RESEARCH ARTICLE

A SNP resource for studying North American moose [version 1; referees: 2 approved, 1 approved with reservations]

Theodore S. Kalbfleisch[1], Brenda M. Murdoch[2], Timothy P. L. Smith[3], James D. Murdoch[4], Michael P. Heaton[3], Stephanie D. McKay[4]

[1]Department of Biochemistry and Molecular Biology, School of Medicine, University of Louisville, Louisville, Kentucky, USA
[2]University of Idaho, Moscow, Idaho, USA
[3]U.S. Meat Animal Research Center, Clay Center, Nebraska, USA
[4]University of Vermont, Burlington, Vermont, USA

v1 **First published:** 10 Jan 2018, **7**:40 (doi: 10.12688/f1000research.13501.1)
Latest published: 10 Jan 2018, **7**:40 (doi: 10.12688/f1000research.13501.1)

Abstract

Background: Moose (*Alces alces*) colonized the North American continent from Asia less than 15,000 years ago, and spread across the boreal forest regions of Canada and the northern United States (US). Contemporary populations have low genetic diversity, due either to low number of individuals in the original migration (founder effect), and/or subsequent population bottlenecks in North America. Genetic tests based on informative single nucleotide polymorphism (SNP) markers are helpful in forensic and wildlife conservation activities, but have been difficult to develop for moose, due to the lack of a reference genome assembly and whole genome sequence (WGS) data.

Methods: WGS data were generated for four individual moose from the US states of Alaska, Idaho, Wyoming, and Vermont with minimum and average genome coverage depths of 14- and 19-fold, respectively. Cattle and sheep reference genomes were used for aligning sequence reads and identifying moose SNPs.

Results: Approximately 11% and 9% of moose WGS reads aligned to cattle and sheep genomes, respectively. The reads clustered at genomic segments, where sequence identity between these species was greater than 95%. In these segments, average mapped read depth was approximately 19-fold. Sets of 46,005 and 36,934 high-confidence SNPs were identified from cattle and sheep comparisons, respectively, with 773 and 552 of those having minor allele frequency of 0.5 and conserved flanking sequences in all three species. Among the four moose, heterozygosity and allele sharing of SNP genotypes were consistent with decreasing levels of moose genetic diversity from west to east. A minimum set of 317 SNPs, informative across all four moose, was selected as a resource for future SNP assay design.

Conclusions: All SNPs and associated information are available, without restriction, to support development of SNP-based tests for animal identification, parentage determination, and estimating relatedness in North American moose.

Open Peer Review

Referee Status: ? ✓ ✓

	Invited Referees 1	2	3
version 1 published 10 Jan 2018	? report	✓ report	✓ report

1 **Joshua M. Miller**, Université du Québec à Montréal, Canada
University of Alberta, Canada

2 **Paul Stothard**, University of Alberta, Canada

3 **Kris J. Hundertmark**, University of Alaska Fairbanks, USA

Discuss this article

Comments (0)

Corresponding authors: Theodore S. Kalbfleisch (ted.kalbfleisch@louisville.edu), Michael P. Heaton (mike.heaton@ars.usda.gov), Stephanie D. McKay (stephanie.mckay@uvm.edu)

Author roles: Kalbfleisch TS: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Murdoch BM**: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Smith TPL**: Funding Acquisition, Investigation, Methodology, Resources, Writing – Review & Editing; **Murdoch JD**: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Heaton MP**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **McKay SD**: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Kalbfleisch TS, Murdoch BM, Smith TPL *et al.* **A SNP resource for studying North American moose [version 1; referees: 2 approved, 1 approved with reservations]** *F1000Research* 2018, **7**:40 (doi: 10.12688/f1000research.13501.1)

Grant information: Funding for this research was provided by the USDA, ARS appropriated projects 5438-32000-033-00D and 5438-31320-012-00D, the USDA National Institute of Food and Agriculture, McIntire-Stennis project 1002300, the University of Vermont College of Agriculture and Life Sciences, the College of Agricultural and Life Sciences at the University of Idaho, with the resources of the University of Louisville's research computing group and the Cardinal Research Cluster.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 10 Jan 2018, **7**:40 (doi: 10.12688/f1000research.13501.1)

Introduction

Alces alces is the largest member of the Cervidae family, and ranges throughout the circumpolar boreal forests of Eurasia and North America[1,2]. The species diverged from the ancestors of domestic cattle and sheep approximately 27 million years ago[3]. Moose are important ecologically, as a large ungulate with strong ecosystem impacts; economically, due to their value for tourism and hunting; and culturally, as a prominent symbol in many regions[4]. Consequently, there is active management of moose populations by wildlife agencies throughout their range in North America. However, management is hampered by a lack of genetic tools for monitoring moose, assessing the genetic health of populations, and even detecting illegal harvesting. Moose populations appear to be declining in some regions, including parts of the Upper Midwest of the United States[5,6], and effective management is often dependent on data that are logistically challenging and/or costly to collect.

Identifying individual animals and measuring relatedness among and within populations are important for effective wildlife management and conservation efforts[7–9]. Identifying individuals can be as simple as observing their unique color patterns, for example in wild dogs (*Lycaon pictus*)[10]. However, this is not practical in species such as moose that have few features with obvious variation between individuals. Moreover, coat color patterns provide little information about genetic relatedness. Association of younger and older animals has been used to infer relationships, for example in swift foxes (*Vulpes velox*) where pups at a den are presumed to be offspring of the attending parents based on the monogamous behaviors they exhibit. However, detailed parentage studies have revealed multiple paternity within swift fox litters[11] and other fox species[12]. Generally, genetic testing provides a more accurate assignment of parentage and supports unique identification of individuals in the vast majority of instances, as well as an estimation of intra- and inter-population genetic variability.

Genetic testing using DNA markers has been applied to human, livestock, and wildlife studies for many years[13–16]. This form of testing first gained popularity with the development of microsatellite short tandem repeat (STR), and mitochondrial genome markers, concurrent with the development of DNA amplification and sequencing technologies. Approximately 5 to 11 microsatellite markers from cattle, sheep, and caribou (*Rangifer tarandus*) have been adapted for moose studies[17–22]. These studies form the basis of our current understanding of North American moose population structure and genetic diversity. DNA technology developments in the past decade have led to the replacement of microsatellite and mitochondrial genome markers with SNP markers because SNPs are more abundant, have greater stability over generations, are more accurately genotyped, and are amenable to automating the genotyping processes[23]. Moreover, panels of SNPs broaden the use of genotyping for management and conservation efforts, because they can provide not only identification of individuals and parentage, but also estimation of inbreeding and relatedness, and detection of admixture between populations of wildlife. For example, an SNP-based approach has been used for conservation efforts in endangered species such as the Iberian lynx (*Lynx pardinus*)[24] and Tasmanian devil (*Sarcophilus harrisii*)[25], as well as more common but wide-ranging species like the brown bear (*Ursus arctos*)[26]. The application of SNP-based approaches, however, requires first the identification of polymorphisms segregating in the populations being studied, and developing assays that support accurate genotyping. Accordingly, a SNP panel spanning the genome of North American moose would be useful for addressing fundamental questions about population genetics in this species.

