Journal of Heredity, 2022, **113**, 479–489
https://doi.org/10.1093/jhered/esac022
Advance access publication 3 May 2022
Genome Resources

American Genetic Association

OXFORD

Genome Resources

A De Novo Chromosome-Level Genome Assembly of the White-Tailed Deer, *Odocoileus Virginianus*

Evan W. London, Alfred L. Roca, Jan E. Novakofski and Nohra E. Mateus-Pinilla

From the Illinois Natural History Survey-Prairie Research Institute, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA (London, Roca, Novakofski, and Mateus-Pinilla); Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA (London, Roca, Novakofski, and Mateus-Pinilla); and Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA (Roca).

Address Correspondence to A.L. Roca at the above address, or e-mail: roca@illinios.edu.
Address Correspondence to N.E. Mateus-Pinilla at the above address, or e-mail: nohram@illinois.edu.

Corresponding Editor: Klaus-Peter Koepfli

Abstract

Cervids are distinguished by the shedding and regrowth of antlers. Furthermore, they provide insights into prion and other diseases. Genomic resources can facilitate studies of the genetic underpinnings of deer phenotypes, behavior, and disease resistance. Widely distributed in North America, the white-tailed deer (*Odocoileus virginianus*) has recreational, commercial, and food source value for many households. We present a genome generated using DNA from a single Illinois white-tailed sequenced on the PacBio Sequel II platform and assembled using Wtdbg2. Omni-C chromatin conformation capture sequencing was used to scaffold the genome contigs. The final assembly was 2.42 Gb, consisting of 508 scaffolds with a contig N50 of 21.7 Mb, a scaffold N50 of 52.4 Mb, and a BUSCO complete score of 93.1%. Thirty-six chromosome pseudomolecules comprised 93% of the entire sequenced genome length. A total of 20 651 predicted genes using the BRAKER pipeline were validated using InterProScan. Chromosome length assembly sequences were aligned to the genomes of related species to reveal corresponding chromosomes.

Key words: annotation, haploid, Illumina, non-model species, Omni-C, Pacific Biosciences

The white-tailed deer (*Odocoileus virginianus*) is 1 of 5 species within the deer family Cervidae that is native to the United States, along with the mule deer (*Odocoileus hemionus*), moose (*Alces americanus*), caribou (*Rangifer tarandus*), and elk (*Cervus canadensis*). White-tailed deer are the most widespread of all Capreolinae (New-world deer), with a range extending from the Arctic Circle in Canada to Peru and Bolivia (Hewitt 2011). In the United States (USA) deer hunting is a growing industry, accounting for $20 billion of value added to the GDP in 2016 (Allen et al. 2018). Additionally, there were 3172 deer farms operating in the United States with an estimated value of $50 million in meat and animal product sales as of 2017 (USDA National Agricultural Statistics Service 2019).

Reference genomes are currently available for 3 North American deer species; mule deer (Lamb et al. 2021), Rocky Mountain elk (Masonbrink et al. 2021), and white-tailed deer (*Odocoileus virginianus texanus*) (Seabury et al. 2011). Using third-generation sequencing (3GS), the Rocky Mountain elk and mule deer genomes have been resolved at the chromosome level (Lamb et al. 2021; Masonbrink et al. 2021). A chromosome-level assembly sequence is a reasonably complete pseudo-molecule with some gaps but consisting primarily of sequenced bases (Genome Reference Consortium 2021). The Rocky Mountain elk and mule deer genomes were both generated using Pacific Biosciences (PacBio), Illumina, and Hi-C sequencing with both assemblies consisting of 35 chromosome-scale scaffolds (Lamb et al. 2021, Masonbrink et al. 2021). However, 3GS was not yet available when the Seabury et al. assembly was generated for white-tailed deer (Seabury et al. 2011). The existing white-tailed deer genome consists of >17 000 small scaffolds generated using second-generation sequencing (2GS).

Third-generation sequencing, such as PacBio, allows for continuous reads of single molecules of DNA ranging in size from 1 to 50 kb (English et al. 2012). Long reads allow for greater overlaps between DNA sequences, resolution of long repeat elements, and the reconstruction of contigs (Pollard et al. 2018). A technique based on chromosome conformation capture, Hi-C (Omni-C) sequencing, is utilized to map associations between sequences originating from the same chromosome (Belton et al. 2012). Applying both 3GS (PacBio) and 2GS sequencing (Illumina, Hi-C) techniques allows for the construction of higher resolution genome assemblies because the high accuracy of 2GS short-reads corrects errors in 3GS long-read sequencing (Mahmoud et al. 2019).

Received December 13, 2021; Accepted May 5, 2022

Having a high-quality genome assembly can empower further studies and genomic resequencing projects at the population level (Fuentes-Pardo et al. 2017). The ranched/farmed white-tailed deer industry is relatively small but has been expanding within the United States, thus creating a demand for genomic resources that can be used to study heritable traits such as body size, antler rejuvenation (Jamieson et al. 2020), and resistance to pathogens (Seabury et al. 2020).

A more complete, 3GS white-tailed deer genome will facilitate future research studies into additional genes that may play a role in diseases of cervids (Masonbrink et al. 2021), including white-tailed deer. For example, all native cervid species in North America are susceptible to chronic wasting disease (CWD), a transmissible spongiform encephalopathy (Rivera et al. 2019), linked to genetic variation in the *PRNP* gene (Robinson et al. 2012; Brandt et al. 2018; Güere et al. 2020). Additionally, according to recent genome-wide association studies (GWAS), other non-*PRNP* loci may play a role in CWD (Seabury et al. 2020), as is the case for other prion diseases such as Creutzfeldt-Jakob disease (Jones et al. 2020).

Furthermore, having a chromosome-level assembly comparative genomics across species. Extrapolation of linkage is dependent on relative chromosome location, and chromosome arrangements may differ across species (Potter et al. 2017). Knowing the chromosome identities in white-tailed deer will allow for the evaluation of gene relationships within and across chromosomes (Kong et al. 1997).

Therefore, there is a need to build on the resolution of the white-tailed deer genome using 3GS and Hi-C scaffolding technologies. The primary aim of this study is to create a de novo chromosome-level deer genome by integrating resources from both 2GS and 3GS platforms as well as Hi-C sequencing for scaffolding. Additionally, chromosome comparisons will be made between white-tailed deer and other mammal species to identify homologous chromosomal regions.

