G3
Genes | Genomes | Genetics

Genome-Wide SNP Discovery and Analysis of Genetic Diversity in Farmed Sika Deer (*Cervus nippon*) in Northeast China Using Double-Digest Restriction Site-Associated DNA Sequencing

Hengxing Ba,* Boyin Jia,* Guiwu Wang,* Yifeng Yang,* Gilead Kedem,† and Chunyi Li*,1
*State Key Laboratory for Molecular Biology of Special Wild Economic Animals, Institute of Special Wild Economic Animals and Plants, Chinese Academy of Agricultural Sciences, Changchun 130112, China and †GENEWIZ Inc., South Plainfield, New Jersey 07080

ABSTRACT Sika deer are an economically valuable species owing to their use in traditional Chinese medicine, particularly their velvet antlers. Sika deer in northeast China are mostly farmed in enclosure. Therefore, genetic management of farmed sika deer would benefit from detailed knowledge of their genetic diversity. In this study, we generated over 1.45 billion high-quality paired-end reads (288 Gbp) across 42 unrelated individuals using double-digest restriction site-associated DNA sequencing (ddRAD-seq). A total of 96,188 (29.63%) putative biallelic SNP loci were identified with an average sequencing depth of 23×. Based on the analysis, we found that the majority of the loci had a deficit of heterozygotes ($F_{IS} > 0$) and low values of H_{obs}, which could be due to inbreeding and Wahlund effects. We also developed a collection of high-quality SNP probes that will likely be useful in a variety of applications in genotyping for cervid species in the future.

KEYWORDS
sika deer
ddRAD-seq
SNP discovery
genetic diversity
STACKS
GenPred
Shared Data Resources
Genomic Selection

Sika deer are an economically valuable species, with some body parts, such as antler, blood, penis, and placenta, used in traditional Chinese medicine. In order to continue the tradition, people have captured sika deer from the wild on a large scale in Jilin province in the northeastern part of China since the Qing dynasty (c. 1733) (Sheng and Ohtaishi 1993). These animals were gradually introduced into many other regions of China. The original captive population in northeast China is generally considered to be the source of the entire sika deer population on farms throughout China (Sheng and Ohtaishi 1993).

Almost all sika deer raised on farms in northeast China are used for antler (velvet antler) production (Sheng and Ohtaishi 1993; Wu *et al.* 2004). China is currently one of the largest producers of velvet antler in the world and sika deer farming in China is still thriving. Nonetheless, there is still concern about the deterioration of farm-bred sika deer (Wu and Zhang 2001). In recent decades, people have been trying to improve velvet antler yield through long-term improved management on farms (McCullough *et al.* 2009). Whether a healthy genetic diversity can be maintained is an open question. Therefore, any genetic management requires detailed knowledge of the genetic diversity.

The genetic diversity of the sika deer wild population has been investigated using two major sources of genomic variation. Mitochondrial genes, such as cytochrome b and the control region, have proven highly informative for investigations into the classification and phylogeny of sika deer with maternal haplogroups (Wu *et al.* 2004, 2005; Lu *et al.* 2006; Takiguchi *et al.* 2012; Krojerova-Prokesova *et al.* 2013; Ba *et al.* 2015, 2016). Autosomal microsatellites have been used extensively to estimate levels of genetic diversity (Tamate *et al.* 2000; Thevenon *et al.* 2004; Krojerova-Prokesova *et al.* 2013; He *et al.* 2014). Although highly polymorphic markers can be very informative, in particular when assessing recent demographic bottlenecks (endangered and captive-bred

doi: https://doi.org/10.1534/g3.117.300082
Manuscript received June 12, 2017; accepted for publication July 18, 2017; published Early Online July 27, 2017.

Supplemental material is available online at www.g3journal.org/lookup/suppl/doi:10.1534/g3.117.300082/-/DC1.
[1] Corresponding author: State Key Laboratory for Molecular Biology of Special Wild Economic Animals, Institute of Special Wild Economic Animals and Plants, Chinese Academy of Agricultural Sciences, Changchun 130112, China. E-mail: lichunyi1959@163.com

species), the previous data are not sufficient to provide a good understanding of the level of genetic diversity in the farmed population in China. This is the first study to report levels of genetic diversity within sika deer using single nucleotide polymorphism (SNP) markers on a global scale.

Their genomic abundance and amenability to cost-effective, high-throughput genotyping, has meant that SNPs are now the most widely used class of genetic markers. Use of genome-wide SNPs constitutes an important genetic tool for investigating genetic diversity in livestock animals (Groeneveld *et al.* 2010). Although genome-wide SNP discovery in white-tailed deer (Seabury *et al.* 2011) and hog deer (Wang *et al.* 2017) has been reported, development of this tool in deer has been hampered in part by the substantial duplicate regions in large parts of the deer genome, and by the lack of a reference genome sequence. Fortunately, double-digest restriction site-associated DNA sequencing (ddRAD-seq) technology (Peterson *et al.* 2012) can provide a flexible and inexpensive platform for the simultaneous discovery of tens of thousands of SNP markers in model and nonmodel organisms. While several SNP genotyping tools have been developed for SNP discovery, including GATK (DePristo *et al.* 2011), STACKS (Catchen *et al.* 2011, 2013), SAMtools (Li 2011) and RADtools (Baxter *et al.* 2011), STACKS is recommended for ddRAD-seq projects, as it uses a multinomial-based likelihood model to call SNPs, which incorporates a bounded SNP model.