Low genetic diversity among North American moose populations has been previously reported[17,18], making development of SNP panels challenging. The low diversity has been attributed in the prevailing theory, to a relatively recent (ca. 11,000–14,000 years ago) colonization from Asia and subsequent founder effect induced by extended range expansion from an original small group of animals[27]. However, no definitive evidence has been presented that refutes an alternative hypothesis, that North American moose experienced a severe population bottleneck at some time in the past[20], as occurred for North American bison (*Bison bison*) populations[28]. Whether a founder effect, bottleneck, or both, the small effective population size simultaneously increases the need for developing genetic tools for management and the challenge of creating SNP marker panels.

Discovery of SNPs in a species has generally been preceded by development of its reference genome assembly. Using this assembly, whole genome sequence (WGS) reads from individual animals can be aligned, and differences between segregating alleles identified. However, creation of a reference assembly still represents a significant barrier for most research communities interested in wildlife species. Fortunately, an alternative approach that uses the reference genomes of related species has been developed[29], and shown to effectively identify high-confidence SNPs likely to be segregating within the target species. Here we report the whole genome sequencing of four moose genomes, each obtained from distant geographic regions of North America, and the use of the cattle and sheep reference genomes to align the sequence data and identify SNPs likely to be segregating among moose populations. A set of criteria was developed to select the potentially most useful set of moose SNPs, and to identify 317 autosomal variants meeting these criteria. The associated sequence information was made freely available, and represents a resource for developing genotyping assays to support moose genetic research.

Methods

Ethical statement

This article contains no studies performed with animal subjects, and thus, no additional institutional ethical permits were required. Samples for DNA extraction were donated by private individuals not associated with this research. These were hunters that had legally harvested moose during the firearm hunting season in their state. No additional approvals were needed, since all hunters obtained valid hunting licenses for the harvesting of moose.

Animal samples

Samples of muscle tissue were obtained from four animals likely comprising three putative subspecies of *A. alces* based on their location in North America: *A. gigas*, *A. shirasi*, and *A. americana*[30]. These animals were harvested at four distinct geographic locations

(Figure 1) and entered as BioSamples in NCBI BioProject Accession PRJNA325061 (Table 1). As is typical, hunters removed the internal organs in the field, the carcasses were chilled, and the meat was subsequently processed for frozen storage. Each of the four owners donated approximately 50 g of frozen tissue from their harvested animal, and that tissue was archived at USMARC for use in this project.

WGS production, alignment, and SNP genotyping

DNA was extracted from muscle with a typical phenol: chloroform method and stored at 4°C in 10 mM TrisCl, 1 mM EDTA (pH 8.0) as previously described[31]. Approximately 5 µg of moose genomic DNA was fragmented by focused-ultrasonication to generate fragments less than 800 bp long (Covaris, Inc. Woburn, Massachusetts USA). These fragments were used to make an indexed, 500 bp paired-end library according to the manufacturer's instructions (TruSeq DNA PCR-Free LT Library Preparation Kits A and B, Illumina, Inc., San Diego, California USA). After construction, indexed libraries were pooled with other indexed samples in groups of four to eight, and sequenced with a massively parallel sequencing machine and high-output kits (NextSeq500, two by 150 paired-end reads, Illumina Inc.). After sequencing, the raw reads were filtered to remove adaptor sequences, contaminating dimer sequences, and low-quality reads. Pooled libraries with compatible indexes were repeatedly sequenced until a minimum of 40 Gb of sequence with greater than Q20 quality was collected for each animal. Previous results showed that this level of coverage provided genotype scoring rates and accuracies that exceeded 99%[29].

The DNA sequence alignment process was similar to that previously reported[29]. Briefly, FASTQ files corresponding to a minimum of 40 Gb of Q20 sequence were aggregated for each animal. The reference assemblies for both UMD3.1[32] and Oar_v3.1 were downloaded from the NCBI genomes download site and indexed for use with the Burrows Wheeler aligner (BWA) version 0.7.12[33]. The fastq files corresponding to R1 and R2 runs for the paired end libraries of each respective animal were aligned individually using the BWA aln algorithm and bovine reference assembly UMD3.1. The R1 and R2 datasets were then merged and collated using BWA sampe. The process was repeated for the mapping of the reads to the ovine Oar_v3.1 reference assembly. The resulting sequence alignment map (SAM) files were converted to binary alignment map (BAM) files, and subsequently sorted using Samtools (version 0.1.18)[34]. PCR duplicates were marked in the BAM files using the Genome Analysis Toolkit (GATK, version 1.5-32-g2761da9)[35]. Regions in the mapped dataset that would benefit from realignment due to small insertions and deletions were identified using the GATK module RealignerTargetCreator, and realigned using the module IndelRealigner. The BAM file produced at each of these steps was indexed using Samtools. The resulting indexed BAM files were made available via the Intrepid Bioinformatics genome browser http://www.intrepidbio.com/, with groups of animals linked at the USMARC WGS browser (mapped to cattle, mapped to sheep). The raw reads were deposited at NCBI BioProject Accession PRJNA325061. Some SNP variants were identified manually by inspecting the target sequence with Integrative Genomics Viewer (IGV) software version 2.1.28[36,37], as described in previously[38]. In these cases, read depth, allele count, allele position in the read, and quality score were taken into account when the manual genotype determination was made.

Variant detection and filtering

The above mapping efforts produced BAM files for the alignments to both UMD3.1, and Oar_v3.1. The BAM files for all four

Figure 1. Locations in North America where the moose used in this study were collected.

Table 1. Whole genome sequence information and alignment statistics for four North American moose.

								Reads mapped (%)	
Species	**Animal identifier**	**BioSample number**	**Location or breed**	**Sex**	**Estimated aligned read depth[a]**	**Gb sequence collected**	**Total reads (millions)**	**Bt_UMD3.1**	**Oar_v3.1**
Alces alces gigas	HM2013 (201524011)	7695254[b]	Paxson, Alaska, USA	M	19.2	64.1	461	13.0	10.4
Alces alces shirasi	JC2001 (200124009)	7695255[b]	Green River Lakes, Wyoming, USA	F	13.7	45.7	327	13.5	10.8
Alces alces americana	R199	7695256[b]	Lowell, Vermont, USA	M	23.2	77.3	653	8.8	6.8
Alces alces shirasi	Clearwater06	7695257[b]	Elk River, Idaho, USA	M	20.0	66.8	549	8.9	6.9
Bos taurus	19969811	5216015[c]	USA Hereford	M	14.2	47.5	314	88.4	nd[d]
Ovis aries	199735001	5216748[e]	USA Rambouillet	M	13.7	42.7	319	22.2	83.8

Key:

[a]Based on the observed linear relationship between aligned read depth (y) and total gigabases (Gb) of genomic sequence collected (x) with a quality score ≥ to 20, where y = 0.3x in cattle and sheep genomes[38,39].

