Genome-wide variation within and between wild and domestic yak

KUN WANG,1 QUANJUN HU,1 HUI MA, LIZHONG WANG, YONGZHI YANG, WENCHUN LUO and QIANG QIU

State Key Laboratory of Grassland Agro-Ecosystem, College of Life Science, Lanzhou University, Lanzhou 73000, China

Abstract

The yak is one of the few animals that can thrive in the harsh environment of the Qinghai-Tibetan Plateau and adjacent Alpine regions. Yak provides essential resources allowing Tibetans to live at high altitudes. However, genetic variation within and between wild and domestic yak remain unknown. Here, we present a genome-wide study of the genetic variation within and between wild and domestic yak. Using next-generation sequencing technology, we resequenced three wild and three domestic yak with a mean of fivefold coverage using our published domestic yak genome as a reference. We identified a total of 8.38 million SNPs (7.14 million novel), 383 241 InDels and 126 352 structural variants between the six yak. We observed higher linkage disequilibrium in domestic yak than in wild yak and a modest but distinct genetic divergence between these two groups. We further identified more than a thousand of potential selected regions (PSRs) for the three domestic yak by scanning the whole genome. These genomic resources can be further used to study genetic diversity and select superior breeds of yak and other bovid species.

Keywords: breeding, genome-wide variation, resequencing, selective sweep, wild yak

Received 6 October 2013; revision received 16 December 2013; accepted 27 December 2013

Introduction

The yak (Bos grunniens) is endemic to the Qinghai-Tibetan Plateau (QTP) and in adjacent high-altitude regions (Wiener et al. 2003). Yak were domesticated by the ancient Qiang people from wild yak, possibly at the beginning of the Holocene (Wiener et al. 2003; Guo et al. 2006; Wang et al. 2010). More than 14 million domestic yak provide food (meat and milk), transport, shelter and fuel for Tibetans and others living at high altitudes. Only a few other large animals can survive the extremely cold, harsh and oxygen-poor conditions in this area (Zi 2003). In addition, 15–20 thousand wild yak still roam the northwestern QTP (Schaller & Liu 1996). In recent decades, there have been attempts to produce hybrids between domestic and wild yak, leading to a successful hybrid breed, the Datong yak, which exhibits greater body size, faster growth rate and higher disease resistance than the other breeds (Wiener et al. 2003). However, genetic variation studies within and between domestic and wild yak remain limited, except those based exclusively on the mitochondrial DNA (Guo et al. 2006; Wang et al. 2010, 2011).

Genome-wide analyses using next-generation sequencing (NGS) technology provide an unprecedented opportunity to examine genetic variation between and within wild and domestic plants and animals (Ellegren & Sheldon 2008; Stapley et al. 2010; Helyar et al. 2011). With a well assembled-reference genome, a genome-wide variation map can be easily obtained by the application of NGS and alignment software (Garvin et al. 2010). Any detected variation can then be used as molecular markers in studies of genetic diversity and breed selection (Kerstens et al. 2009; Huang et al. 2010; Lai et al. 2010; Larkin et al. 2012). Moreover, strong artificial selection will increase the frequency of favourable alleles at loci affecting interesting beneficial traits (Kim & Stephan 1999). These selection sweep regions can be approximately identified by finding the change in diversity between wild and domestic populations (Meuwissen et al. 2013). For example, resequencing genomes of domestic and wild silkworms discovered 354 candidate genes linked to domestication, which are probably associated with silk gland development, energy metabolism and reproduction (Xia et al. 2009). Similarly, a genomic comparison of domestic and wild pigs identified a strong selection signature associated with several genes (e.g. NR6AI, PLAG1 and LCORL) involved in the elongation of the back and an increased number of vertebrae (Rubin
et al. 2012). In addition, Huang et al. (2012) found 55 domestication loci by comparing the genome-wide variation between cultivated and wild rice individuals.

Studies of genomic variation in large livestock species, such as horse and cattle, have mostly concentrated on comparisons between different domestic breeds because their wild progenitors have become extinct (Larkin et al. 2012; Petersen et al. 2013). The last wild cattle (Bos primigenius), died in Europe in 1627 (Edwards et al. 2007) while the progenitors of modern horses became extinct in the 19th century (Outram 2007) because their wild progenitors have become extinct (Larson 2007). Comparisons between different domestic breeds provide us not only with valuable breeding resources, but also with the first opportunity to examine the origin and domestication of the large livestock. With the availability of a domestic yak genome (Qiu et al. 2012) and the establishment of the yak genome database (Hu et al. 2012), we were able to perform the whole-genome resequencing of wild and domestic yak. We aimed to examine genetic variation within and between domestic and wild yak. These genomic resources will be useful for studies of genetic diversity and genetic improvement in yak and other bovid species.