Methods

Biological Sample Collection

A muscle tissue sample was selected from the Illinois tissue research archive used in previous CWD genetic studies (Brandt 2018; Rivera 2019; Ishida 2020) . Illinois white-tailed deer was traditionally classified as *Odocoileus virginianus borealis*, although the population is an admixture of deer relocated from adjacent regions (Pietsch 1954; Perrin-Stowe 2020). The criteria for choosing the sample included: being stored for fewer than 5 months, cold-weather field conditions during sample collection, and sustained storage at −20 °C. A Male was chosen to sequence both X and Y chromosomes. The selected male white-tailed deer originated from Jo Daviess County and was sampled in February 2020.

Nucleic Acid Library Preparation

Circulomics Nanobind High-Molecular-Weight DNA Extraction

High-molecular-weight DNA was extracted from 0.5 g of muscle tissue using the Nanobind Tissue Big DNA Kit (Curculomics, Baltimore MD). Briefly, tissue was disrupted using a tight-fitting 1.0-mL Dounce homogenizer before lysing with proteinase K. Following homogenization, the solution was centrifuged at 3000 × *g* at 4 ° C for 5 min to pelletize debris and proteins. The supernatant containing the DNA was transferred to a low-bind microfuge tube. Isopropanol and the magnetic Nanobind disk were then added to the supernatant and gently mixed. The disk, containing bound DNA, was washed 3 times using a magnetic tube rack to prevent DNA shearing. Finally, DNA was eluted from the disk using 75 µL of elution buffer. The resulting DNA concentration was quantified using a Qubit 4 fluorometer (ThermoFisher Scientific, Waltham, MA), and the DNA length was quantified using a Fragment Analyzer™ (Advanced Analytical Technologies, Inc.).

Third- and Second-Generation DNA Sequencing

To generate long-read sequencing libraries, DNA was sheared with a gTube to an average fragment length of 30 kb prior to conversion into a library following the SMRTbell Express Template Prep Kit 2.0 protocol from Pacific Biosciences. The library was sequenced on 2 SMRT cells on a PacBio Sequel II with 24-hr movies. Short-read shotgun genomic libraries were prepared using the Hyper Library construction kit from Kapa Biosystems (Roche, Penzberg, Germany). Libraries were sequenced on the Illumina NovaSeq 6000 equipped with an SP flowcell using 2 × 150 bp paired-end reads. Chromatin conformation capture sequencing libraries were prepared using the Omni-C kit from Dovetail Genomics. Libraries were pooled; quantitated by qPCR and sequenced on the Illumina NovaSeq 6000 equipped with an SP flowcell using 2 × 150 nt paired-end reads. Library preparation and sequencing were conducted by the Roy J. Carver Biotechnology Center of the University of Illinois at Urbana-Champaign (UIUC).

Pre-processing Reads

Sequencing reads that were much shorter or longer than the expected read length for 3GS were removed from the read data sets before genome assembly to reduce misassembles and false contigs. Specifically, PacBio long reads >5000 bp (Hufnagel et al. 2020) were retained from the data sets using Fastp (Chen et al. 2018) to improve the final assembly, and reads greater than 50 kb were removed to reduce the potential for chimeric molecular sequencing templates. All 2GS Adaptor sequences were removed using bcl2fastq (https://support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html). Thereafter, sequences were filtered for a minimum read length of 50 bp and a minimum PHRED score of 30 to ensure short read accuracy using Fastp. Omni-C reads were subsampled using Seqtk (https://github.com/lh3/seqtk) to include only 300 million read pairs based on the recommended protocol provided by Dovetail Genomics for the Omni-C kit (Dovetail Genomics, Scotts Valley, CA). We list the programs and versions used throughout the assembly and analysis pipeline in (Table 1).

De Novo Genome Assembly and Error Correction

Assembly with Wtdbg2

Filtered PacBio reads were assembled using the Wtdbg2 assembler (Ruan and Li 2020) at successive coverage threshold intervals: 50×, 70×, 90×, 100×, 110×, 135×, and 150× (Supplementary Table S1). Furthermore, analysis was conducted using the 90x threshold coverage assembly because the increase of coverage threshold from 70× to 90× produced the most substantial gains in "longest contig length" (Supplementary Table S1), while limiting excess coverage from the higher error-rate long reads. Genome assembly

Table 1. Bioinformatics software used for assembly and analysis

	Software	Version
Assembly and error correction		
Long-read filtering	Fastp	0.20.0
De novo Assembly	Wtdbg2	2.5
Contig polishing (long reads)	https://github.com/PacificBiosciences/gcpp	8.0.0
Short-read pre-processing	Bcl2fastq2	2.20
Short-read filtering	Fastp	0.20.0
Contig polishing (short reads)	Pilon	1.2.2
Contig deduplication	Purge-Haplotigs	1.1.1
Contamination screen	BLAST+	2.10.1
Scaffolding		
Omni-C™ read filtering	https://github.com/lh3/seqtk	0.3.0
Arima genomics mapping pipeline	https://github.com/lh3/bwahttp://samtools.sourceforge.net/ https://github.com/broadinstitute/picard	0.7.17 1.11 2.10.1
Omni-C™ scaffolding	SALSA2	2.2
Omni-C™ contact map	Juicebox HiCExplorer	1.11.08 2.2.1.1
Scaffold deduplication	Purge-Haplotigs	1.1.1
Benchmarking		
Genome completeness	BUSCO	4.1.4
Synteny with other species	https://github.com/JustinChu/JupiterPlot	1.0
Annotation		
Repeat assessment	RepeatMasker	4.1.1
Protein alignments	ProtHint	2.5.0
RNA alignments	STAR	2.7.6a
Gene prediction	BRAKER	2.1.6
Prediction filtering	Interproscan	5.52-86

Software presented in relative order of use in the pipeline. See citations in-text.

and polishing were conducted on the Biocluster at the Carl R. Woese Institute for Genomic Biology at UIUC.

Error Correction

Two polishing steps were performed using long and short sequencing reads to improve assembly accuracy. Wtdbg2 performed a single round of consensus polishing (Ruan and Li 2020) using the binned sequences. An additional round of long-read consensus polishing was completed by aligning PacBio reads to the 90× Wtdbg2 assembly using the Arrow algorithm from the PacBio SMRTLink software package (Pacific Biosciences). Short-read consensus polishing rounds were conducted with the filtered Illumina reads using Pilon (Walker et al. 2014) in conjunction with UniCycler (Wick et al. 2017) to execute 10 iterative rounds of Pilon polishing. Pilon polishing with a single round was also conducted to address potential overcorrection. The assembly was examined using BLAST+ for contaminant sequences that may have been introduced through the DNA extraction sequencing process. The core UniVec database (http://ftp.ncbi.nih.gov/pub/UniVec) was downloaded from NCBI on March 10, 2021, and a nucleotide BLAST search was performed against the assembly contigs. All contaminant sequences were excised. The contig from which each was excised was split at the removal sites into 2 separate contigs (NCBI 2016). Excision of contaminating bp was performed using the Emacs text editor using the start and end positions of the alignment output from the nucleotide BLAST search. Furthermore, the genome was assessed for potential contamination during final submission to the GenBank Genome database.