[b]NCBI BioProject number 325061

[c]NCBI BioProject number 324822

[d]Not determined

[e]NCBI BioProject number 324837

animals were analyzed simultaneously for variation against both the UMD3.1 and the Oar_v3.1 genomes. The GATK UnifiedGenotyper was used with the genotype mode (-gt_mode) flag set to DISCOVERY, and the likelihood model (-glm) flag was set to BOTH in order to identify both single nucleotide variants, and small insertions and deletions. The maximum number of alternate alleles (--max_alternate_alleles) flag was set to allow only three. Other than those mentioned, default parameters were used. Samtools was used to generate a pileup file containing the measured allele and depth of coverage at each position for all four animals. Variant sites in the four moose were filtered for having a minimal read depth of ten, and a minimum genotype quality score of 30. The SNPs were filtered for having a minor allele frequency (MAF) of 0.5, with both homozygous genotypes present among four animals. Fifty bases of flanking DNA sequence on either side of the targeted moose SNP were analyzed for nucleotide alleles that were homozygous in all four moose yet different from the cattle or sheep reference sequences. These nucleotide sites were flagged as potential moose "species-specific" alleles and the 101 bp of context sequence was edited to create a moose consensus reference sequence. The 101 bp of moose consensus sequence derived from the alignment of one reference genome was then tested for alignment to the other reference genome. Moose SNPs with MAFs of 0.5, and having been derived independently from alignment to both reference genomes, were manually assigned genome-wide bins based on their chromosome and proximity as inferred by alignment with the cattle genome. The goal of assigning markers to bins was to minimize linkage while allowing automated SNP assay design software the opportunity to select the best candidate marker for each distinct genomic region. All of these conservative filters were intended to maximize marker informativity in North American moose populations, and minimize potential technical difficulties with SNP assay designs that rely on oligonucleotide hybridization for genotype detection.

Results

An average of 63.5 Gb total genome sequence was collected for four moose. Based on similar estimates in cattle and sheep, this would correspond to an average read depth of 19-fold coverage if aligned to a moose reference genome of similar quality (Table 1). However, when cattle and sheep reference genomes were used, an average of 11.0 and 8.7% of the moose reads were aligned, respectively. For comparison, the same alignment method was performed with sets of bovine and ovine genomic sequences and resulted in 88.4% and 83.8% reads aligned to their respective genome assemblies (Table 1). For cross-species comparison, 22.2% of the ovine set of genomic sequence reads were aligned to the bovine assembly. Although the moose read depth was low when averaged across the entire genome of cattle or sheep, at conserved genome regions it was consistent with the expected average read depth of 19-fold. Thus, the moose read depth in conserved genomic regions appeared to be sufficient for identifying polymorphic sites and accurately assigning variant alleles.

Alignment of moose reads to the cattle and sheep genomes identified approximately 48.3 million and 39.7 million sites that differed from the reference assemblies, respectively. These included SNPs, insertions and deletions, and sites where

moose-associated nucleotide differences occurred. The latter sites were defined as having homozygous genotypes in the four moose, with alleles differing from those in cattle or sheep (Figure 2). After stringent filtering for read depth and alignment quality, there were 1,095,371 and 813,006 moose variants identified with the respective cattle and sheep genome assemblies (Table 2). Approximately 96% of these were homozygous moose-associated nucleotide differences (Supplementary file S1 and Supplementary file S2). The remaining 46,005 and 36,934 variants were moose SNPs identified by the respective cattle and sheep alignments (Supplementary file S3 and Supplementary file S4). The MAF distribution of the moose SNPs was similar for both sets with the large group having a 0.125 MAF (approximately 37%, Table 2). The most informative moose SNPs (i.e., "highly informative") were defined as those with a 0.5 MAF and both homozygous genotypes present among any of the four moose, and are candidate SNPs that may have arisen to a high MAF prior to the species arrival in North America (Figure 3). There were 1,341 and 1,014 of these moose SNPs identified with the cattle and sheep alignments, respectively (Table 2).

Candidate moose SNPs were further excluded when the flanking sequences in one reference genome were not uniquely identified in the other. This left 773 and 552 highly informative moose SNPs identified in conserved regions of the cattle and sheep genomes, respectively (Supplementary Table S1 and Supplementary Table S2). Of these 1,325 highly informative SNPs, 1,008 were unique between the two sets, while 317 were common to both sets. The latter represents the most informative moose SNPs, with the highest flanking sequence conservation, due to their independent alignment to both reference genomes (Table 2).

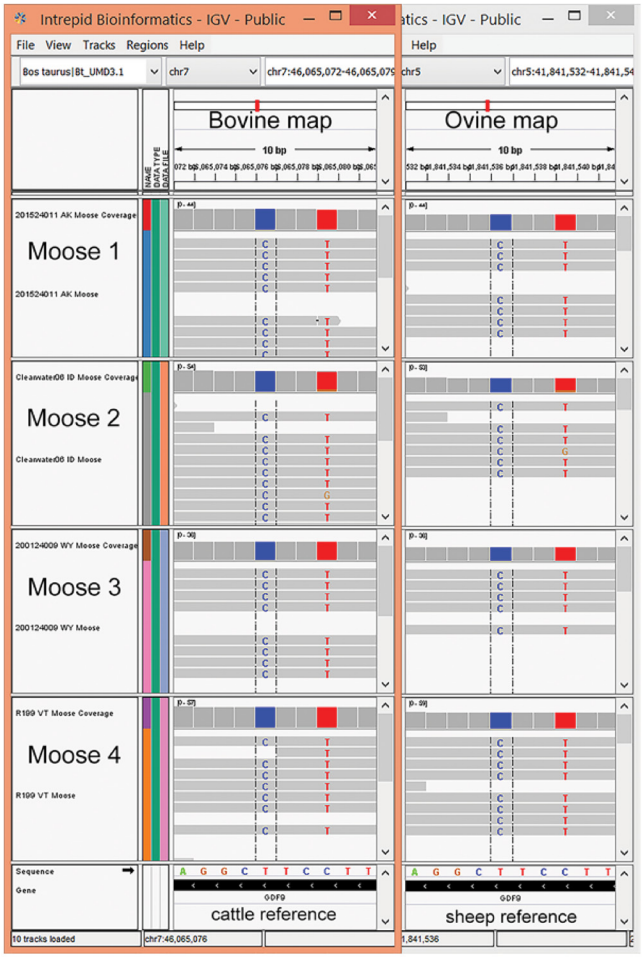

Figure 2. Computer screen images of species-associated nucleotide differences in moose. Overlapping computer screen images of moose WGS data aligned to bovine and ovine reference genomes, respectively, showing two moose-associated nucleotides in the *GDF9* gene. The bovine UMD3.1 and ovine Oar_v3.1 map positions for the variant sites are chr7:46,065,076 - 46,065,079 and chr5: 41,841,536 - 41,841,536,539, respectively.

Table 2. North American moose SNPs identified by aligning WGS to bovine and ovine reference genomes.

		Reference assembly	
Step	**Progressively filtered variant sets**	**Bt_UMD3.1**	**Oar_v3.1**
1	Variants passing depth and quality filters[a]	1,095,371	813,006
2	Moose-associated nucleotide differences[b]	1,049,080	775,700
2	SNPs	46,005	36,934
3	SNPs with 0.125 MAF	16,815	13,698
3	SNPs with 0.250 MAF	12,502	10,106
3	SNPs with 0.375 MAF	13,369	10,710
3	SNPs with 0.500 MAF	3,319	2,420
4	SNPs with 0.500 MAF and both homozygotes present	1,341	1,014
5	SNPs with conserved flanking regions[c]	773	552
6	SNPS with independent alignment to both reference genomes[d]	317	317

Key:

[a]Autosomal chromosome alignment with minimum read depth of ten and minimum genotyping quality score of 30. There were approximately 60,600 and 58,200 additional moose SNPs in the respective UMD3.1 and Oar_v3.1 alignments that were heterozygous in all four moose. However, these were excluded from the SNP counts because this artifact is caused by sequence read misalignment.

[b]These sites are difference from the reference and homozygous in all four moose.

[c]101 bp flanking regions of SNPs with 0.5 MAF that could be unambiguously identified by BLAT alignment to the other reference assembly (Table S1 and Table S2). These regions also contained no SNPs among the four moose.