Materials and methods

Sample collection and DNA sequencing

Three wild yak were collected in the Hoh Xil, and all of them have taxonomic characters specific to the wild yak: long hair and large skeleton (Wang et al. 2010). Three domestic yak from Hezuo in Gansu, Ruoergai in Sichuan and Lasa in Tibet were also sampled. Genomic DNA was extracted using the standard phenol–chloroform method (Sambrook 2001). For each individual, 5 μg of DNA was fragmented to an insert size of 500 bp. DNA fragments were then treated according to the Illumina DNA sample preparation protocol: fragments were end-repaired, A-tailed, ligated to paired-end adaptors and PCR amplified for library construction. Sequencing was performed on the Illumina HiSeq 2000 platform, and 90-bp paired-end reads were generated.

Reads mapping and SNP identification

Reads with more than 10% unidentified nucleotides or more than 65% bases with Q20 ≤ 7 were filtered out with a custom perl script. We mapped paired-end reads to the reference genome using BWA (Li & Durbin 2009) with the default parameters. Only uniquely mapped and paired aligned reads were used for SNP identification. We then sorted the Binary Alignment/Map (BAM) files using SAMTOOLS (Li et al. 2009) and applied the ‘rmdup’ command to remove the PCR duplications generated in the process of sequencing. To reduce the inaccuracy alignment caused by insertion and deletion (InDels), we realigned reads that located in regions around InDels using GATK (McKenna et al. 2010) for each sample, separately. The known SNPs (Qiu et al. 2012) were used to recalibrate the base quality in the BAM files. The resulted mapping files were employed simultaneously to genotype the six samples with ‘UnifiedGenotyper’ in GATK (DePristo et al. 2011). The highest scoring 10% resulting SNP calls were used as a training set for variant quality recalibration to filter unreliable SNPs. Among the remaining high-quality sites, we only retained SNPs that were both successfully called and covered by at least two reads in each individual (Ellegren et al. 2012).

Population genetic analysis

The software PLINK (Purcell et al. 2007) was used to perform the principal component analysis (PCA) on the SNPs of the six individuals with default parameters. The reads of cattle (SRA: ERP000712; Zhan et al. 2011) were used as an outgroup to genotype the ancestral state. The genetic distance matrix was calculated based on pairwise identity-by-state (IBS) distance. The phylogenetic tree was constructed using the software PHYLIP (Felsenstein 2009) and visualized with MEGA5 (Tamura et al. 2011). The population-differentiation statistics (F_{ST}) and the linkage disequilibrium (LD) were calculated by VCF tools (Danecek et al. 2011).

Detection of Insertions and Deletions (InDels) and structural variants (SVs)

We used GATK to identify InDels using the recalibrated BAM files generated in the previous steps. Top ten of the best-quality InDels were used as the training set to model InDels with variant quality score recalibration (VQSR). The model was applied to recalibrate the quality of InDels, and only high-quality InDels were retained. We put six BAM files together to detect structural variants using BreakDancer package (Chen et al. 2009). Any SV for each individual with support of less than two reads was filtered out.

Identification of Potential Selected Regions (PSRs) and spatial autocorrelation test

We performed a whole-genome scan using the sliding window method with a size of 50 kb to calculate the ratio of the diversity in the wild population (\(\pi_w\)) to the diversity in the domestic population (\(\pi_d\)). The windows with the highest 5% ratio were selected as the potential selected regions (Xia et al. 2009; Vasemagi et al. 2012; Li et al. 2010). Any InDels or SVs were realigned in these regions using GATK to tolerate the inaccuracy alignment caused by InDels or SVs. The program PLINK was used to call SNPs from the re-aligned reads with default parameters. Any SNPs with minor allele frequency (MAF) ≤ 0.05 were filtered out.

SNPs were subjected to a sliding window method with a size of 50 kb and a step size of 10 kb to calculate the ratio of diversity between the wild and domestic population (\(\pi_w/\pi_d\)). The windows with the highest 5% ratio were selected as the potential selected regions (Xia et al. 2009; Vasemagi et al. 2012; Li et al. 2010). Any InDels or SVs were realigned in these regions using GATK to tolerate the inaccuracy alignment caused by InDels or SVs. The program PLINK was used to call SNPs from the re-aligned reads with default parameters. Any SNPs with minor allele frequency (MAF) ≤ 0.05 were filtered out.