Genome Deduplication and Scaffolding

Removal of Haplotigs and Artifacts

The software Purge-Haplotigs (Roach et al. 2018) was used to remove haplotig and artifact assembly fragments. Artifacts were defined as contigs with greater than 80% of their sequence being above the high or below the low sequencing coverage thresholds. The threshold of 80% (default setting) was previously shown to be sufficient to purge putative artifacts and organelle contigs from the assembly (Roach et al. 2018). Contig coverage histograms were generated by aligning the filtered PacBio reads to the assembly using Minimap2 (v.) (Li 2018). The histogram (Supplementary Figure S1A) was generated using the *purge_haplotigs hist* command with the long-reads aligned back to the genome in a BAM file as input. The low coverage threshold was set to 15, and the high coverage threshold was set to 190. The midpoint coverage between the haploid and diploid peaks was set to 55. Coverage thresholds were derived from the histogram peaks in Supplementary Figure S1A and their midpoint.

Arima Genomics Mapping Pipeline

Subsampled Omni-C paired reads were aligned to the deduplicated contig assembly using bwa index and bwa mem (Li and Durbin 2009). Aligned read pairs were sorted by position using SAMtools (Li et al. 2009) and filtered for 5' ends using the *filter_five_end.pl* script (https://github.com/ArimaGenomics/mapping_pipeline). Reads were also filtered with SAMtools using a minimum mapping quality of 10. Read groups and duplicate reads were added using Picard (https://github.com/broadinstitute/picard).

Scaffolding, Contig Reassignment, and Haplotig Removal

Mapped Omni-C reads were used as input for Salsa (Ghurye et al. 2019) scaffolding. Salsa was run in correction mode, allowing the use of mapping information to detect misassemblies in the input contigs. The contacts between scaffolds were visualized using Juicebox (Robinson et al. 2018). The second round of deduplication was conducted using Illumina paired-end reads using Purge-Haplotigs (Roach et al. 2018). In short, Illumina reads were aligned to scaffolds using Minimap2 and a read-depth histogram was created using the *purge_haplotigs hist* command (Supplementary Figure S1B). Scaffolds were filtered based on a low coverage threshold of 5, and a high coverage threshold of 90. The midpoint threshold was set to 25. Coverage thresholds were derived from the histogram in Supplementary Figure S1B peaks and their midpoint.

Chromosome-Level Pseudomolecule Curation

The scaffold chromatin contact matrix was visualized with HiCExplorer (Ramirez et al. 2018) and specific scaffold–scaffold contact graphs were examined using Juicebox. Based on the contact graphs, scaffolds were joined into pseudomolecules when orientation could be determined. The orientation of the largest scaffold in each pseudomolecule was assumed to be in the forward direction. Smaller scaffolds were reversed as necessary based on the contact information. Final chromosomes were aligned to the chromosome assemblies of 6 other species using MiniMap2. The species in order of largest to smallest chromosome number were *Cervus canadensis* (GCA_019320065.1, Masonbrink et al. 2021), *Cervus nippon* (GWHANOY00000000, Xiumei et al. 2021), *Cervus elaphus* (GCA_002197005.1, Bana et al. 2018), *Bos taurus* (GCA_000003205.1, Mehta et al., 2009), *Ovis aries* (GCA_011170295.1, Li et al. 2021), and *Homo sapiens* (GCA_000001405.28, Schneider et al. 2017). The species *C. canadensis*, *C. nippon*, and *C. elaphus* have 68 autosomes; whereas *B. taurus* and *O. aries* have 58 and 52 autosomes, respectively. All species had sequences for both X and Y chromosomes except for *C. nippon*, for which the Y-chromosome sequence was not available at the time of publication. Sex chromosomes were determined based on alignment with the other species.

Genome Annotation

Genomic annotation used multiple available databases for gene prediction. Gene models were predicted using the BRAKER annotation pipeline with transcript and protein evidence via GeneMark ETP+ (Altschul et al. 1990; Lomsadze et al. 2005; Stanke et al. 2008; Camacho et al. 2009; Barnett et al. 2011; Hoff et al. 2016, 2019). RNA alignments were examined using GeneMark (Lomsadze et al. 2014). Proteins were aligned to the genome using ProtHint (Brůna et al. 2020), which combines the Splan (Gotoh 2008; Iwata and Gotoh 2012) and DIAMOND (Buchfink et al. 2015) protein aligners. Prior to annotation, the genome was masked with RepeatMasker (Smit et al. 2013) using Cetartiodactyla and ancestral repeat sequences in the RepBase Update repeat database (Bao et al. 2015). Cetartiodatyla includes cetaceans and even-toed ungulates (Price et al. 2005). Soft-masking of repeat sequences using RepeatMasker was used to increase annotation speed and accuracy (Hoff et al. 2019).

Available RNA-seq data for white-tailed deer were downloaded from the NCBI Sequence Read Archive. RNA-Seq data have been generated in previous studies from multiple tissue types including retropharyngeal lymph node (SRX4604241), liver (SRX2175788, SRX2175791), antler (SRX2175789), bone (SRX2175790), lung (SRX2175792), brain (SRX2175793), muscle (SRX2175794), testis (SRX2175795, SRX2175797), and pedicle (SRX2175796). All RNA reads were trimmed using Trim Galore (Martin et al. 2011) using the default settings to remove adapter sequences and sequences with an average Phred score below 30. Following trimming, RNA was aligned to the genome using STAR (Dobin and Gingeras 2015) and sorted into bam files using SAMtools (Li et al. 2009). All RNA dataset BAM files were then merged into a single input file for BRAKER.