[d]SNP that were independently identified by alignment in each reference genome and manually grouped into 216 chromosomal bins for assay design (Table S3).

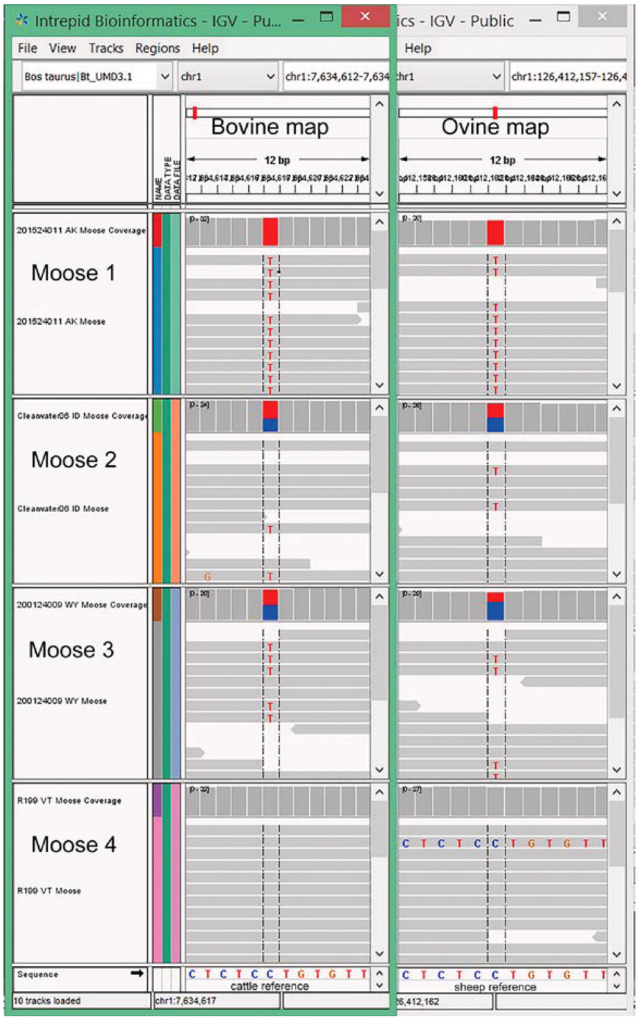

Figure 3. Computer screen images of a highly informative moose SNP. Overlapping computer screen images of moose WGS data aligned to bovine and ovine reference genomes, respectively, showing a highly informative SNP. Screenshots from IGV software showing one of 317 moose SNPs with a 0.5 MAF, both homozygous genotypes present, and aligned to genomic regions conserved in all three species. The bovine UMD3.1 and ovine Oar_v3.1 map positions are chr1:7,634,617 and chr1: 126,412,162, respectively.

The alignment coordinates of the 1,325 highly informative SNPs were analyzed for genome-wide distribution patterns that may indicate ascertainment biases caused by the variant selection. Overall, the distribution of SNP sites in the sets with 773, 552, and the 317 intersecting markers, appeared to be widespread in the cattle and sheep genomes and generally appropriate for genome-wide estimates (Figure 4). However, some SNP clustering was observed as the set of 317 had a mean and median spacing of 5.3 and 2.1 Mb, respectively (Supplementary Figure S1A). To facilitate SNP genotype assay design, the clustered SNPs were manually grouped into 216 bins with a mean size of approximately 8.1 Mb (median 5.9 Mb, Supplementary Figure S1B). Thus, SNP assay designs could be directed to each bin, with the option to use any SNP from that bin for multiplex assay design (Table S3).

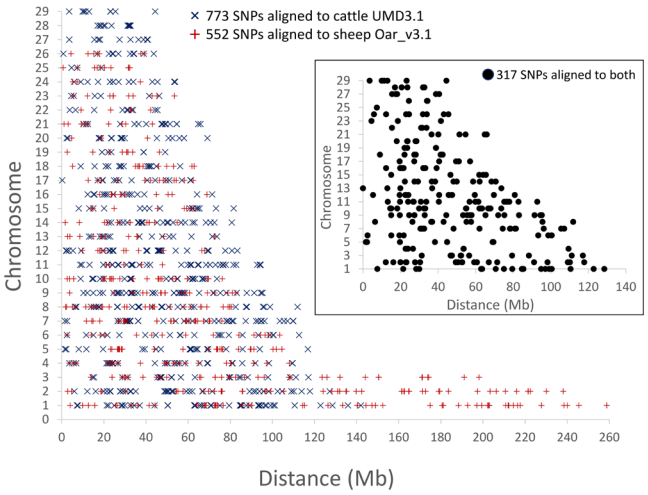

Figure 4. Genome-wide distribution of highly informative moose SNPs. The distribution of moose SNPs with 0.5 MAF relative to the cattle and sheep chromosomal locations (see Table S1 and Table S2 for marker details). The inset shows the chromosomal map positions with the cattle UMD3.1 reference assembly.

Genotype analysis for the 773 and 552 moose SNPs derived from the cattle and sheep alignments, respectively, showed that each moose had approximately the same proportion of opposing homozygous genotypes (Figure 5A, Supplementary Table S4 and Supplementary Table S5). However, there were significant differences in the ratio of heterozygous genotypes to homozygous genotypes (Figure 5B). The Alaskan moose had the most favorable average heterozygosity ratio (1.26), followed by the moose from Wyoming and Idaho (1.10 and 1.06, respectively), and the Vermont moose (0.68). Note that the numerical value of the ratios calculated from these SNP is likely an underestimate of the within-animal genome-wide heterozygosity, because there may be ascertainment bias resulting from targeting of SNP discovery to genomic regions conserved between three species. The SNP allele sharing between each of the four moose was analyzed with the sets of 773 and 552 markers to obtain a genome-wide measurement of their relatedness. This was possible because the method for selecting each of these SNPs was not dependent on which two of the four moose were heterozygous. The pair of moose from Alaska and Idaho had the highest proportion of shared alleles (0.430 and 0.397), while the Alaska and Vermont pair had the lowest (0.255 and 0.279, Table 3). Together, the genotype results with these sets of 773 and 552 SNPs indicate that there was a west-to-east pattern of decreasing genetic diversity in the four moose used in this study.

The combined set of 1,008 highly informative moose SNPs were also evaluated for their relative proximity to genes in the annotated reference assemblies of cattle and sheep. In the sets of 773 and 552 moose SNPs, 256 and 181 were present within genes, respectively (Table S6 and Table S7). Some genes contained more than one polymorphism, and thus, there were 221 and 178 total cattle and sheep genes, respectively, with highly informative SNPs. Of these genes with moose SNPs, 84 were identified in both cattle