Identification of Potential Selected Regions (PSRs) and spatial autocorrelation test

We performed a whole-genome scan using the sliding window method with a size of 50 kb to calculate the ratio of the diversity in the wild population (\(\pi_w\)) to the diversity in the domestic population (\(\pi_d\)). The windows with the highest 5% ratio were selected as the potential selected regions (Xia et al. 2009; Vasemagi et al. 2012; Li et al. 2010). Any InDels or SVs were realigned in these regions using GATK to tolerate the inaccuracy alignment caused by InDels or SVs. The program PLINK was used to call SNPs from the re-aligned reads with default parameters. Any SNPs with minor allele frequency (MAF) ≤ 0.05 were filtered out.

Identification of Potential Selected Regions (PSRs) and spatial autocorrelation test

We performed a whole-genome scan using the sliding window method with a size of 50 kb to calculate the ratio of the diversity in the wild population (\(\pi_w\)) to the diversity in the domestic population (\(\pi_d\)). The windows with the highest 5% ratio were selected as the potential selected regions (Xia et al. 2009; Vasemagi et al. 2012; Li et al. 2010). Any InDels or SVs were realigned in these regions using GATK to tolerate the inaccuracy alignment caused by InDels or SVs. The program PLINK was used to call SNPs from the re-aligned reads with default parameters. Any SNPs with minor allele frequency (MAF) ≤ 0.05 were filtered out.
et al. 2013; Qi et al. 2013). We used LastZ (Harris 2007) to obtain information on the synteny between yak and cattle (UMD 3.1) and plot the ratio ($\pi_w / \pi_d$) from yak into cattle chromosome window by window.

We used a simple method to test the spatial autocorrelation of PSRs (Payseur et al. 2002). The nonoverlap windows were regarded as the basic unit and assumed to be isolated from each other. A window with the highest 5% ratio ($\pi_w / \pi_d$) was treated as high ratio window (HRW), and the percentage of HRWs that had at least another HRW neighbour was regarded as spatial autocorrelation level. The windows of each scaffold were shuffled 1000 times to evaluate expected value and standard error of the percentage in random situation. Based on the normal frequency distribution, Z-test was used to test the statistical significance.

**SNP visualization**

The SNP information of our resequencing data was added into the yak genome database for visualization (http://me.lzu.edu.cn/yak/). A custom perl script was used to convert the SNPs from variant call format (VCF) into general feature format (GFF). Then, we used bp_load_gff.pl to import the GFF-formatted files into the MySQL relational database as recommended by the Gbrowse tutorial.

**Results and discussion**

**Data production**

A total of 102 gigabases of high-quality sequences with 1.14 billion reads were generated. We mapped the reads of six samples to the yak reference genome with BWA. A summary of the resequencing data is presented in Table 1. After eliminating unknown bases in the yak genome, at least 95% of the genome was covered by reads from a single individual and most regions (85%) of the genome were covered commonly by all individuals. The average unique mapping rate of these reads was 88%, resulting in an average 5.5-fold coverage.

<table>
<thead>
<tr>
<th>Sample IDs</th>
<th>Sample names</th>
<th>Number of raw reads</th>
<th>Number of clear reads</th>
<th>Number of mapped reads</th>
<th>Mapping rate (%)</th>
<th>Unique mapping rate (%)</th>
<th>Coverage (%)</th>
<th>Effective depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1†</td>
<td>W1</td>
<td>185 466 760</td>
<td>175 159 334</td>
<td>173 239 766</td>
<td>98.90</td>
<td>87.18</td>
<td>95.23</td>
<td>5.42</td>
</tr>
<tr>
<td>W2</td>
<td>W2</td>
<td>199 499 682</td>
<td>185 493 094</td>
<td>184 139 478</td>
<td>99.27</td>
<td>85.90</td>
<td>95.84</td>
<td>5.50</td>
</tr>
<tr>
<td>W3</td>
<td>W3</td>
<td>182 567 864</td>
<td>171 842 422</td>
<td>170 371 908</td>
<td>99.14</td>
<td>86.97</td>
<td>95.32</td>
<td>5.23</td>
</tr>
<tr>
<td>D1</td>
<td>D1</td>
<td>178 052 372</td>
<td>168 514 758</td>
<td>167 644 438</td>
<td>99.48</td>
<td>88.25</td>
<td>95.84</td>
<td>5.32</td>
</tr>
<tr>
<td>D2</td>
<td>D2</td>
<td>193 416 260</td>
<td>176 81 2800</td>
<td>174 328 320</td>
<td>98.59</td>
<td>89.69</td>
<td>97.17</td>
<td>5.87</td>
</tr>
<tr>
<td>D3</td>
<td>D3</td>
<td>197 323 732</td>
<td>172 305 695</td>
<td>171 200 435</td>
<td>99.36</td>
<td>89.92</td>
<td>96.90</td>
<td>5.68</td>
</tr>
</tbody>
</table>

†W1, W2 and W3 are wild individuals. D1, D2 and D3 represent domestic individuals.