Following the guidance of BRAKER pipeline D (https://github.com/Gaius-Augustus/BRAKER), protein sequences from humans (n = 20 396) and artiodactyls (n = 8 931) present in the SwissProt database (Boutet et al. 2007) were used as evidence from "closely related" species. Vertebrate protein sequences present in the orthologous gene database, OrthoDB (n = 4 937 339) (Kriventseva et al. 2019), were used as evidence from more distantly related species. All protein sequences were aligned to the genome using the ProtHint pipeline within GeneMark-EP (Brůna et al. 2020), which provides an output file that BRAKER can use to incorporate protein information. BRAKER merges the external evidence from RNA-Seq and protein alignments for use as input to the Augustus gene prediction software (Stanke et al. 2006; Keller et al. 2011), which outputs the final general feature format files containing the locations and features of predicted genes. Only genes supported by RNA and protein sequence data were used for further analysis.

The longest coding sequences for each supported gene predicted by BRAKER were translated into amino acids and queried against the InterProScan Gene3D and Pfam protein databases (Jones et al. 2014; Lewis et al. 2018; Blum et al. 2021; Mistry et al. 2021). Sequences with matches were retained within the BRAKER annotation file and predicted genes without corresponding matches were removed. Additionally, retroelements with identified reverse-transcriptase domains were removed from the protein-coding gene annotations.

Assessing Completeness and Synteny

To assess the completeness of the assembly, BUSCO (Manni et al. 2021) searches were conducted following successive steps of the assembly and analysis pipeline. All BUSCO searches were conducted using the Cetartiodactyla lineage dataset from OrthoDB (Kriventseva et al., 2019). Synteny between the white-tailed deer pseudomolecule assembly and the Rocky Mountain elk assembly (GCA_019320065) was

visualized using JupiterPlot (https://github.com/JustinChu/JupiterPlot). The software performs alignments between reference chromosomes and query scaffolds using Minimap2, runs in assembly mode drawing the alignment links in a circular diagram.

Results

Sequencing and Assembly

3GS, 2GS, and Omni-C Sequencing Metrics

Genomic sequencing used 13 µg of DNA with an average fragment length of 54.8 kb. Two single-molecule real-time sequencing cells produced 32 932 198 reads (cell 1: 13 756 097; cell 2: 19 176 101) for a total of 390.6 gigabases (Gb) of DNA sequence (cell 1: 174.6 Gb; cell 2: 215.9 Gb). Long reads used for assembly were between 5 and 50 kb in length and totaled 23 002 345 bp (cell 1: 9 554 592; cell 2: 13 447 753), covering 345.9 Gb of sequence (cell 1: 153.6 Gb; cell 2: 192.3 Gb). A single lane of paired-end Illumina sequencing produced 1 049 534 322 paired-end reads for a total of 157.4 Gb of DNA sequence. Short reads used for error correction and deduplication totaled 1 004 706 776; thus representing 149.2 Gb of the total sequence. A single lane of paired-end Illumina sequencing of the Omni-C library produced a total of 967 979 604 paired reads for a total of 145.1 Gb of DNA sequence. The 600 million paired-end reads were subsampled from the total Omni-C sequencing output for a total of 90 Gb of sequence.

Contig- and Scaffold-Level Assembly with Deduplication

The 90x coverage Wtdbg2 assembly represented a plateau in assembly quality while limiting the input of "noisy" long reads and was used for further analysis (Supplementary Table S1). Wtdbg2 produced 5 506 contigs from filtered PacBio reads. Deduplication of the contig assembly produced 984 haplotigs and 2103 artifact sequences for a total of 27.9 and 34.2 Mb, respectively. A single contig was found to have a contamination vector based on a BLAST search of the UniVec database and no contamination was found by GenBank submission staff. The final contig assembly consisted of 2420 contigs with a total length of 2 461 348 864 bp. The N50 of the contig assembly was 21.7 Mb with an L50 of 32 contigs (Table 2 and Figure 1A). Scaffolding by Salsa with Omni-C reads was able to join 312 contigs into 156 scaffolds. Additionally, Salsa detected 8 contigs with mis-assemblies based upon Omni-C mapping information. Misassembled contigs were separated into 16 sequences before being joined into scaffolds. The N50 of the scaffold assembly was 51.4 Mb, with an L50 of 18 sequences (Figure 1A). A strong diagonal "self-associated" signal was observed in the Hi-CExplorer plot of the Omni-C contact matrix (Figure 1B), with minimal non-self associations. Deduplication of the scaffold assembly with Purge-Haplotigs revealed 637 duplicated haplotype scaffolds and 972 artifact scaffolds for a total of 17.9 and 18.6 Mb, respectively. The final scaffold assembly consisted of 191 contigs joined into 508 scaffolds with a total un-gapped length of 2.42 Gb (2 424 791 208 bp). A total of 36 scaffolds were joined into 12 chromosome groups based on HiC associations and the remaining 24 chromosomes consisted of single scaffolds (Table 3 and Figure 1C). The 36 chromosome pseudomolecules had an ungapped length of 2 258 487 866 bp (Table 3), representing 93% of the complete genomic sequence assembled in this study. The number of annotated genes per chromosome and the corresponding chromosomes of other species are shown in Table 3. The number of annotated genes per chromosome and corresponding chromosomes of other species are shown in Table 3. Chromosomal fissions were inferred if multiple chromosomes in the *Odocoileus virginianus* assembly aligned to the same chromosome in another organism Gray cells in (Table 3). Similarly, fusions were inferred if a single

Table 2. Assembly statistics and BUSCO scores for white-tailed deer

	***O. v. borealis*(contig level)**	***O. v. borealis*(scaffold-level)**
Total length (bp)	2 461 348 864	2 424 946 708
Number of sequences	2420	508
Number of "N" gaps	n/a	311
% "N"	n/a	0.006%
Largest sequence (bp)	108 025 303	108 602 581
Smallest sequence (bp)	1939	2657
Average length (bp)	1 017 086.3	4 773 517.1
N50 (bp)	21 776 300	52 482 646
L50 (# of sequences)	32	18
N90 (bp)	3 308 695	10 477 849
L90 (# of sequences)	134	49
BUSCO† (*n* = 13 335)		
C: complete	93.2% (12 433)	93.2% (12 424)
S: single copy	90.9% (12 128)	91.0%(12 129)
D: duplicated	2.3% (305)	2.2% (295)
F: fragmented	0.4% (53)	0.4% (51)
M: missing	6.4% (849)	6.4% (860)

†Single-copy orthologous genes from the 22 species in the Cetartiodactyla lineage dataset.