In this study, we employed the ddRAD-seq technology to achieve the first genome-wide SNP discovery for farmed sika deer that are widely distributed in northeast China. The STACKS pipeline was applied for SNP calling. Using newly developed markers, we investigated the genetic diversity of farmed populations. We also developed a collection of SNP probes for genotyping in the future.

MATERIALS AND METHODS

Sample collection

A total of 42 unrelated animals were selected from eight farms (farm codes: AD, CB, DF, NM, SP, SY, XF, and XK) in northeast China, with six from NM and XK farms and five from the other six farms. Whole-blood samples were extracted from the jugular vein using EDTA vacuum tubes and were stored at −20 °C until DNA extraction. The genomic DNA extractions from whole blood were performed using a blood DNA kit (Qiagen) according to the manufacturer's instructions. Each DNA sample was evaluated by gel electrophoresis for the presence of high-molecular-weight DNA and then stored at −80°C until ddRAD-seq library construction.

ddRAD-seq library preparation and sequencing

The procedure was performed as described in previous studies (DaCosta and Sorenson 2014) with some modifications. First, the double-digest reactions were carried out in a volume of 25 µl containing ~0.8 µg of genomic DNA, 5U of PstI and MseI, and 1× buffer 3.1 (NEB). The reaction mixture was incubated at 37 °C for 2 hr and 65 °C for 30 min Amplification and sequencing adapters with a unique barcode (5 or 6 bp) were ligated on to the digested DNA. Each sample was then amplified via PCR in a 50 µl reaction volume, containing 70–100 ng of adaptor-ligated DNA fragments, and amplified with 22 cycles following the manufacturer's protocol. Samples were run on a 2% agarose gel, and DNA in the 300–450 bp size range (with indices and adaptors) was excised using a gel extraction kit (Qiagen). Each sample library was pooled in equal amounts and quantified using Agilent 2100 (Agilent Technologies) and real-time quantitative PCR, and then paired-end 101 bp sequencing was performed using the Illumina HiSeq4000 platform (BGI, Shenzhen, China).

Available sika deer genome reference

The available sika deer genome reference (SK-REF) included 387,939 contigs, comprising a total of 2.73 Gbp. All sequences were longer than 500 bp, with a N50 length of 38.4 kbp. The percentage of repeat sequences detected by Repeatmasker-4.0.6 (http://www.repeatmasker.org) was 44.14% and the GC level was 42.11% (H. Ba, unpublished data).

RAD-seq data analysis and SNP identification

Given the main quality score parameters (-c -q -r) of the process_radtags program in STACKS v.1.41 (Catchen *et al.* 2011, 2013), raw reads were filtered and separated by barcode. The filtered paired-end reads were uniquely mapped to SK-REF using BOWTIE v2.0 (Langmead and Salzberg 2012) (disallowing gaps and suppressing unpaired and discordant alignments), followed by SAMtools v1.2 (Li 2011) to convert to a BAM file. Putative orthologous loci were assembled for each animal using the BAM file as the input of the pstacks function with the minimum depth of coverage above 5. Two mismatches were allowed between animals when assembling the catalog of ddRAD loci using the cstacks program. Matches of individual ddRAD loci to the catalog were searched using the sstacks program. Loci that have low depth of coverage or high sequencing error will exhibit poor log likelihood scores that are highly negative. The rxstacks program (-conf_filter -conf_lim 0.75 -prune_haplo -lnl_lim -8.0) was used to make corrections to genotype calls in individuals based on data from a population-wide examination. Subsequently, the cstacks and sstacks programs were again run to rebuild and match to the catalog. The populations program was used to filter the dataset to contain SNPs found in at least 75% of individuals and with a minor allele frequency ≥0.05.

Estimates of genetic diversity within population

Observed heterozygosity (H_{obs}), expected heterozygosity (H_{exp}) and inbreeding coefficient of an individual relative to the subpopulation (F_{IS}) were calculated using the populations program in STACKS v.1.41. Deviation from the Hardy–Weinberg equilibrium (HWE) was assessed by performing a Fisher's exact test with the HardyWeinberg R package (Graffelman and Camarena 2008) for each SNP marker. The False discovery rate correction (Benjamini–Hochberg) was performed using the p.adjust program in the R package.

Population structure analysis

Bayesian clustering analysis implemented in STRUCTURE 2.3.4 (Pritchard *et al.* 2000) was used for estimating the number of populations/groups (k) represented by the dataset. Three iterations were run per K (number of populations) for $K = 2$ or $K = 3$ using an admixture model. Each run consisted of a burn-in of 100,000 MCMC steps, followed by 500,000 replications. Population structure was also examined by carrying out principal component analysis using SMARTPCA within EIGENSOFT (Patterson *et al.* 2006).

Development of SNP probes for genotyping assay

A custom script was run to develop high-quality SNP probes. In summary, sufficient 50 bp flanking sequences on either side of the SNPs were derived from the SK-REF genome and then were filtered according to four filtration criteria: (a) no repetitive sequences; (b) SNP-free within flanking sequence; (c) one of two alleles in accordance with the base in the reference; (d) flanking sequence on the SK-REF ≥50bp.

Data availability

Raw (adapter trimmed) Illumina ddRAD-seq sequences: NCBI Short Read Archive (project accession: SRP105008). Statistics describing

Table 1 Filtering results for STACKS pipeline

Quality Filtering Steps	Number of ddRAD Loci	%
Populating observed ddRAD loci for 42 animals	7,576,300	100.0
Removing ddRAD loci below the minimum depth of 5×	7,456,091	98.4
Removing ddRAD loci below the log likelihood threshold of −8	1,989,000	26.3
Removing ddRAD loci below the minimum constraint of 32 (75%) individuals	324,564	4.3
Putative biallelic SNP (MAF ≥0.05)	98,166	1.3

different properties of each sequenced individual and all high-quality SNP probes are available in the Supplemental Material, Table S1 and Table S2.