**SNP identification and annotation**

With strict quality control and filtering, we obtained a total of 8.38 million SNPs, 7.14 million of which were identified for the first time. Among all SNPs, 6.51 million were located in intergenic regions, 1.79 million in intron regions and 83 280 in coding regions (Table 2). The transition-to-transversion (Ti/Tv) ratio was 2.39 for the genome and 3.06 for coding regions only. The observed ratio of heterozygous-to-homozygous SNPs for every individual was approximately 1.5 (Table S1, Supporting information), higher than that observed for an individual human genome (Levy et al. 2007).

We divided the 8.38 million SNPs into three categories: 5.11 million found in both wild and domestic yak, 1.59 million specific to domestic yak and 1.69 million specific to wild yak (Fig. 1a). No region showed significant enrichment of the specific SNPs (100 kb window with chi-square test compared with genome background) for either wild or domestic yak. The sequence diversity ($\pi$) was estimated at $9.70 \times 10^{-4}$ in wild yak and $9.67 \times 10^{-4}$ in domestic yak, which were slightly lower than that in all individuals ($10.12 \times 10^{-4}$). The level of differentiation between them, $F_{ST}$, was estimated to be 0.06, which is lower than that between dogs and their wolf ancestors (about 0.22; Axelsson et al. 2013) and humans from different continents (about 0.11 between East-Asia and Europe; Nelis et al. 2009). The

---

Table 1: Summary of the yak resequencing data

Table 2: The distribution of SNPs in the yak genome
high proportion of shared SNPs, similar diversity between single population and all individuals and the low FST revealed a small genetic divergence between domestic and wild yak. However, the principal component analysis (PCA) indicated that wild and domestic yak are significantly separated along the first principal component, which accounts for the greatest proportion of variability (Fig. 2a). The neighbour-joining tree also indicated the wild and domestic yak can be divided into two groups (Fig. 2b). Moreover, we observed that linkage disequilibrium decays slower in domestic yak than in wild yak (Fig. 3), which may be caused by the population bottleneck event during the process of yak domestication.

**Short insertions and deletions (InDels) and Structural Variants (SVs)**

Short insertions and deletions are the second most abundant type of genetic variation in the whole genome (Wang et al. 2008). After being carefully filtered, a total of 182 222 insertions and 201 019 deletions were obtained (Fig. 1b). Within all InDels, 313 645 were found in both domestic and wild yak, 35 142 were specific to domestic yak and 34 454 were specific to wild yak. The ratio of heterozygous-to-homozygous InDels was one-third that of the SNPs, consistent with a previous study of humans (Montgomery et al. 2013). We recovered 1008 InDels (0.26%) in the coding sequences of the 852 genes. These coding InDels were enriched for sizes that were multiples of three sites (3n) (Fig. S1, Supporting information). This might be the result of purifying selection against frame-shifts in the coding regions (Zhan et al. 2011). While the genome-wide SNP:InDel ratio is 22:1, the SNP:InDel ratio for coding regions is 83:1, as would be expected because of the more deleterious effect of InDels on the protein-coding regions.

We detected a total of 126 352 reliable structural variants. Among these, 72 481 were insertions, 52 454 were deletions, 713 were inversions, and 704 were chromosomal translocations (Fig. 1c). Within these SVs, 108 478 were present in both wild and domestic yak, 4963 were specific to domestic yak and the other 12 911 were specific to the wild yak. There was no significant difference
between the lengths of these three types of SVs (Fig. 1d). In total, the SVs covered 10,526 genes and the coding sequences of 1923 genes. Genome-wide analyses indicated that 90.39% and 8.61% of these SVs were located in the intergenic regions and the intronic regions, respectively. Compared with the randomly selected equivalent regions in the genome, we observed a significant bias ($P < 2.2 \times 10^{-16}$, chi-squared test with 1000 replicates by Monte Carlo simulation) towards intergenic regions, suggesting that SVs occurred more frequently in regions with a low functional constraint.