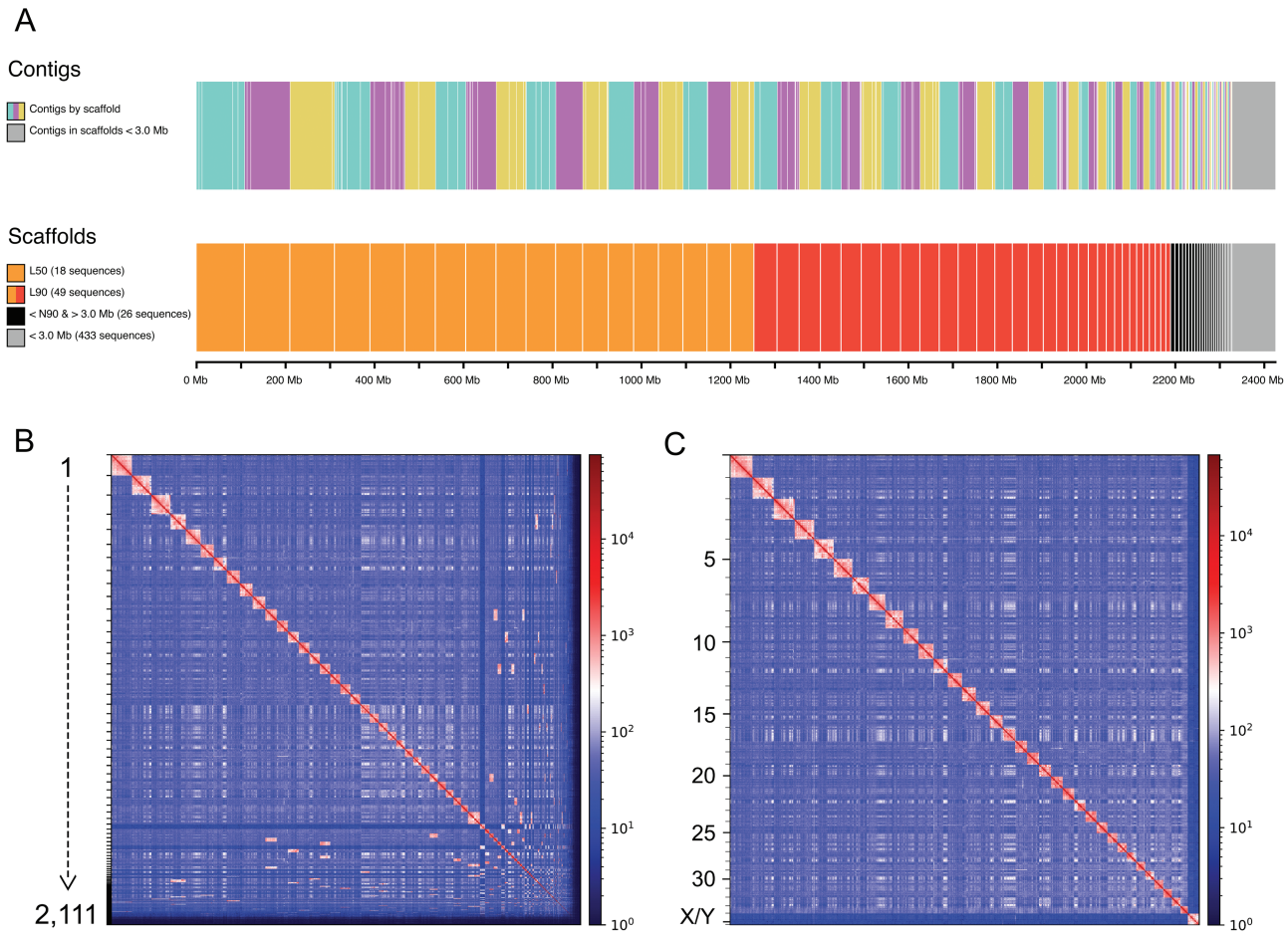

Figure 1. Contig, scaffold, and chromosome-level assemblies of the white-tailed deer genome. (**A**) Scaffolds are arranged by size (bottom) and their component contigs are arranged by scaffold (top). The largest scaffolds representing 50% (orange) and 90% (orange + red) of the assembly are indicated with color, leaving the remaining 10% of the assembly (black + gray). Scaffolds below 3 Mb (gray) are not visually separated. The number of contigs per scaffold is presented in Table 3. (**B**) Scaffold contact map generated from chromatin conformation capture Omni-C sequencing and visualized with HiCExplorer. Scaffold-scaffold contacts are shown increasing from blue to white, to red, and the strong diagonal signal represents scaffold self-association based on nuclear proximity. (**C**) Contact map for chromosome-sized pseudomolecules sequences manually curated into chromosomes.

chromosome in the *Odocoileus virginianus* assembly aligned to multiple chromosomes in another organism Bolded cells in(Table 3).

Genome Analysis

Annotation of Genes and Repetitive Elements

RepeatMasker identified 3 499 765 total interspersed repetitive elements in the Ovbo_1.0 assembly occupying a total of 1 034 014 200 bp. The genome had an average repeat density of 42.69% per scaffold, with the largest 36 scaffolds having a repeat density of 42.09%. Initial analysis using BRAKER predicted 46 152 complete genes, of which 37 684 were supported by external RNA or protein evidence. Validation of gene predictions with InterProScan supported 26 648 predicted genes. Of these supported genes, 5997 contained reverse transcriptase domains and were removed from the annotation set, for a final count of 20 651 protein-coding genes (Table 3).

Assessing Completeness Using BUSCO and Synteny

Initial BUSCO scores following assembly by Wtdbg2 were 89.6% complete genes (88.0% single-copy; 1.6% duplicated) and BUSCO was re-run following each step of analysis (Supplementary Table S2). The final BUSCO scores following scaffold deduplication were 93.2% complete genes (91.0% single-copy, 2.1% duplicated). Synteny comparisons between the white-tailed deer and Rocky Mountain elk assemblies showed single chromosomal fission of the Rocky Mountain elk chromosome 1 into the white-tailed deer chromosomes 12 and 17 (Table 3 and Supplementary Figure S2). Despite the greatly enhanced contiguity of the reference genome assembly achieved herein via 3GS sequencing, it should also be noted that the BUSCO scores from Seabury et al. 2011 are higher (93.7%) than those achieved in this study (93.2%); thereby reflecting the quality and precision of the previous 2GS assembly.