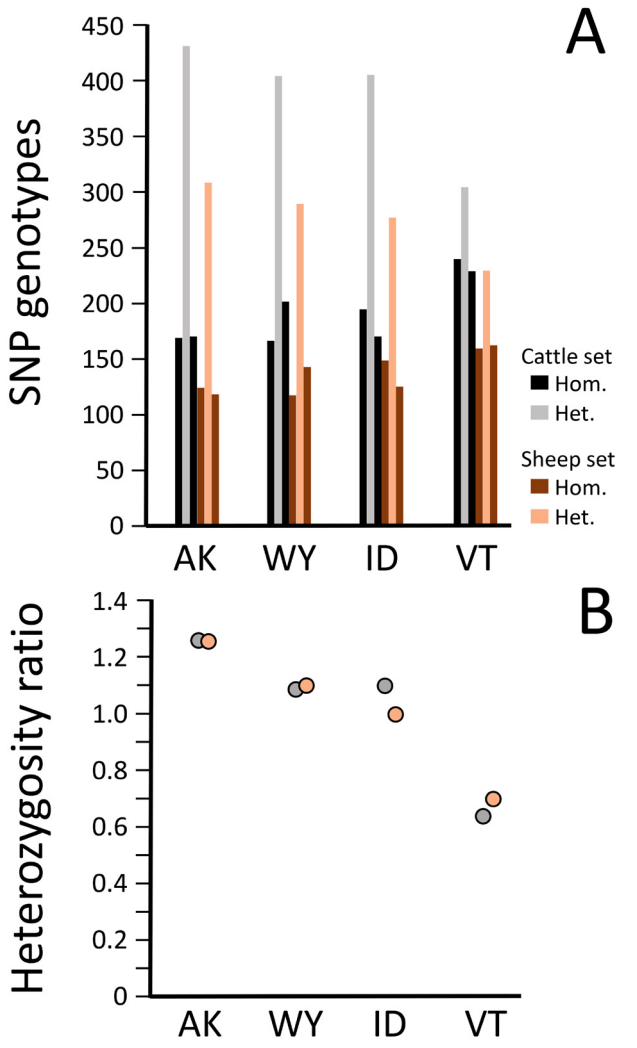

Figure 5. Heterozygosity analysis of moose genomes. The ratio of heterozygous to homozygous genotypes from four moose was evaluated with sets of 773 and 552 markers SNPs with 0.5 MAF (see Table S1 and Table S2 for marker details). (**A**) Genotype counts for each of the four animals with the 773 moose SNPs identified in the alignment to cattle (Table S1) and the 552 moose SNPs identified in the alignment to sheep (Table S2). (**B**) The heterozygosity ratios calculated for each of the four animals from the 773 SNP set (grey circles); and the 552 SNP set (tan circles). The ratio consisted of the number of heterozygous sites divided by the combined number of homozygous sites.

and sheep alignments. In addition, there were a number of informative SNPs in noteworthy genes that did not pass the read depth and quality score filters. For example, the prion gene (*PRNP*) affects susceptibility to spongiform encephalopathies such as chronic wasting disease in cervids. By manually viewing the *PRNP* coding sequence with IGV software, a coding SNP with 0.5 MAF and both homozygotes present was identified (M217I, Table 4). Thus, the publicly searchable and viewable moose WGS presented here represents a novel genomics resource that may facilitate candidate gene-based research in this species.

Discussion

We sequenced four moose from regions that span the United States, to approximately 19-fold genome coverage, and aligned them to the cattle and sheep reference genomes. Approximately 10% of moose sequences were aligned and used to identify more than 40 k moose SNPs in this cross-species approach. The relatively low alignment rate may be a reflection of the 27 million year average molecular divergence time between moose and non-cervid members of the Pecora infraorder[3]. In spite of the alignment rate, 1,008 highly informative moose SNPs were identified for future use in developing DNA-based genetic tests to support forensic and wildlife conservation activities. These 1,008 moose SNPs were derived from the intersection of two overlapping sets aligned to cattle (773 SNPs) and sheep (552 SNPs) reference genome assemblies. The 1,008 moose SNPs were refined to a minimal subset of 317 moose SNPs found in the most highly conserved genome regions. All of these markers are publicly available and ready for validation on a variety of SNP genotyping technology platforms. An important first step in evaluating these SNPs will be characterizing their MAFs in wild populations of North American moose. The online whole genome moose sequences, together with reference genotypes (Supplementary Table S4 and Supplementary Table S5) and DNA from these four moose, provide the opportunity for immediate design, testing, and validation of these candidate parentage SNPs.

Genotype information from the 1,008 moose SNPs was useful for measuring genome-wide differences in DNA sequence diversity among the four individuals. Measurements of heterozygosity and allele sharing showed that the Alaskan moose was the most diverse, the Vermont moose was the least, with the moose from Idaho and Wyoming being intermediate. This is consistent with a species that crossed the Bering Land Bridge into Alaska and radiated outward from west to east across North America. SNPs

Table 3. Proportion of heterozygous sites shared between pairs of moose from Alaska (AK), Idaho (ID), Wyoming (WY), and Vermont (VT), USA.

	Proportion							
	SNPs used from cattle alignment (n = 773)				SNPs used from sheep alignment (n = 552)			
Source	AK	ID	WY	VT	AK	ID	WY	VT
AK	1.000	-	-	-	1.000	-	-	-
ID	0.430	1.000	-	-	0.397	1.000	-	-
WY	0.378	0.323	1.000	-	0.391	0.317	1.000	-
VT	0.255	0.268	0.325	1.000	0.279	0.281	0.320	1.000

Table 4. Moose gene variants identified by viewing selected genes in IGV.

Bovine UMD3.1				Moose genotype[a]				Allele frequency	
BTA	Position	Gene	Feature	AK	WY	VT	ID	A1	A2
5	45,830,842[b]	*IFNG*	Intron 1	G[c]	S	C	S	0.50	0.50
13	47,415,079[b]	*PRNP*	CDS, M217I[d]	K	K	G	T	0.50	0.50
2	6,587,679	*ANKAR*	Intron 2	W	A	W	W	0.63	0.38
2	6,219,619	*MSTN*	Exon 5	G	C	S	C	0.63	0.38
3	22,970,184	*PDE4DIP*	Exon 16	T	T	C	Y	0.63	0.38
5	66,599,950	*IGF1*	CDS, I27V[e]	C	T	Y	T	0.63	0.38
7	22,883,135	*ICAM1*	Intron 1	Y	Y	C	Y	0.63	0.38
8	78,244,133	*UBQLN1*	Exon 3	Y	Y	T	T	0.75	0.25
16	27,304,417	*TLR5*	Exon 2	C	G	G	S	0.63	0.38
26	20,694,701	*DNMBP*	Exon 17	W	W	A	W	0.63	0.38

Key:

[a]Based on samples from four individuals sourced from Alaska (AK), Wyoming (WY), Vermont (VT), and Idaho (ID), USA.

[b]Highly-informative moose parentage SNPs with 0.5 MAF and both homozygous genotypes present among the four moose.

[c]Homozygotes are denoted with the one-letter nucleotide code. Heterozygotes are denoted with IUPAC/IUBMB ambiguity codes: R = a/g, Y = c/t, M = a/c, K = g/t, S = c/g, W = a/t[40].

[d]The *PRNP* codon number 217 refers to the number system in cattle. In moose, this codon is at position 209.

[e]The *IGF1* codon is in exon 2 and the numbering for codon 27 is the same in cattle as in moose.

have been previously used to estimate genome diversity in other species with low genetic diversity like the European bison (*B. bonasus*)[41] and the Tasmanian devil (*S. harrisii*)[25]. A caveat with our results is the overall heterozygosity of each moose may be underestimated due to ascertainment bias for highly informative SNPs in highly conserved genomic regions. In other words, variation in conserved moose genome regions may occur at a lower rate than that in non-conserved regions. In spite of this potential ascertainment bias, the results suggest that combinations of these markers may be useful in detecting population structure.

An important unanswered question is: how informative will these SNPs be in moose populations? Population-wide data to address this question will require development and application of genotyping assays, and assembly of pertinent samples for testing, which was beyond the scope and resources of the present report. The data presented here, which identify polymorphisms with alternate homozygous genotypes in a limited sample of only four individuals, suggest that the SNP selected represent variation that existed prior to arrival of moose in North America.