**Screening and annotation of domestication-related loci**

The domestication and artificial selection led to a reduction in nucleotide diversity and altered allele frequency. We measured the ratio of the genetic diversity in wild yak to that in domestic yak ($\pi_w/\pi_d$) across the yak genome using a sliding window method. The 5% of windows with the highest ratio ($\pi_w/\pi_d > 2.12$), a total of 2485, were considered to contain selective signatures (Fig. 4). The adjacent windows were merged into a single region, which resulted in 1910 potential selected regions (PSRs) (Table S2, Supporting information), 104 of which comprised more than three windows. The longest PSR was located at scaffold740_1, spanning from 2.5 to 4.2 Mb. Compared with the overall genome, PSRs showed higher gene density (1.37 times higher than the background, chi-square test $P < 2.2 \times 10^{-16}$), probably reflecting more important positions than the other regions.

One major difference between genetic drift and artificial selection is that genetic drift appears randomly in the genome (Lynch et al. 2011), while artificial selection results in low diversity within linkage blocks (Pritchard et al. 2010). We therefore investigated the spatial autocorrelation of PSRs and found a positive and significantly higher spatial autocorrelation in PSRs than in random situation ($Z$-test $P < 2.2 \times 10^{-16}$). This demonstrated that PSRs were tending to cluster together, which was similar to the effect of recent selection sweep (Neher et al. 2013). Besides, we noticed in domestic yak that the

---

**Fig. 2** Genome-wide relationship between domestic and wild yak. (a) The PCA of the six yak. Wild and domestic yak are separated along the first principal component (PC1). (b) Neighbour-joining tree for the six yak. The number adjacent to each branch is the bootstrap value.
level of LD was much higher in PSRs than in the genome background (Fig. 3), showing strong signatures of direct artificial selection.

To gain insights into the functional significance of the genes located in PSRs, we performed Gene Ontology (GO) enrichment analyses by comparing them to the genome background. We found that the GO terms acute inflammatory response (GO:0006953), antibiotic transporter activity (GO:0042895) and defence response (GO:0006952) were significantly enriched (chi-square test \( P < 0.05 \)). This would probably have been the result of a rapid response to new pathogens associated with frequent contact with humans during domestication. The genes of PSRs probably cause great differences in morphology, physiology and behaviour. We detected one body stature-related gene (\( \text{PLAG1} \)) and two milk characteristic-related genes (\( \text{DGAT1} \) and \( \text{ABCG2} \)), which had been identified previously as crucial domestication-related genes (Grisart \textit{et al.} 2002; Cohen-Zinder \textit{et al.} 2005; Karim \textit{et al.} 2011). However, it should be noted that all these findings need to be tested further based on sampling more wild and domestic individuals.

Yak genetic variation database

We have added the newly obtained data into the yak genome database (Hu \textit{et al.} 2013) for data visualization. We integrated our variation data into Gbrowse, a popular genome browser used in the Generic Model Organism Database (GMOD) project. All data have been organized into a Postgresql relational database, which is efficient in retrieving data from indexed files. An overview of the browser window is presented in Fig. S2 (Supporting information).

Conclusion

In this study, we report a large genome variation data set for wild and domestic yak using a DNA resequencing-based strategy. We identified a total of 8.38 million SNPs.
(7.14 million of these are novel), 182 222 insertions, 201 019 deletions and 126 352 different types of structural variants within all individuals. We integrated these variation data into the yak genome database, resulting in a comprehensive compilation of yak genetic variation identified to date. We found low genetic differentiation between wild and domestic yak although they seemed to aggregate into separate clusters. In addition, we found more than one thousand candidate genomic regions that might have been exposed to strong artificial selection during domestication. The identified candidate genes may be economically important. Despite the limitation of the sampled individuals, all these findings are very interesting and deserve to be carefully examined and tested based on more samples in future. Overall, this report of the yak genome-wide SNPs should prove useful for marker-assisted breeding and for the study of genetic diversity in both yak and other bovid species.

Acknowledgements

Financial support was provided by the National High Technology Research and Development Program of China (863 Program, 2013AA102505-3-2), National Natural Science Foundation of China (31320252, 31209946), the Fundamental Research Funds for the Central Universities (lzujbky-2009-k05, lzujbky-2013-205) and the International Collaboration 111 Projects of China, the 985 and 211 Projects of Lanzhou University.

References


Purcell S, Neale B, Todd-Brown K et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics, 81, 559–575.


Q.Q. and K.W. conceived this study. L.W. and Y.Y. collected samples and extracted DNA. K.W., W.L. and H.M. processed the data. Q.H. updated the YGD. All authors read and approved the final manuscript.

Data Accessibility

Supporting Information
Additional Supporting Information may be found in the online version of this article:
Fig. S1 The relationship between size and number of InDels in coding regions.
Fig. S2 Visualization of genome-wide variation using Gbrowse.
Table S1 Basic information for each individual.
Table S2 List of 1910 potential selected regions in the yak genome.