Discussion

Our chromosome-level assembly of the white-tailed deer genome will serve as a valuable resource for future ruminant and cervid research including molecular phylogeny and comparative evolutionary studies. By employing multiple sequencing technologies, including Illumina short-reads, Omni-C reads, and Pacific Biosciences long reads,

Table 3. Genome annotations and homology for the 36 chromosome pseudomolecules of white-tailed deer

Chrom. ID	Ungapped length (bp)	No. of gaps	No. of genes	No. of repeats	*Cervus canadensis*	[a]*Cervus elaphus*	*Cervus nippon*	[a]*Bos taurus*	[a]Ovis aries	*Homo sapiens*
1	108 600 581	4	721	174 661	3	18	4	4	4	7
2	102 048 420	2	929	173 101	5	11	6	11	1	9
3	100 279 162	1	1 253	169 281	4	9	5	7	5	5
4	93 958 800	7	628	158 735	7	19	8	1	1	3
5	93 570 283	16	956	164 814	2	20	3	3	1	1
6	89 349 494	3	813	150 017	6	12	7	10	7	15
7	85 956 676	3	583	141 968	8	15	9	**26/28**	22	10
8	80 668 930	2	385	133 362	9	30	10	12	10	13
9	78 136 789	7	685	134 814	10	23	1	13	13	20
10	73 421 497	8	704	130 149	11	1	11	15	15	11
11	72 668 630	4	589	123 588	13	14	13	16	12	1
12	68 288 379	2	574	118 932	1	16	2	17	17	22
13	68 111 889	5	304	109 301	15	33	14	2	2	2
14	67 564 244	4	319	117 421	16	25	15	20	16	5
15	66 412 021	3	332	115 247	12	21	12	14	9	8
16	61 986 249	6	472	107 329	17	13	16	21	18	15
17	60 095 371	8	1 077	101 312	1	5	2	19	11	17
18	57 744 214	5	336	93 516	14	29	18	8	2	9
19	57 482 545	4	252	93 221	20	28	26	9	8	6
20	57 216 540	5	1 059	98 539	18	4	1	18	14	**16/19**
21	56 275 464	4	254	91 774	19	**6/17**	17	6	6	4
22	55 840 698	2	246	91 928	23	27	21	24	23	18
23	53 708 925	2	546	96 459	21	22	20	5	3	12
24	52 991 459	1	465	88 100	25	3	23	5	3	12
25	51 970 072	1	210	88 431	27	31	24	1	1	21
26	47 961 987	1	376	76 790	22	24	19	22	19	3
27	45 470 101	4	603	74 352	28	7	25	23	20	6
28	44 772 963	4	506	77 294	29	2	28	29	21	11
29	43 582 846	2	249	81 117	26	6	27	6	6	4
30	43 498 002	3	426	77 730	24	33	22	2	2	1
31	43 483 628	4	329	77 828	30	16	29	8	2	9
32	41 958 503	5	221	66 679	31	32	30	27	26	8
33	40 612 519	2	659	77 187	32	10	31	25	24	16
34	35 913 106	0	238	57 077	33	26	32	9	8	6
X	54 563 062	18	340	97 953	X	X	X	X	X	X
Y	2 343 217	3	11	4 570	Y	Y	—[b]	X	X	X

Table 3. Continued

Chrom. ID	Ungapped length (bp)	No. of gaps	No. of genes	No. of repeats	*Cervus canadensis*	[a]*Cervus elaphus*	*Cervus nippon*	[a]*Bos taurus*	[a]Ovis aries	*Homo sapiens*
Placed	2 258 507 266		18 869	3 834 577	—	—	—	—	—	—
Unplaced	166 333 442	-	1 782	281 882	—	—	—	—	—	—
Total	2 424 840 708		20 651	4 116 459	—	—	—	—	—	—

Gray cells—multiple chromosomes in the *Odocoileus virginianus* assembly aligned to the same chromosome in another organism. Bold cells—a single chromosome in the *Odocoileus virginianus* assembly aligned to multiple chromosomes in another organism.
[a]Chromosomes (chrom.) for this species are not numbered in order of size.
[b]No Y chromosome sequence available for *Cervus nippon*.

the contiguity, and accuracy of the assembly were able to surpass those of previously generated Capreolinae (New World deer) genomes; *Rangifer tarandus* (GCA_014898785), and *Odocoileus hemionus* (GCA_004115125). This work, resulting in the Ovbor_1.0 assembly, used currently available long-read 3GS and Omni-C technologies to produce a scaffold N50 of 52 Mb, which is 60 times longer than the scaffold N50 of the existing *Odocoileus virginianus texanus* assembly (GCA_002102435; Seabury et al. 2011) generated before 3GS became available. The current assembly has an average of fewer than 5 gaps per chromosome and will serve as a valuable reference genome for genomic studies in white-tailed deer and other cervids.

The assembly produced during step 1 by Wtdbg2 produced large contiguous sequences, with an NG50 of 21 Mb. This is comparable to the human Wtdbg2 assembly presented by Ruan and Li (2020), which had an NG50 of 18 Mb. Arrow long-read polishing was able to extend and error-correct contigs produced by Wtdbg2. Pilon polishing was performed iteratively; however, the most accurate error correction was complete after a single iteration (Supplementary Table S2). This may be compared with the study by Nguyen et al. (2020) where 4 rounds of pilon polishing were required, following the use of Oxford Nanopore technology. Oxford Nanopore reads use complex electrical signals, and long-range errors can occur (Rang et al. 2018). By contrast, in PacBio reads, errors are characterized by insertions and deletions, and a single round of pilon polishing produced an assembly with a higher BUSCO score (Supplementary Table S2). Furthermore, each round of pilon polishing can take 1– 2 days to complete depending on the size of the genome; thereby reducing the time spent on error correction and improving overall pipeline efficiency.

BUSCO results indicated that only a small percentage of the sequence was duplicated. During purging by Purge-Haplotigs, it was only necessary to purge 98.6 Mb of sequence, which was less than 5% of the total genome length and was comprised of putative artifact and haplotig sequences. By contrast, other genome assemblies require almost 50% of the genome sequence to be purged (Roach et al. 2018). Thus, the assembly produced by Wtdbg2 contained primarily collapsed haplotype sequences without high levels of duplication. Furthermore, only 8 contigs had missassemblies that were able to be detected by Salsa, (i.e., <0.1% of Wtdbg2 contigs), indicating that almost all contigs were in the chromosome order implicated by Omni-C.