RESULTS

ddRAD sequencing

After trimming the barcodes and filtering low-quality bases, ddRAD sequencing generated a total of over 1.45 billion high-quality paired-end reads (288 Gbp) across the 42 animals. Of these clean reads, >0.93 billion (65.63%) were aligned uniquely to the SK-REF genome. As variation among animals is often observed in a pooled ddRAD sequencing library, the number of aligned reads per individual ranged from 0.55 to 5.3 million (1.01–9.78 Gbp, with an average of 4.1 Gbp) (Table S1).

SNP discovery

Using the Stacks pipeline, we initially obtained 7,576,300 candidates of the ddRAD loci from all individuals. Quality filtering (see Methods) reduced this to a total of 324,564 (4.28%), which were shared by >75% of the individuals, corresponding to genomic size of 0.032 Gbp (~1.1% of the genome) (Table 1). Within these 324,564 ddRAD loci of ~100 bp, we detected 96,188 (29.63%) putative biallelic SNPs with an average sequencing depth of 23×. For each individual, the number of SNPs varied from 56,388 (58.6%) to 89,845 (93.4%), and the sequencing depth ranged from 9× to 50×. The number of identified heterozygous

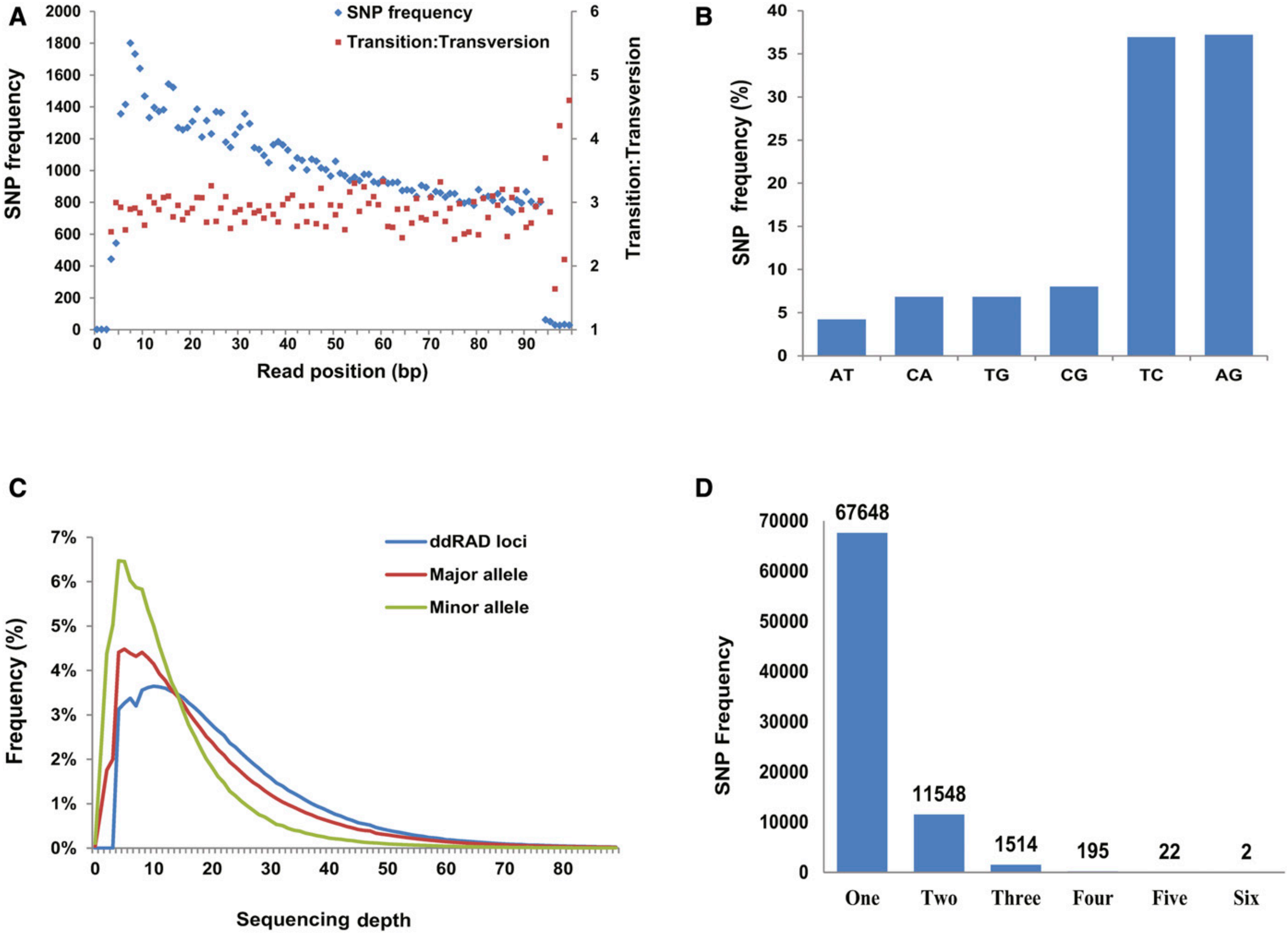

Figure 1 Evaluation of putative SNP quality. (A) Distribution of SNP frequency on ddRAD tag positions. Distribution of the identified Ts/Tv is also illustrated. (B) Distribution of six types of substitutions. (C) Distribution of SNP frequency in major allele, minor allele, and ddRAD loci against their sequencing depth. (D) Distribution of SNP frequency of the number of SNPs in ddRAD loci tags.