Conclusions

These moose SNPs and associated sequence information are available for use without restriction, and provide a basis for developing commercial SNP-based "parentage" SNP DNA tests for validation in North American moose populations.

Data availability

FASTQ files for the four moose combined are available in NCBI SRA, with contiguous accession numbers SRX3218250 - SRX3218281.

The SRA accession numbers for each individual are:

SRX3218264 - SRX3218271, Alaska moose HM2013;

SRX3218254 - SRX3218259 and SRX3218262 - SRX3218263, Wyoming moose JC2001; SRX3218250 - SRX3218253 and SRX3218272 - SRX3218275, Vermont moose R199; and SRX3218260 - SRX3218261 and SRX3218276 - SRX3218281, Idaho moose Clearwater06.

The data are part of NCBI BioProject Accession PRJNA325061.

In addition, access to the aligned sequences is available via the USDA: http://www.ars.usda.gov/Research/docs.htm?docid=25590 (moose aligned to cattle), and http://www.ars.usda.gov/Research/docs.htm?docid=25712 (moose aligned to sheep).

Download access to the BAM files is available at the Intrepid Bioinformatics sites: http://server1.intrepidbio.com/FeatureBrowser/customlist/record?listid=7919250313 (moose aligned to cattle), and http://server1.intrepidbio.com/FeatureBrowser/customlist/record?listid=7919250315 (moose aligned to sheep).

Competing interests

No competing interests were disclosed. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. The USDA is an equal opportunity provider and employer.

Grant information

Funding for this research was provided by the USDA, ARS appropriated projects 5438-32000-033-00D and 5438-31320-012-00D, the USDA National Institute of Food and Agriculture, McIntire-Stennis project 1002300, the University of Vermont College of Agriculture and Life Sciences, the College of Agricultural and Life Sciences at the University of Idaho, with the resources of the University of Louisville's research computing group and the Cardinal Research Cluster.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We thank J. Carnahan of USMARC for her outstanding technical assistance; H. Moertl (Alaska), B. Carnahan (Wyoming), C. Alexander of the Vermont Fish and Wildlife Department, and Y. Doe and D. Davenport (Idaho) for providing, or assisting with acquisition of moose samples. This work was funded, in part, by the University of Vermont College of Agriculture and Life Sciences, the College of Agricultural and Life Sciences at the University of Idaho, with the resources of the University of Louisville's research computing group and the Cardinal Research Cluster, and we thank H. Simrall of U. Louisville for his assistance.

Supplementary materials

Table S1. List of 773 moose parentage SNPs aligned to bovine reference assembly UMD3.1.

Click here to access the data.

Table S2. List of 552 moose parentage SNPs aligned to ovine reference assembly Oar_v3.1.

Click here to access the data.

Table S3. List of 317 candidate moose parentage SNPs grouped into 216 bins for use in multiplex assay design.

Click here to access the data.

Table S4. Genotypes for 773 moose SNPs aligned to the bovine reference assembly UMD3.1.

Click here to access the data.

Table S5. Genotypes for 552 moose SNPs aligned to the ovine reference assembly Oar_v3.1.

Click here to access the data.

Table S6. List of 256 moose parentage SNPs occurring within genes annotated in bovine reference assembly UMD3.1.

Click here to access the data.

Table S7. List of 181 moose parentage SNPs occurring within genes annotated in ovine reference assembly Oar_v3.1.

Click here to access the data.

Figure S1. Distributions for 317 moose parentage SNPs and their 216 marker groups.

Click here to access the data.

File S1. VCF file with 1,095,371 moose-specific variants aligned to the cattle UMD3.1 reference assembly.

Click here to access the data.

File S2. VCF file with 813,006 moose-specific variants aligned to the sheep Oar_v3.1 reference assembly.

Click here to access the data.

File S3. VCF file with 46,005 moose SNPs aligned to the cattle UMD3.1 reference assembly.

Click here to access the data.

File S4. VCF file with 36,934 moose SNPs aligned to the sheep Oar_v3.1 reference assembly.

Click here to access the data.