Our genome annotation produced by BRAKER and validated with InterProScan expanded the set of annotations on chromosome-sized sequences. This annotation will provide further genomic context allowing for the assessment of chromosomal rearrangements and evolutionary relationships in white-tailed deer. Although annotations were validated using human protein sequences, research has shown that lineage-specific traits such as antler growth have their genetic basis in genes (referred to as headgear genes) that are shared across mammalian lineages; therefore, it is unlikely that InterProScan validation led to a loss of lineage-specific genes. Some genes have been shown to be under positive selection in ruminants with headgear traits (OLIG1 and OTOP3), while others have been shown to be highly expressed in headgear (i.e., SOX10, NGFR, ALX1, VCAN, COL1A1) (Chen et al. 2019 and Wang et al. 2019). This annotation information may also facilitate future syntenic comparisons utilizing further gene-based synteny. Chromosome fissions and fusions were detected between

the white-tailed deer genome and the other species that were compared (Table 3). Identification of chromosomal arrangements will inform the assumptions made about gene linkage and synteny.

Future genome-wide association studies will be able to make alignments to the chromosome-level scaffolds of the 3GS Ovbor_1.0 white-tailed deer reference assembly. Having a chromosome-level assembly with few gaps will empower future population genomic sequencing to characterize genetic diversity within the deer population that could identify underlying genetic disease resistance loci and assist with current conservation efforts.

Supplementary Material

Supplementary data are available at *Journal of Heredity* online.

Funding

This project was supported by the U.S. Fish and Wildlife Service [Federal Aid in Wildlife Restoration (W-146-R)]. With additional funding from the Illinois Natural History Survey – Prairie Research Institute and the Office of the Vice Chancellor of Research, at the University of Illinois Urbana-Champaign.

Conflict of Interest

The authors declare there is no conflict of interest.

Acknowledgments

We thank the Illinois Department of Natural Resources biologists for their efforts in conducting surveillance for chronic wasting disease and for allowing us to sample the animal used in this study. We thank Dr Alvaro Hernandez and the staff of the Roy J. Carver biotechnology center at UIUC for their genomic sequencing and consultation. We thank Kimberly Walden and Dr Christopher Fields of the UIUC HPCBio facility for their consultancy and assistance with the genome assembly pipeline. We thank Dr Julian Catchen of the Department of Evolution, Ecology, and Behavior at UIUC for providing insight into bioinformatics methods and analyses. Additionally, we thank Dr Daniel Raudabaugh for his courtesy reviews and genomic discussions.

Data Availability

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAJQKH000000000. Illumina, Omni-C, and PacBio sequencing reads have been deposited in NCBI Sequence Read Archive (SRR17118554, SRR17118555, SRR17162326, SRR17162327). The GFF file produced by BRAKER is provided as data in Supplementary Material. The GFF file following validation by InterProScan is provided as data in Supplementary Material.

References

Allen T, Olds E, Southwick R, Scuderi B, Howlett D, Caputo L. 2018. Hunting in America: an economic force for conservation. *National Shooting Sports Foundation*. 2018 Edition:10.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403–410.

Bana NA, Nyiri A, Nagy J, Frank K, Nagy T, Stéger V, Schiller M, Lakatos P, Sugár L, Horn P, et al. 2018. The red deer Cervus elaphus genome CerEla1.0: sequencing, annotating, genes, and chromosomes. *Mol Genet Genomics*. 293:665–684.

Bao W, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 6:11.

Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. 2011. BamTools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*. 27:1691–1692.

Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. 2012. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*. 58:268–276.

Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, et al. 2021. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res*. 49(D1):D344–D354. doi:10.1093/nar/gkaa977.

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. 2007. UniProtKB/Swiss-Prot. *Methods Mol Biol*. 406:89–112.

Brandt AL, Green ML, Ishida Y, Roca AL, Novakofski J, Mateus-Pinilla NE. 2018. Influence of the geographic distribution of prion protein gene sequence variation on patterns of chronic wasting disease spread in white-tailed deer (Odocoileus virginianus). *Prion*. 12:204–215.

Brůna T, Lomsadze A, Borodovsky M. 2020. GeneMark-EP+: Eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinf*. 2:lqaa026.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 12:59–60.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinf*. 10:421.

Chen S, Zhou Y, Chen Y, Gu J. 2018. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 34:i884–i890.

Chen L, Qiu Q, Jiang YU, Wang K, Lin Z, Li Z, Bibi F, Yang Y, Wang J, Nie W, Su W. 2019. Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science*. 364:eaav6202.

Dobin A, Gingeras TR. 2015. Mapping RNA-seq reads with STAR. *Curr Protoc Bioinformatics*. 51:11.14.1–11.14.19.

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, et al. 2012. Mind the gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One*. 7:e47768.

Fuentes-Pardo AP, Ruzzante DE. 2017. Whole-genome sequencing approaches for conservation biology: advantages, limitations and practical recommendations. *Mol Ecol*. 26:5369–5406.

Genome Reference Consortium. 2021. *Assembly terminology - Genome Reference Consortium*. Retrieved September 17, 2021, from https://www.ncbi.nlm.nih.gov/grc/help/definitions

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. 2019. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol*. 15:e1007273.

Gotoh O. 2008. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Res*. 36:2630–2638.

Güere ME, Våge J, Tharaldsen H, Benestad SL, Vikøren T, Madslien K, Hopp P, Rolandsen CM, Røed KH, Tranulis MA. 2020. Chronic wasting disease associated with prion protein gene (*PRNP*) variation in Norwegian wild reindeer (*Rangifer tarandus*). *Prion* 14(1):1–10. doi:10.1080/19336896.2019.1702446.

Hewitt DG. 2011. *Biology and management of white-tailed deer*. CRC Press. https://books.google.com/books?id=k4DOBQAAQBAJ

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*. 32:767–769.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-genome annotation with BRAKER. *Methods Mol Biol*. 1962:65–95.

Hufnagel DE, Hufford MB, Seetharam AS. 2020. SequelTools: a suite of tools for working with PacBio Sequel raw sequence data. *BMC Bioinf*. 21:429.

Ishida Y, Tian T, Brandt AL, Kelly AC, Shelton P, Roca AL, Novakofski J, Mateus-Pinilla NE. 2020. Association of chronic wasting disease susceptibility with prion protein variation in white-tailed deer (*Odocoileus virginianus*). *Prion*. 14:214–225.

Iwata H, Gotoh O. 2012. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Res*. 40:e161.

Jamieson A, Anderson SJ, Fuller J, Côté SD, Northrup JM, Shafer ABA. 2020. Heritability estimates of antler and body traits in white-tailed deer (*Odocoileus virginianus*) from genomic-relatedness matrices. *J Hered*. 111:429–435.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 30:1236–1240.