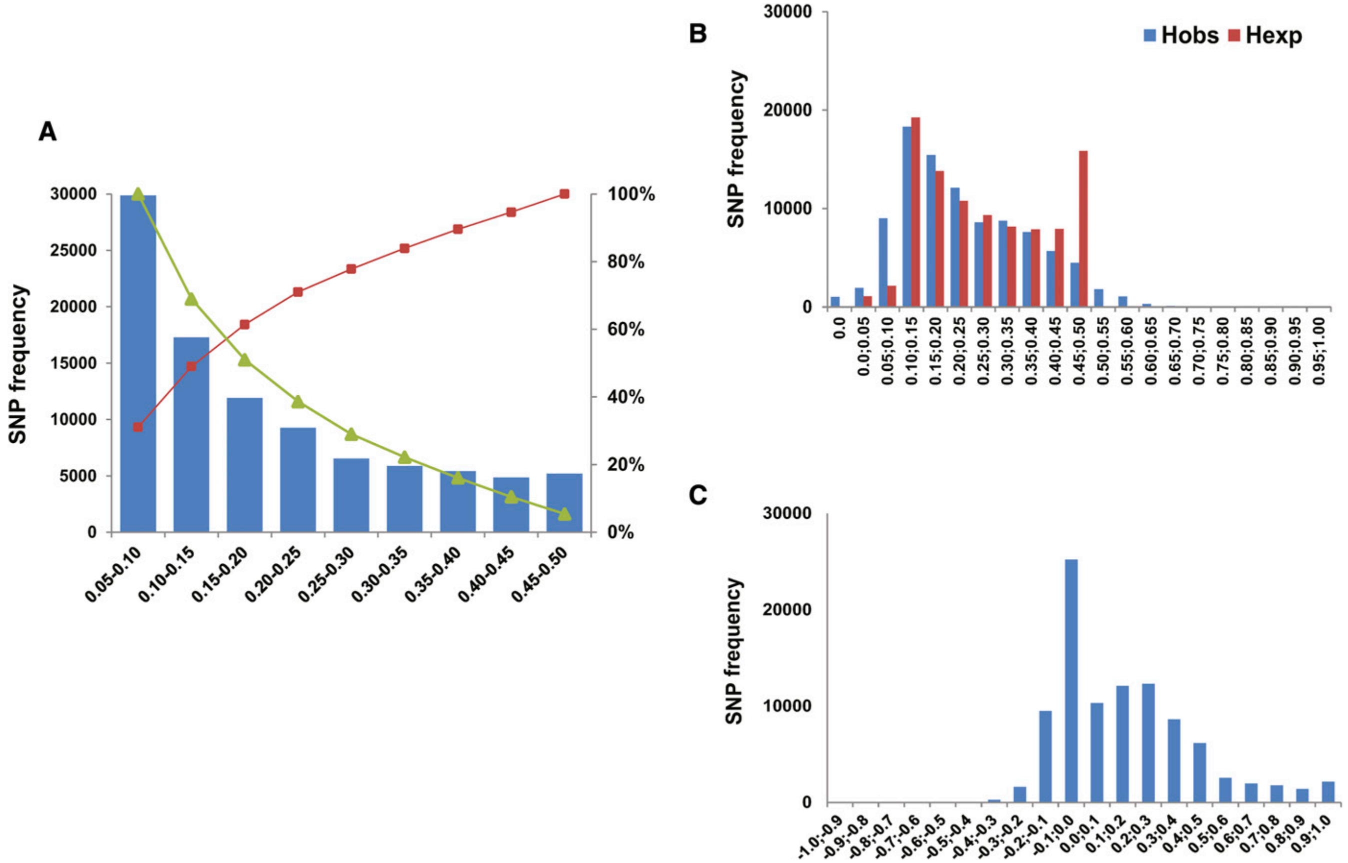

Figure 2 Population genetic parameters for putative SNP loci. (A) Frequency of SNPs across the MAF bins. The line depicts the cumulative density function across the MAF bins. (B) Frequency of SNPs across the heterozygosity bins. (C) Frequency of SNPs across the F_{IS} bins.

SNPs varied from 11,780 to 22,641 with an average of 19,078, corresponding to 0.38–0.71 heterozygous SNPs per kilobase pair with an average of 0.61 detected per individual animal (Table S1). With a read rate within individuals averaging 83%, this suggests that the rate could be ~0.74 per kbp.

Through evaluation of SNP frequency, substitution types and sequencing depth, the results showed that: (a) SNP frequency decreased from 5′ to 3′ end in all loci excluding both ends; however, transition/transversion (Ts/Tv) ratios were very stably distributed over the loci except for the last five positions at the 3′ end (Figure 1A); (b) frequency of A/T substitution is relatively less than other types of substitutions (Figure 1B), which could be attributed to lower coverage in AT-rich region; (c) minor alleles had relatively lower sequencing depth, but the difference between the depths of the minor and major alleles was not statistically significant (Wilcox paired test P value = 0.056) (Figure 1C); (d) a majority of loci contained one SNP and only 1733 (2.14%) were identified with >2 SNPs (Figure 1D).

Genetic diversity

Genetic parameters evaluated using putative SNPs included MAF, H_{obs}, H_{exp}, F_{IS}, and deviation from HWE for the studied population. The distribution of MAF showed that the SNPs with MAF <0.15 were overrepresented at nearly half of the SNPs (49.0%), and SNPs were evenly distributed in high-MAF bins ranging from 0.25 to 0.5 (29%) (Figure 2A). The average MAF of all SNP loci was 0.20.