References

1. Franzmann AW: **Alces alces.** *Mammalian Species.* 1981; **154**: 1–7.
Publisher Full Text
2. Karns PD: **Population distribution, density and trends.** In: Franzmann AW, Schwartz CC, editors. *Ecology and management of the North American moose.* Washington, D.C., USA: Smithsonian Institution Press, 1998; 125–39.
3. Kumar S, Stecher G, Suleski M, *et al.*: **TimeTree: A Resource for Timelines, Timetrees, and Divergence Times.** *Mol Biol Evol.* 2017; **34**(7): 1812–9.
PubMed Abstract | Publisher Full Text
4. Timmermann HR, Rogers AR: **Moose: competing and complementary values.** *Alces.* 2005; **41**: 85–120.
Reference Source
5. Lenarz MS, Fieberg J, Schrage MW, *et al.*: **Living on the edge: viability of moose in northeastern Minnesota.** *J Wildl Manage.* 2010; **74**(5): 1013–23.
Publisher Full Text
6. Murray DL, Cox EW, Ballard WB, *et al.*: **Pathogens, nutritional deficiency, and climate influences on a declining moose population.** *Wildlife Monographs.* 2006; **166**(1): 1–30.
Publisher Full Text
7. Allendorf FW, Hohenlohe PA, Luikart G: **Genomics and the future of conservation genetics.** *Nat Rev Genet.* 2010; **11**(10): 697–709.
PubMed Abstract | Publisher Full Text
8. Frankham R, Ballou JD, Briscoe DA: **Introduction to conservation genetics.** 2nd ed. Cambridge, UK; New York: Cambridge University Press, 2010; xxiii: 618.
Reference Source
9. Luikart G, England PR, Tallmon D, *et al.*: **The power and promise of population genomics: from genotyping to genome typing.** *Nat Rev Genet.* 2003; **4**(12): 981–94.
PubMed Abstract | Publisher Full Text
10. Creel S, Creel NM: **The African wild dog behavior, ecology, and conservation.** Princeton, New Jersey, USA.: Princeton University Press; 2002.
Reference Source
11. Kitchen AM, Gese EM, Waits LP, *et al.*: **Multiple breeding strategies in the swift fox, *Vulpes velox*.** *Anim Behav.* 2006; **71**(5): 1029–38.
Publisher Full Text
12. Weston Glenn JL, Civitello DJ, Lance SL: **Multiple paternity and kinship in the gray fox (*Urocyon cinereoargenteus*).** *Mamm Biol.* 2009; **74**(5): 394–402.
Publisher Full Text
13. Arif IA, Khan HA, Bahkali AH, *et al.*: **DNA marker technology for wildlife conservation.** *Saudi J Biol Sci.* 2011; **18**(3): 219–25.
PubMed Abstract | Publisher Full Text | Free Full Text
14. DeYoung RW, Honeycutt RL: **The molecular toolbox: genetic techniques in wildlife ecology and management.** *J Wildl Manage.* 2005; **69**(4): 1362–84.
Publisher Full Text
15. Schlotterer C: **The evolution of molecular markers--just a matter of fashion?** *Nat Rev Genet.* 2004; **5**(1): 63–9.
PubMed Abstract | Publisher Full Text
16. Vignal A, Milan D, SanCristobal M, *et al.*: **A review on SNP and other types of molecular markers and their use in animal genetics.** *Genet Sel Evol.* 2002; **34**(3): 275–305.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Broders HG, Mahoney SP, Montevecchi WA, *et al.*: **Population genetic structure and the effect of founder events on the genetic variability of moose, *Alces alces*, in Canada.** *Mol Ecol.* 1999; **8**(8): 1309–15.
PubMed Abstract | Publisher Full Text
18. Hundertmark KJ: **Reduced Genetic Diversity in Two Introduced and Isolated Moose Populations in Alaska.** *Alces.* 2009; **45**: 137–42.
Reference Source
19. Sattler RL, Willoughby JR, Swanson BJ: **Decline of heterozygosity in a large but isolated population: a 45-year examination of moose genetic diversity on Isle Royale.** *PeerJ.* 2017; **5**: e3584.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Schmidt JI, Hundertmark KJ, Bowyer RT, *et al.*: **Population Structure and Genetic Diversity of Moose in Alaska.** *J Hered.* 2009; **100**(2): 170–80.
PubMed Abstract | Publisher Full Text
21. Wilson GA, Strobeck C, Wu L, *et al.*: **Characterization of microsatellite loci in caribou *Rangifer tarandus*, and their use in other artiodactyls.** *Mol Ecol.* 1997; **6**(7): 697–9.
PubMed Abstract | Publisher Full Text
22. Wilson PJ, Grewal S, Rodgers A, *et al.*: **Genetic variation and population structure of moose *(Alces alces)* at neutral and functional DNA loci.** *Can J Zool.* 2003; **81**(4): 670–83.
Publisher Full Text
23. Sachidanandam R, Weissman D, Schmidt SC, *et al.*: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.** *Nature.* 2001; **409**(6822): 928–33.
PubMed Abstract | Publisher Full Text
24. Kleinman-Ruiz D, Martinez-Cruz B, Soriano L, *et al.*: **Novel efficient genome-wide SNP panels for the conservation of the highly endangered Iberian lynx.** *BMC Genomics.* 2017; **18**(1): 556.
PubMed Abstract | Publisher Full Text | Free Full Text
25. Wright B, Morris K, Grueber CE, *et al.*: **Development of a SNP-based assay for measuring genetic diversity in the Tasmanian devil insurance population.** *BMC Genomics.* 2015; **16**: 791.
PubMed Abstract | Publisher Full Text | Free Full Text
26. Norman AJ, Street NR, Spong G: ***De novo* SNP discovery in the Scandinavian brown bear (*Ursus arctos*).** *PLoS One.* 2013; **8**(11): e81012.
PubMed Abstract | Publisher Full Text | Free Full Text
27. Hundertmark KJ, Shields GF, Udina IG, *et al.*: **Mitochondrial phylogeography of moose (*Alces alces*): late pleistocene divergence and population expansion.** *Mol Phylogenet Evol.* 2002; **22**(3): 375–87.
PubMed Abstract | Publisher Full Text
28. Hedrick PW: **Conservation genetics and North American bison (*Bison bison*).** *J Hered.* 2009; **100**(4): 411–20.
PubMed Abstract | Publisher Full Text
29. Kalbfleisch T, Heaton MP: **Mapping whole genome shotgun sequence and variant calling in mammalian species without their reference genomes [version 2; referees: 2 approved].** *F1000Res.* 2013; **2**: 244.
PubMed Abstract | Publisher Full Text | Free Full Text
30. Hundertmark KJ, Shields F, Bowyer RT, *et al.*: **Genetic relationships deduced from cytochrome-*b* sequences among moose.** *Alces.* 2002; **38**: 113–22.
Reference Source
31. Heaton PM, Grosse MW, Kappes MS, *et al.*: **Estimation of DNA sequence diversity in bovine cytokine genes.** *Mamm Genome.* 2001; **12**(1): 32–7.
PubMed Abstract | Publisher Full Text
32. Zimin AV, Delcher AL, Florea L, *et al.*: **A whole-genome assembly of the domestic cow, *Bos taurus*.** *Genome Biol.* 2009; **10**(4): R42.
PubMed Abstract | Publisher Full Text | Free Full Text
33. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2010; **26**(5): 589–95.
PubMed Abstract | Publisher Full Text | Free Full Text
34. Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–9.
PubMed Abstract | Publisher Full Text | Free Full Text
35. McKenna A, Hanna M, Banks E, *et al.*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res.* 2010; **20**(9): 1297–303.
PubMed Abstract | Publisher Full Text | Free Full Text
36. Robinson JT, Thorvaldsdóttir H, Winckler W, *et al.*: **Integrative genomics viewer.** *Nat Biotechnol.* 2011; **29**(1): 24–6.
PubMed Abstract | Publisher Full Text | Free Full Text
37. Thorvaldsdóttir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** *Brief Bioinform.* 2013; **14**(2): 178–92.
PubMed Abstract | Publisher Full Text | Free Full Text
38. Heaton MP, Smith TP, Carnahan JK, *et al.*: **Using diverse U.S. beef cattle genomes to identify missense mutations in *EPAS1*, a gene associated with high-altitude pulmonary hypertension [version 1; referees: 2 approved].** *F1000Res.* 2016; **5**: 2003.
PubMed Abstract | Publisher Full Text | Free Full Text
39. Heaton MP, Smith TP, Freking BA, *et al.*: **Using sheep genomes from diverse U.S. breeds to identify missense variants in genes affecting fecundity [version 1; referees: 2 approved].** *F1000Res.* 2017; **6**: 1303.
PubMed Abstract | Publisher Full Text | Free Full Text
40. NC-IUB: **Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Nomenclature Committee of the International Union of Biochemistry (NC-IUB).** *Proc Nat Acad Sci U S A.* 1986; **83**(1): 4–8.
PubMed Abstract | Free Full Text
41. Tokarska M, Marshall T, Kowalczyk R, *et al.*: **Effectiveness of microsatellite and SNP markers for parentage and identity analysis in species with low genetic diversity: the case of European bison.** *Heredity (Edinb).* 2009; **103**(4): 326–32.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Referee Status: ? ✓ ✓

Version 1

Referee Report 06 February 2018

doi:10.5256/f1000research.14659.r29739

✓ **Kris J. Hundertmark**
Department of Biology and Wildlife, Institute of Arctic Biology, University of Alaska Fairbanks, Fairbanks, AK, USA

This study presents SNP discovery and variation of moose in North America, which are very valuable data. First, moose are an understudied species and their genetics can tell us a lot about the phylogeography of species after the LGM. As the authors state, moose in North America have generally low neutral nuclear diversity, and are less diverse than their European counterparts (information from Siberia, if it exists, has not been published in English). They are also less diverse than many northern species and many species of cervids, such as caribou, white-tailed deer, and elk. Hypotheses have been proposed for these observations and have been tested with mtDNA, but only recently have nuclear markers been used to study phylogeography in European moose. No such tests have been conducted in North America. This study provides new tools to spur moose research, although I agree with Reviewer 2 that microsatellites and mtDNA still have their uses.

The inclusion of 4 individuals in this study is good, as is the broad geographic range of the samples across North America. The methods are appropriate, although as the first reviewer points out, RADSeq could have been used in SNP discovery and has some advantages over the methods used, but I believe that this report on moose is secondary to the authors' overall goals and we are fortunate that Kalbfleish et al. decided to pursue this and share their findings. The comparison to reference genomes, although from bovids not very closely related to moose, yielded some advantages, including identifying SNPs in important functional genes, such as the PRNP locus. Nonetheless, a white-tailed deer genome was made public in summer 2017 and although these authors may not be willing to start over by comparison of their data to a very similar genome (same subfamily and same number of chromosomes) the prospect exists and should be pursued. I can see the point of Reviewer 2 concerning Fig. 4 and assumed synteny but I believe I got the intended message from the figure as is. But I think mentioning somewhere that North American moose have 34 pairs of autosomes when discussing the success of mapping SNPs to the reference genomes would be appropriate.