Jones E, Hummerich H, Viré E, Uphill J, Dimitriadis A, Speedy H, Campbell T, Norsworthy P, Quinn L, Whitfield J, et al. 2020. Identification of novel risk loci and causal insights for sporadic Creutzfeldt-Jakob disease: a genome-wide association study. *Lancet Neurol*. 19:840–848.

Keller O, Kollmar M, Stanke M, Waack S. 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*. 27:757–763.

Kong A, Cox NJ. 1997. Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet*. 61:1179–1188.

Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. 2019. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res*. 47:D807–D811.

Lamb S, Taylor AM, Hughes TA, Mcmillan BR, Larsen RT, Khan R, Weisz D, Dudchenko O, Aiden EL, Edelman, NB, Frandsen PB. 2021. *De novo* chromosome-length assembly of the mule deer (*Odocoileus hemionus*) genome. *Gigabyte*. 2021:1–13.

Lewis TE, Sillitoe I, Dawson N, Lam SD, Clarke T, Lee D, Orengo C, Lees J. 2018. Gene3D: extensive prediction of globular domains in proteins. *Nucleic Acids Res*. 46:D1282.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 34(18):3094–3100. doi:10.1093/bioinformatics/bty191.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and samtools. *Bioinformatics*. 25:2078–2079.

Li R, Yang P, Li M, Fang W, Yue X, Nanaei HA, Gan S, Du D, Cai Y, Dai X, et al. 2021. A Hu sheep genome with the first ovine Y chromosome reveal introgression history after sheep domestication. *Sci China Life Sci*. 64:1116–1130.

Lomsadze A, Burns PD, Borodovsky M. 2014. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res*. 42:e119.

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*. 33:6494–6506.

Mahmoud M, Zywicki M, Twardowski T, Karlowski WM. 2019. Efficiency of PacBio long read correction by 2nd generation Illumina sequencing. *Genomics*. 111:43–49.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol*. 38:4647–4654.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*. 17:10–12.

Masonbrink RE, Alt D, Bayles DO, Boggiatto P, Edwards W, Tatum F, Williams J, Wilson-Welder J, Zimin A, Severin A, et al. 2021. A pseudomolecule assembly of the Rocky Mountain elk genome. *PLoS One*. 16:e0249899.

Mehta J, Starmer C, Sugden R, Schelling T, Kahneman D, Stanovich K, West R, Rubinstein A, Jung R, Haier R. et al. 2009. The genome sequence of Taurine cattle: a window to ruminant biology and evolution. *Science* 324:522–28.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res*. 49:D412–D419.

Nguyen SV, Greig DR, Hurley D, Donoghue O, Cao Y, McCabe E, Mitchell M, Schaffer K, Jenkins C, Fanning S. 2020. Yersinia canariae sp. nov., isolated from a human yersiniosis case. *Int J Syst Evol Microbiol*. 70:2382–2387.

National Center for Biotechnology Information. 2016. The UniVec Database. In *NCBI*. https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/.

Pietsch LR. 1954. White-tailed deer populations in Illinois. *Biological Notes*. 34:1–24.

Perrin-Stowe TIN, Ishida Y, Terrill EE, Hamlin BC, Penfold L, Cusack LM, Novakofski J, Mateus-Pinilla NE, Roca AL. 2020. Prion Protein Gene (PRNP) sequences suggest differing vulnerability to chronic wasting disease for florida key deer (odocoileus virginianus clavium) and columbian white-tailed deer (O. v. leucurus). *J Hered*. 111:564–572.

Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. 2018. Long reads: their purpose and place. *Hum Mol Genet*. 27: R234–R241.

Potter S, Jason GB, Mozes PKB, Janine ED, Mark K, Mark DBE, Craig M. 2017. Chromosomal speciation in the genomics era: disentangling phylogenetic evolution of rock-wallabies. *Front Genet*. 8. Article no.: 10. doi:10.3389/fgene.2017.00010.

Price SA, Bininda-Emonds OR, Gittleman JL. 2005. A complete phylogeny of the whales, dolphins and even-toed hoofed mammals (Cetartiodactyla). *Biol Rev Camb Philos Soc*. 80:445–473.

Ramírez F, Vivek B, Laura A, Kin CL, Björn AG, José V, Bianca H, Asifa A, Thomas M. 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun*. 9(1):1–15. doi:10.1038/s41467-017-02525-w.

Rang FJ, Kloosterman WP, de Ridder J. 2018. From Squiggle to Basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol*. 19(1):90. doi:10.1186/s13059-018-1462-9.

Rivera NA, Brandt AL, Novakofski JE, Mateus-Pinilla NE. 2019. Chronic wasting disease in cervids: prevalence, impact and management strategies. *Vet Med: Res Rep* 10:123–139. doi:10.2147/vmrr.s197404.

Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinf*. 19:460.

Robinson SJ, Samuel MD, O'Rourke KI, Johnson CJ. 2012. The role of genetics in chronic wasting disease of North American cervids. *Prion*. 6:153–162.

Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. 2018. Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst*. 6:256–258.e1.

Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nat Methods*. 17:155–158.

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, et al. 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res*. 27:849–864.

Seabury CM, Bhattarai EK, Taylor JF, Viswanathan GG, Cooper SM, Davis DS, Dowd SE, Lockwood ML, Seabury PM. 2011. Genome-wide polymorphism and comparative analyses in the white-tailed deer (*Odocoileus virginianus*): a model for conservation genomics. *PLoS One*. 6:e15811.

Seabury CM, Oldeschulte DL, Bhattarai EK, Legare D, Ferro PJ, Metz RP, Johnson CD, Lockwood MA, Nichols TA. 2020. Accurate

genomic predictions for chronic wasting disease in U.S. white-tailed deer. *G3 (Bethesda)*. 10:1433–1441.

Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2013-2015 http://www.repeatmasker.org.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*. 24:637–644.

Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinf*. 7:62.

United States Department of Agriculture National Agricultural Statistics Service. 2019. United States summary and state data. *2017 Census of Agriculture*, 28.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 9:e112963.

Wang Y, Zhang C, Wang N, Li Z, Heller R, Liu R, Zhao Y, Han J, Pan X, Zheng Z, Dai X. 2019. Genetic basis of ruminant headgear and rapid antler regeneration. *Science*. 364:eaav6335.

Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol*. 13:e1005595.

Xiumei X, Ai C, Wang T, Li Y, Liu H, Hu P. 2021. The first high-quality reference genome of Sika deer provides insights for high-tannin adaptation. BioRxiv preprint article.