Of 96,188 SNPs, 11,876 (12.35%, at $P \leq 0.05$) were shown to deviate from HWE. Following application of the Benjamini–Hochberg correction, 8205 (5.53%) SNPs remained significant ($P \leq 0.05$). The excess of H_{obs} values was observed in the frequency distribution from 0 to 0.15, and the deficit was from 0.4 to 0.5 (Figure 2B). The deficit of heterozygotes ($F_{IS} > 0$) is shown in Figure 2C, indicating a high level of inbreeding. The average H_{obs}, H_{exp}, and F_{IS} calculated across all 96,188 SNPs were 0.24, 0.28, and 0.16, respectively.

Genetic structure: In order to assess whether there is a stratification within the farmed sika deer population, population structure was estimated using all SNPs (96,188) across >32 animals (≥75%). Since the STRUCTURE model assumes that the loci are independent and at HWE, we also performed STRUCTURE analysis on a filtered SNP dataset (13,739) at HWE ($P \geq 0.05$) with one SNP on each ddRAD loci across >38 animals (≥90%). Graphic displays of the consistent results from two SNP datasets provided a meaningful explanation of the genetic structure and levels of admixture for the farmed populations (Figure 3, A and C). At K = 2, the two clusters were distinct; four animals (SP1, SP3, SP4, and SP5) were clearly distinguished and, of these, SP3 and SP4 represented an admixture pattern (two clusters of genetic background). At K = 3, another cluster of two animals (AD5 and DF4) was evident. Similar results were also observed using the PCA approach based on these two SNP datasets. Although PC1 and PC2 accounted for only 2.4% and 1.9% of total variation, respectively, four animals (SP1, SP3, SP4, and SP5) were clearly distinguishable based on PC1, and two animals (AD5 and DF4) based on PC2 (Figure 3, B and D).

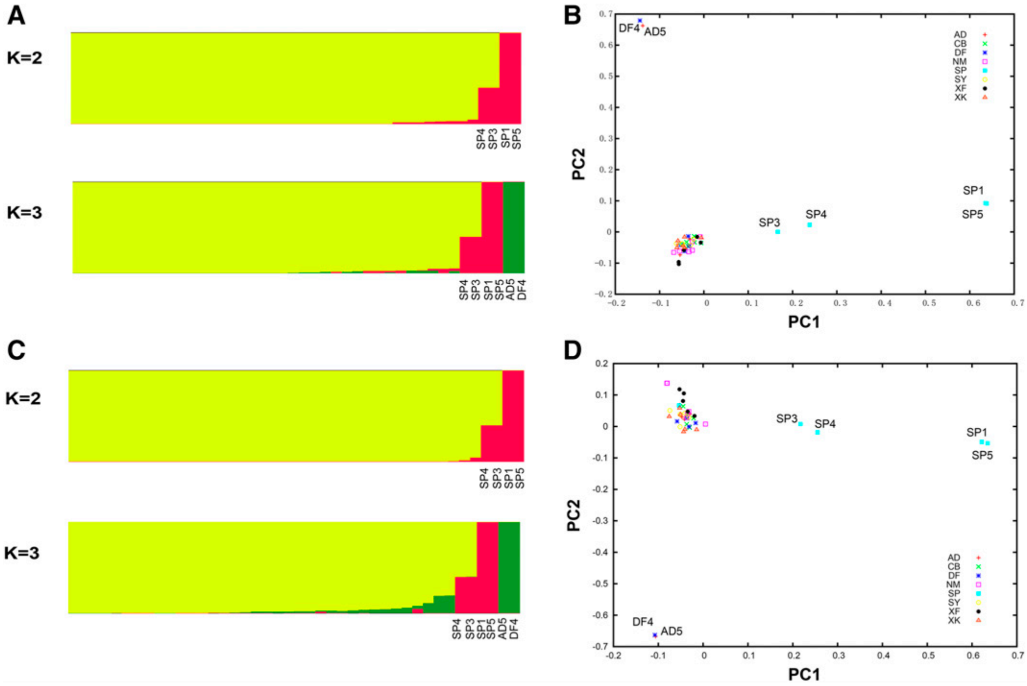

Figure 3 Population genetic structure. (A) STRUCTURE analysis for $K = 2$ and 3 for 96,188 SNPs. (B) PCA analysis for 96,188 SNPs. (C) STRUCTURE analysis for $K = 2$ and 3 for 13,739 SNPs. (D) PCA analysis for 13,739 SNPs.

SNP probes

We selected 27,009 high-quality SNPs probes according to four filtration criteria. The Ts/Tv ratio was 3.72. The MAF frequency distribution was accordant between the 96,188 and 27,009 SNP datasets (Figure 4), suggesting that the repetitive sequences have no effect on the change of MAF frequency distribution for the discovered SNPs. These SNP probes (101 bp) were matched with the bovine genome assembly (UMD3.1) via BLASTN, and their alignment resulted in 58.42% (15,778/27,009) probes producing 50,447 *E*-value informative hits (*E* value $1e^{-20}$). As expected, 96.24% (15,185/15,778) of these matched SNP probes produced one unique alignment to a bovine genome, with an average identity of 94.39%. Only 3.76% (593/15,778) probes hit multiple cattle genome positions (35,262).

