Bioinformatics. 2015. Preprint Available: figshare. 2018.
33. Laetsch DR, Blaxter ML. BlobTools: Interrogation of genome assemblies. F1000Research. 2017;6(1287):1287.
34. Donath A, Jühling F, Al-Arab M, Bernhart SH, Reinhardt F, Stadler PF, et al. Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. Nucleic acids research. 2019;47(20):10543-52.
35. Ou S, Su W, Liao Y, Chougule K, Agda JR, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome biology. 2019;20(1):1-18.
36. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic acids research. 2013;41(10):e108-e.
37. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. 2011;29(7):644-52.
38. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols. 2013;8(8):1494-512.
39. Henschel R, Lieber M, Wu L-S, Nista PM, Haas BJ, LeDuc RD, editors. Trinity RNA-Seq assembler performance optimization. Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond; 2012.
40. Liu R, Dickerson J. Strawberry: Fast and accurate genome-guided transcript reconstruction and quantification from RNA-Seq. PLOS Computational Biology. 2017;13(11):e1005851.
41. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology. 2015;33(3):290.
42. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature protocols. 2016;11(9):1650.
43. Song L, Sabunciyan S, Florea L. CLASS2: accurate and efficient splice variant annotation from RNA-seq reads. Nucleic acids research. 2016;44(10):e98-e.
44. Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: incorporating protein homology information into gene prediction with GeneMark-EP and AUGUSTUS. Plant and Animal Genomes XXVI. 2018.
45. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24(5):637-44.
46. Mapleson D, Venturini L, Kaithakottil G, Swarbreck D. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. GigaScience. 2018;7(12):giy131.

47. Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. GigaScience. 2018;7(8):giy093.
48. Consortium U. UniProt: a worldwide hub of protein knowledge. Nucleic acids research. 2019;47(D1):D506-D15.
49. Gremme G. GenomeThreader Gene Prediction Software. 2014.
50. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols. 2012;7(3):562.
51. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, et al. InterPro in 2017—beyond protein family and domain annotations. Nucleic acids research. 2016;45(D1):D190-D9.
52. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236-40.
53. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015:btv351.
54. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular biology and evolution. 2017;35(3):543-8.
55. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. GigaScience. 2020;9(3).
56. Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Research. 2012;40(7):e49-e.
57. Giudicelli V, Chaume D, Lefranc M-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic acids research. 2005;33(suppl_1):D256-D61.
58. Zimin AV, Puiu D, Luo M-C, Zhu T, Koren S, Marçais G, et al. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome research. 2017;27(5):787-92.
59. Proost S, Fostier J, De Witte D, Dhoedt B, Demeester P, Van de Peer Y, et al. i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. Nucleic acids research. 2011;40(2):e11-e.

Figure 1. Synteny and HiC plot of Elk chromosomes. A. Gene-based synteny between *C. elaphus hippelaphus* and *C. canadensis*. B. HiC plot of elk chromosomes in JuiceBox. C. Gene-based synteny between *B. taurus* and *C. canadensis*.

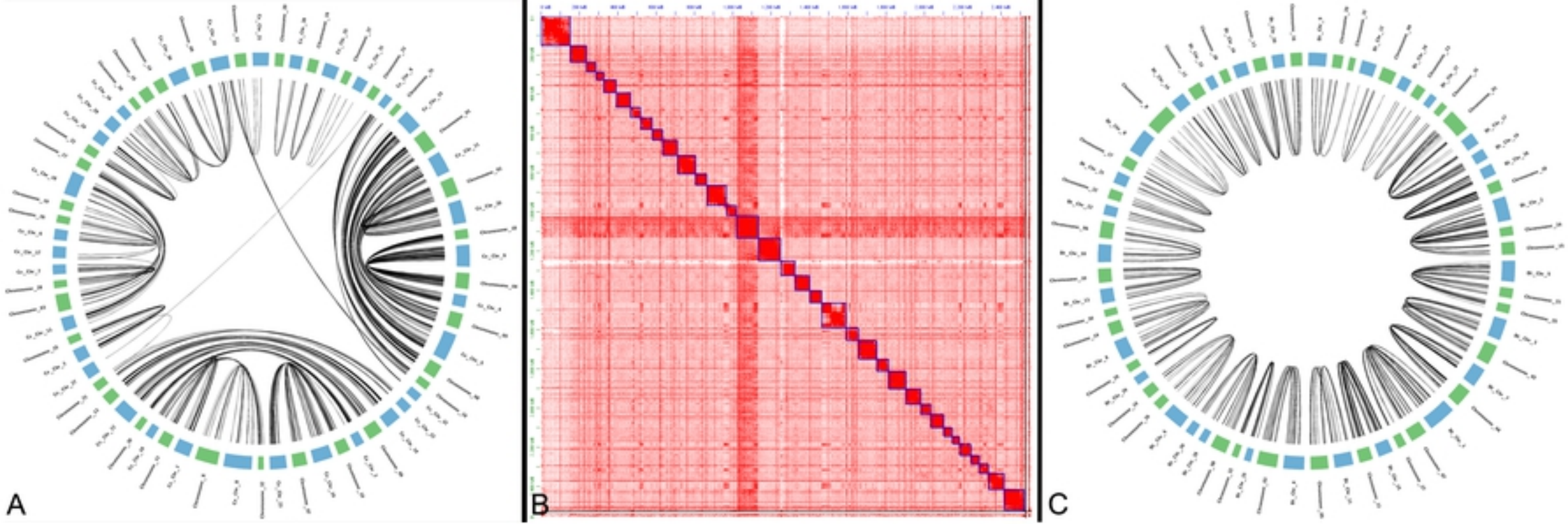

Figure 1