I find the last sentence interesting in that the authors believe the SNP variants they found existed prior to moose entering North America. Considering that event likely happened within the last 15,000 years that is a very reasonable statement, and may cause some to think that SNPs may contain little information about geographic variation in moose and all that goes with it. Given the morphological and behavioral differences between, say, Alaskan moose and those in the eastern continent, however, it is obvious that moose have evolved rapidly in that time, despite limited genetic diversity due to Pleistocene bottlenecks and founder effects. How that translates to current SNP diversity and what the latter may be able to tell us about moose are exciting questions to be asked, for which this manuscript sets the stage.

Minor comments:

2nd paragraph of Introduction: should be “among individuals” not “between individuals.”

Third paragraph of Introduction: should be “a SNP-based approach” not “an SNP-based approach.”

Last paragraph of Introduction, last sentence: should it be “has been made freely available” rather than “was made freely available?”

Last paragraph of WGS Production …, last words of 2nd-to-last sentence: Should be “previously.” not “in previously.”

Discussion, last sentence: should this be “SNPs” instead of SNP?”

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Moose evolution, population and spatial genetics of moose and other large mammals

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 30 January 2018

doi:10.5256/f1000research.14659.r30045

Paul Stothard
Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB, Canada

This well-written manuscript describes the discovery and characterization of single nucleotide polymorphisms (SNPs) in moose. These SNPs will be valuable for future research and conservation efforts, as they can be used for animal identification, assigning parentage, and estimating intra- and inter-population genetic variability. The SNP discovery and genotyping are done using well-established and clearly described approaches, and care has been taken to avoid false positive SNPs (requiring that both homozygous genotypes be observed for example). The raw sequencing data is available through the SRA database, while the aligned reads (BAM) files are available via the USDA and Intrepid Bioinformatics sites. The final filtered variants are provided with flanking sequence in the supplementary materials. Larger, less-filtered collections of variants are provided in the form of VCF files, again in the supplementary materials. The clear, detailed manuscript and the raw data and progressively filtered results will make is easy for others to reproduce or make use of the results of this work.

Minor comments

1. In the last paragraph of the introduction the authors note the challenge of creating a whole genome assembly for use in SNP discovery, and that the use of an existing reference genome from a related species can be an effective alternative. In the discussion section the authors mention that the cross-species mapping approach, however, has the drawback of targeting conserved regions of the genome. I am curious as to why the authors chose not to employ a technique like RADseq for SNP discovery, as it does not depend on the availability of a reference genome and would allow them to better assess minor allele frequency, through the inclusion of more individuals. Also, RADseq would be equally effective at targeting conserved and non-conserved regions. The reasons for not using RADseq (or its potential value in future studies) could be addressed in the introduction.

2. The figure legends for Figure 2 and Figure 3 could be expanded slightly to explain to readers not familiar with IGV which elements represent reads, coverage, reference sequence. Also, in Figure 3 it isn't clear to me why the sequence of one read is shown (moose 4, second read from top, aligned to the sheep reference).

3. In the Methods section "fastq" is written as "FASTQ" and "fastq".

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Genetics, genomics, bioinformatics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 30 January 2018

doi:10.5256/f1000research.14659.r30007

? **Joshua M. Miller** [1,2]

[1] Université du Québec à Montréal, Montreal, Quebec, Canada

[2] University of Alberta, Edmonton, Canada

In this work Kalbfleisch *et al.* develop a novel set of SNP loci for moose. To do so the authors generated genomic sequence data from 4 moose across their range in the United States, aligned the reads to both cow and sheep reference genomes, and then applied a series of filters to select a subset of loci in highly conserved regions. The resulting set of loci showed a gradient of diversity decreasing from west to east, consistent with hypothesized colonization. The authors state that these loci will serve as a resource for future management and conservation applications. I think that this study has laudable goals and adds a valuable resource for moose conservation. However, there are some issues the analyses and presentation that need to be clarified.

Overall comments

- The alignment and filtering procedures seem overly restrictive. Authors note that only ~10% of their reads aligned to the cow and sheep reference genomes (not surprising given the levels of divergence among the species) meaning that 90% of the data was essentially thrown away. Why not start with a *de novo* assembly of the moose sequence?
    - Along these lines there are several recent papers that detail a hybrid procedure for genome construction that begins with *de novo* assembly and then apply cross-species alignment for scaffolding, e.g. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1911-6
- If part of the goals is to eventually use these loci for “population genetic applications” (e.g. looking for population structure and assigning individuals to those populations) it strikes me that selecting loci that are in such heavily conserved regions may result in biased estimates. Can the authors comment on this? Doesn’t this procedure result in a lot of “moose specific” variants being missed?
- I can understand that it is not the objective of this paper to create a draft genome sequence for the moose (though the authors note in several places that this would be useful, and the data generated here seem appropriate to attempt this), but then more justification as to why this was not attempted needs to be given.
- Did the authors check and make sure that the “highly conserved genomic regions” are not in repetitive elements? A blast search of these regions should do the trick.

Specific comments

- In the sentence in the introduction starting with “DNA technology developments…” saying that SNPs have “replaced” microsatellites and mitochondrial DNA is a gross overstatement. There are many studies still using these markers and there are many applications where these markers may be preferable. Would be better to say something along the lines that SNPs have gained in use.

- As currently written, the sentence in the introduction starting with “Moreover, panels pf SNPs broaden…” is incorrect. Microsatellites and mitochondrial DNA also provide means of measuring inbreeding and relatedness among individuals. This should be rephrased to state that SNPs may be better at these estimates, likely due to the factors listed in the previous sentence (abundance, accuracy, etc.)
- I would couch the statement that previous estimates of low genetic diversity in moose will make it hard to discover SNPs. This may be true, but it has not been tested and estimates of diversity from 5-11 microsatellites likely will not reflect genome-wide patterns.
- Remove “in the prevailing theory,”
- I find it striking that in the final paragraph of the introduction there is no mention of the plethora of genotype-by-sequencing approaches that are used for SNP discovery and population genetic analyses but do not require a reference genome, e.g Peterson *et al.* (2012)[1]. This should be addressed.
- When you state the sub-species names in the Methods refer readers to Table 1 so that they know which occurs where relative to your samples.
- Which program or programs were used to filter the raw reads?
- State what “UMD3.1” and “Oar_v3.1” are when you first introduce them in the methods
- Is binning relative to the cow genome actually informative for moose? Are the karyotypes comparable? How many chromosomes does the moose have?
- Along this same line, I think having the variants locations for both UMD3.1 and Oar_v3.1 plotted on top of each other in Figure 4 is misleading. It suggests that the locations are syntenic, which is not true. For instance, the plot implies that no variants were discovered on chromosomes 27, 28, and 29 in the sheep genome, when really this is impossible as sheep only have 27 autosomes. I would suggest re-doing this figure with separate panels for each figure, though see the caveat above.
- In Figure 5A why are there two bars for each of the “Hom” counts but only one for the “Het”?

References

1. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE: Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species.*PLoS One*. 2012; **7** (5): e37135 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.