DISCUSSION

The ddRAD-seq technique is an efficient and cost-effective means of SNP discovery, providing thousands of high-quality SNPs, even in the absence of an available genome sequence (Peterson *et al.* 2012). Using the ddRAD-seq technique, we reported here the first genome-wide set of 96,188 novel SNPs in sika deer from 42 unrelated animals selected from eight farms.

In order to evaluate the application of SNPs reported in our study, the SNP quality was screened based on the data characteristics of next-generation sequencing technology. Sequencing errors are often found at the 3′ end of the sequence reads (Dohm *et al.* 2008). If a substantial proportion of the SNPs in the dataset were the result of sequencing errors, the number of SNPs toward the end of the reads should increase. However, our results showed that the number of SNPs identified actually decreased from the 5′ to the 3′ end (Figure 1A), indicating that base errors at the 3′ end were not identified as SNPs. An underrepresentation at both ends of the reads (positions 1–3 and 95–100) was attributed to the *Mse*I and *Pst*I restriction sites and the effect of barcode trimming. However, the Ts/Tv ratio is very stably distributed over the loci except at the very ends of the reads, and was not affected by the relatively higher base error rate at the ends of the reads.

The expected Ts/Tv ratio can vary with the targeted region (*i.e.*, whole genome, exon, specific genes) and can also vary greatly in the CpG islands and GC content of the region (Dohm *et al.* 2008). For example, the Ts/Tv ratio of 3.5 is typical of coding regions in human genomes (Le and Durbin 2011). Higher Ts/Tv ratios are also observed in other reduced representation libraries for SNP discovery (Kraus *et al.* 2011; Le and Durbin 2011). Therefore, it is not unexpected that the observed Ts/Tv ratio is relatively higher in the SNP dataset. The number of A/T substitutions is relatively lower than other types of substitutions (Figure 1B). This could be attributed to the fact that higher coverage needs to be attained to discover real SNPs in AT-rich regions (Dohm *et al.* 2008). Matching the base content of a restriction site to a reference genome can also influence substitution types (Davey *et al.* 2011).

To date, the SK-REF database comprises ~90% of the sika deer genome, and is incomplete. The unique alignment of ddRAD reads does not guarantee that there is no other similar sequence in the remainder of the genome. This type of genome mapping could result in abnormally high coverage in the paralogous or duplicated sequence, which will likely generate a large number of false-positive SNPs with excessive heterozygotes (Malhis and Jones 2010). Our results show that the depths of almost all alleles fall into a reasonable range (less than threefold average depth, 42× for minor alleles and 60× for major alleles) (Figure 1C). This is an acceptable depth range for SNP discovery that had been identified in other reduced representation libraries of sequencing projects, including cattle (Eck *et al.* 2009), porcine (Amaral *et al.* 2009) and turkey (Kerstens *et al.* 2009). We did not observe excess heterozygotes, in which the number of SNPs with $H_{obs} > 0.5$ is relatively lower in the SNP dataset (Figure 2B). In addition, only 1733 (2.14%) ddRAD loci were identified with >2 SNPs, which was also a good indicator of correct homologous loci matching (Figure 1D).

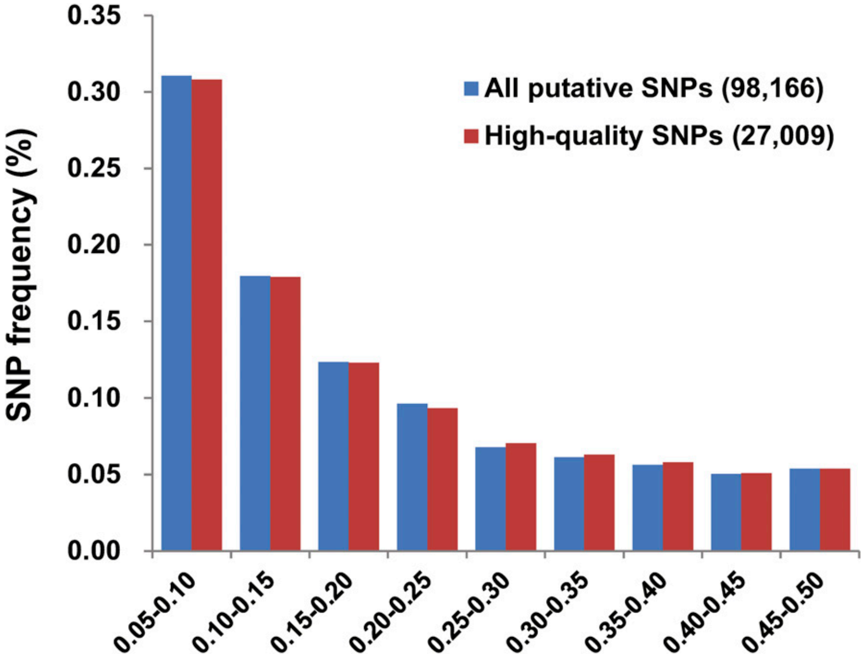

Figure 4 Comparison of SNP frequency distribution between all putative SNPs and high-quality SNPs.

The average frequency of heterozygous SNPs per kilobase pair is 0.74 in the genome of sika deer individuals. As a comparison, this value is slightly higher than that of milu (a highly endangered species that is susceptible to genetic drift and inbreeding because of small population size), but almost <2 times those of cow and panda (Table 2). The relatively low level of genetic variability in the farmed population could be due to ongoing inbreeding within a narrow gene pool and genetic drift. Additionally, the 42 individuals sequenced were from eight farmed populations that may be small and divergent in allele frequencies. For instance, almost 5.53% (8205) of SNPs (Benjamini–Hochberg correction $P \leq 0.05$) deviate from HWE, which may be caused by the different subpopulations sampled. The higher values of F_{IS} and the lower-than-expected values of H_{obs} could be also explained by subpopulation structure (Wahlund effect). The STUCTURE and PCA results further provide evidence of subpopulation structure in farmed sika deer, despite the lack of overall variation and the fact that the PC1 and PC2 accounted for such a small proportion (2.4% and 1.9%, respectively) of the total variation. The small-scale farming population and the application of the technique of artificial insemination over a long period have encouraged inbreeding and genetic drift for velvet antler production. Such population events also compromise the genetic structure of the original population. Hence, our results could help in investigating the structure of farmed populations so such disadvantages can be better managed.

SNP markers are not only readily applied to population genetic analyses, but also suitable for parentage testing and assessment of breed composition. Genome-wide association analysis may offer the opportunity to identify the genomic regions and mutations that underpin production traits such as velvet antler weight. One attractive feature of association studies is that pedigrees are not necessary, so potentially a larger number of the farmed animals may be amenable to this type of analysis (Seabury *et al.* 2011). However, high-quality phenotypes are critical, and this will require major investment.

There may be potential to apply genomic selection for the genetic improvement of farmed sika deer. The small effective population size may make this an attractive proposition. The option of introducing a different subpopulation would also enable the exploitation of heterosis (hybrid vigor).

Overall, we have discovered ~27,009 high-quality SNPs in 1% of the sika genome using ddRAD-seq technology. A total of 58% (15,778/27,009) probes were uniquely matched on to the bovine genome (BLASTN E value $1e^{-20}$), which could be an indication of the sharing of SNP genotypes between the genomes. Only 593 SNP probes of sika deer could be aligned multiple times to the bovine genome, which may represent a duplicated and/or expanded region on the cattle genome. In comparison, Haynes and Latch (2012) successfully genotyped a lower proportion of loci (39%) in mule deer using BovineSNP50 BeadChip, and Powell *et al.* (2016) directly captured 60% exon data from bovine genome on mule deer genome, matching rates almost concordant with our data. As deer and cattle diverged 27–32 MYA (Hassanin and Douzery 2003; Guha *et al.*

■ **Table 2 Information of SNP frequency among four species**

Species	Study Population	Sample Size	Average SNP Frequency in Individual (kb^{-1})	Average Depth	Reference
Sika Deer	Deer from eight farms	42	0.74	23×	This study
Milu (Pere David's Deer)	Captive population	5	0.51	NA[a]	Zhu *et al.* (2016)
Panda	Wild population	34	1.32	4.7×	Zhu *et al.* (2016)
Holstein-Friesian Cows	Breeding population	32	1.35	14.3×	Szyda *et al.* (2015)

[a] NA, not applicable.

2007), the matching of sika deer SNPs discovered in the bovine genome would likely be useful to study genomic evolution and phylogenetic analysis across species within Cervidae, given the high cross-species matching success rate. Although, to date, there is no valuable reference genome within the Cervidae family and map positions are unknown, SNP assays derived from nonrepetitive loci (contigs) could be designed for high-density genotyping across cervid species, providing a reasonable genome analysis tool for future research.

ACKNOWLEDGMENTS

We thank Peter Fennessy and Jia Yankai for reading the paper and giving valuable comments. This work was funded by the Chinese National Science Foundation (31402035) and partially funded by Jilin Province Science Foundation (20170101003JC).

LITERATURE CITED

Amaral, A. J., H. J. Megens, H. H. Kerstens, H. C. Heuven, B. Dibbits *et al.*, 2009 Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. BMC Genomics 10: 374.

Ba, H., F. Yang, X. Xing, and C. Li, 2015 Classification and phylogeny of sika deer (Cervus nippon) subspecies based on the mitochondrial control region DNA sequence using an extended sample set. Mitochondrial DNA 26: 373–379.

Ba, H., L. Wu, Z. Liu, and C. Li, 2016 An examination of the origin and evolution of additional tandem repeats in the mitochondrial DNA control region of Japanese sika deer (Cervus Nippon). Mitochondrial DNA A DNA Mapp. Seq. Anal. 27: 276–281.

Baxter, S. W., J. W. Davey, J. S. Johnston, A. M. Shelton, D. G. Heckel *et al.*, 2011 Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. PLoS One 6: e19315.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko, 2013 Stacks: an analysis tool set for population genomics. Mol. Ecol. 22: 3124–3140.

Catchen, J. M., A. Amores, P. Hohenlohe, W. Cresko, and J. H. Postlethwait, 2011 Stacks: building and genotyping Loci de novo from short-read sequences. G3 1: 171–182.

DaCosta, J. M., and M. D. Sorenson, 2014 Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. PLoS One 9: e106713.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen *et al.*, 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12: 499–510.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491–498.

Dohm, J. C., C. Lottaz, T. Borodina, and H. Himmelbauer, 2008 Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 36: e105.

Eck, S. H., A. Benet-Pages, K. Flisikowski, T. Meitinger, R. Fries *et al.*, 2009 Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. Genome Biol. 10: R82.

Graffelman, J., and J. M. Camarena, 2008 Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. Hum. Hered. 65: 77–84.

Groeneveld, L. F., J. A. Lenstra, H. Eding, M. A. Toro, B. Scherf *et al.*, 2010 Genetic diversity in farm animals–a review. Anim. Genet. 41 (Suppl. 1): 6–31.

Guha, S., S. P. Goyal, and V. K. Kashyap, 2007 Molecular phylogeny of musk deer: a genomic view with mitochondrial 16S rRNA and cytochrome b gene. Mol. Phylogenet. Evol. 42: 585–597.

Hassanin, A., and E. J. Douzery, 2003 Molecular and morphological phylogenies of ruminantia and the alternative position of the moschidae. Syst. Biol. 52: 206–228.

Haynes, G. D., and E. K. Latch, 2012 Identification of novel single nucleotide polymorphisms (SNPs) in deer (Odocoileus spp.) using the BovineSNP50 BeadChip. PLoS One 7: e36536.

He, Y., Z. H. Wang, and X. M. Wang, 2014 Genetic diversity and population structure of a Sichuan sika deer (Cervus sichuanicus) population in Tiebu Nature Reserve based on microsatellite variation. Dongwuxue Yanjiu 35: 528–536.

Kerstens, H. H., R. P. Crooijmans, A. Veenendaal, B. W. Dibbits, A. W. T. F. Chin *et al.*, 2009 Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. BMC Genomics 10: 479.

Kraus, R. H., H. H. Kerstens, P. Van Hooft, R. P. Crooijmans, J. J. Van Der Poel *et al.*, 2011 Genome wide SNP discovery, analysis and evaluation in mallard (Anas platyrhynchos). BMC Genomics 12: 150.

Krojerova-Prokesova, J., M. Barancekova, I. Voloshina, A. Myslenkov, J. Lamka *et al.*, 2013 Dybowski's sika deer (Cervus nippon hortulorum): genetic divergence between natural primorian and introduced Czech populations. J. Hered. 104: 312–326.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–359.

Le, S. Q., and R. Durbin, 2011 SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Res. 21: 952–960.

Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993.

Lu, X. P., F. W. Wei, M. Li, G. Yang, and H. Liu, 2006 Genetic diversity among Chinese sika deer (Cervus nippon) populations and relationships between Chinese and Japanese sika deer. Chin. Sci. Bull. 51: 433–440.

Malhis, N., and S. J. M. Jones, 2010 High quality SNP calling using Illumina data at shallow coverage. Bioinformatics 26: 1029–1035.

McCullough, E. D. R., S. Takatsuki, and K. Kaji, 2009 Sika deer: biology and management of native and introduced populations, pp. 526–531 in *Sika Deer in Mainland China*, edited by McCullough, D. R., Z.-G. Jiang, and C.-W. Li. Springer, Tokyo.

Patterson, N., A. L. Price, and D. Reich, 2006 Population structure and eigenanalysis. PLoS Genet. 2: 2074–2093.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra, 2012 Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7: e37135.

Powell, J. H., S. J. Amish, G. D. Haynes, G. Luikart, and E. K. Latch, 2016 Candidate adaptive genes associated with lineage divergence: identifying SNPs via next-generation targeted resequencing in mule deer (Odocoileus hemionus). Mol. Ecol. Resour. 16: 1165–1172.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Seabury, C. M., E. K. Bhattarai, J. F. Taylor, G. G. Viswanathan, S. M. Cooper *et al.*, 2011 Genome-wide polymorphism and comparative analyses in the white-tailed deer (Odocoileus virginianus): a model for conservation genomics. PLoS One 6: e15811.

Sheng, H. L., and N. Ohtaishi, 1993 *Deer of China: Biology and Management*. Elsevier Science, Amsterdam.

Szyda, J., M. Fraszczak, M. Mielczarek, R. Giannico, G. Minozzi *et al.*, 2015 The assessment of inter-individual variation of whole-genome DNA sequence in 32 cows. Mamm. Genome 26: 658–665.

Takiguchi, H., K. Tanaka, K. Ono, A. Hoshi, M. Minami *et al.*, 2012 Genetic variation and population structure of the Japanese sika deer (Cervus nippon) in the Tohoku District based on mitochondrial D-loop sequences. Zoolog. Sci. 29: 433–436.

Tamate, H. B., A. Okada, M. Minami, N. Ohnishi, H. Higuchi *et al.*, 2000 Genetic variations revealed by microsatellite markers ina small population of the sika deer (Cervus nippon)on Kinkazan Island, northern Japan. Zoolog. Sci. 17: 47–53.

Thevenon, S., L. T. Thuy, L. V. Ly, F. Maudet, A. Bonnet *et al.*, 2004 Microsatellite analysis of genetic diversity of the Vietnamese sika deer (Cervus nippon pseudaxis). J. Hered. 95: 11–18.

Wang, W., H. Yan, J. Yu, J. Yi, Y. Qu *et al.*, 2017 Discovery of genome-wideSNPs by RAD-seqand the genetic diversity of captive hog deer (Axis porcinus). PLoS One 12: e0174299.

Wu, H., Q. H. Wan, and S. G. Fang, 2004 Two genetically distinct units of the Chinese sika deer (Cervus nippon): analyses of mitochondrial DNA variation. Biol. Conserv. 119: 183–190.

Wu, H., Q. H. Wan, S. G. Fang, and S. Y. Zhang, 2005 Application of mitochondrial DNA sequence analysis in the forensic identification of Chinese sika deer subspecies. Forensic Sci. Int. 148: 101–105.

Wu, P. J., and E. D. Zhang, 2001 The resource conservation and utilization of wild sika deer in China. Zhong Yao Cai 24: 552–554 (in Chinese).

Zhu, L., X. Zhang, C. Deng, J. Ding, Y. Ren *et al.*, 2016 Comparative genomics and metagenomics analyses of endangered Père David's deer (Elaphurus davidianus) provide insights into population recovery. bioRxiv DOI: 10.1101/073528.

Communicating editor: D-J. de Koning