Massively Parallel Sequencing of Chikso (Korean Brindle Cattle) to Discover Genome-Wide SNPs and InDels

Jung-Woo Choi¹, Xiaoping Liao², Sairom Park³, Heoyn-Jeong Jeon⁴, Won-Hyong Chung⁵, Paul Stothard², Yeon-Soo Park⁶, Jeong-Koo Lee³, Kyung-Tai Lee⁴, Sang-Hwan Kim⁷, Jae-Don Oh⁷, Namshin Kim⁵, Tae-Hun Kim⁴, Hak-Kyo Lee^{7,*}, and Sung-Jin Lee^{3,*}

Since the completion of the bovine sequencing projects, a substantial number of genetic variations such as single nucleotide polymorphisms have become available across the cattle genome. Recently, cataloguing such genetic variations has been accelerated using massively parallel sequencing technology. However, most of the recent studies have been concentrated on European *Bos taurus* cattle breeds, resulting in a severe lack of knowledge for valuable native cattle genetic resources worldwide. Here, we present the first whole-genome sequencing results for an endangered Korean native cattle breed, Chikso, using the Illumina HiSeq 2,000 sequencing platform. The genome of a Chikso bull was sequenced to approximately 25.3-fold coverage with 98.8% of the bovine reference genome sequence (UMD 3.1) covered. In total, 5,874,026 single nucleotide polymorphisms and 551,363 insertion/deletions were identified across all 29 autosomes and the X-chromosome, of which 45% and 75% were previously unknown, respectively. Most of the variations (92.7% of single nucleotide polymorphisms and 92.9% of insertion/deletions) were located in intergenic and intron regions. A total of 16,273 single nucleotide polymorphisms causing missense mutations were detected in 7,111 genes throughout the genome, which could potentially contribute to variation in economically important traits in Chikso. This study provides a valuable resource for further investigations of the genetic mechanisms underlying traits of interest in cattle, and for the development of improved genomics-based breeding tools.

INTRODUCTION

Korean brindle cattle, known as Chikso, are one of the four indigenous cattle breeds in the Korean peninsula. Chikso have been maintained at very low population sizes and raised in limited areas in South Korea, such as Ulleung Island, Gyeongbuk and Hongcheon, Kangwon in Korea (Choi, 2009; Food Agricultural Organization (FAO), 2012; Jo et al., 2012). The name Chikso is derived from 'Chik', referring to its striped black hair belts on a yellowish brown hair background resembling the kudzu vine, while 'so' means cattle in Korean. The Chikso was also termed 'Ho-Ban-Woo' (tiger cattle) because of its resemblance to a tiger's coat color pattern (Fig. 1) (FAO, 2012). Historical records indicate that Chikso was used mainly as draft and pack animals as it was considered good fortune to have these animals under your roof. Recently, however, Chikso have received attention as beef cattle as demands for safe meat from native cattle breeds have increased in South Korea. At the beginning of the 20th century, policies to unify various coat colors in cattle breeds were enforced, leading to a loss of diverse genetic resources in cattle in the Korean peninsula (Choi, 2009). This decreased genetic diversity has not been properly restored, partly because of the focus on Korean brown cattle as a representative beef cattle breed in recent decades. As a result, the current population of Chikso is at high risk of extinction and Chikso are classified as an endangered species by the FAO (NIAS, 2012). Recently, Chikso have increased in value, especially in the context of conserving such a valuable native genetic resource, and as a new niche beef market in Korea. However, there is a severe lack of genetic information and few genomic investigations have been performed on Chikso cattle, leading to gaps in our knowledge concerning the degree of inbreeding in the current population and their genetic relationships with other

¹Centre for Genetic Improvement of Livestock, Animal & Poultry Science, University of Guelph, Guelph, Ontario N1G 2W1, Canada, ²Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB T6G 2P5, Canada, ³College of Animal Life Sciences, Kangwon National University, Chuncheon 200-701, Korea, ⁴Division of Animal Genomics and Bioinformatics, National Institute of Animal Science, Rural Development Administration, Suwon 441-706, Korea, ⁵Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Korea, ⁶Gangwon Provincial Livestock Research Center, Hoengseong 225-830, Korea, ⁷Genomic Informatics Center and Institute of Genetic Engineering, Hankyong National University, Ansong 456-749, Korea

*Correspondence: breedlee@empas.com (HKL), sjlee@kangwon.ac.kr (SJL)

Received December 28, 2012; revised June 19, 2013; accepted June 24, 2013; published online August 1, 2013

Keywords: Chikso, InDel, massively parallel sequencing, SNP

Korean native cattle breeds.

The completion of the international bovine sequencing and HapMap projects (Bovine Genome Sequencing Analysis Consortium et al., 2009; Bovine HapMap Consortium et al., 2009) have led to substantial numbers of genetic variation, such as single nucleotide polymorphisms (SNPs), becoming widely available for the cattle genome. In particular, recent advances in massively parallel sequencing technology (aka 'Next Generation Sequencing: NGS') have been used successfully to catalog such genetic variations by whole-genome resequencing of diverse cattle breeds in a cost-effective and reasonably accurate manner. For example, using the Illumina Genome Analyzer II platform, Eck et al. (2009) reported approximately 2.4 million SNPs in the sequence of a Fleckvieh bull. The same sequencing platform was used to sequence Japanese native cattle, Kuchinoshima-Ushi, leading to the identification of 6.3 million putative SNPs (Kawahara-Miki et al., 2011). In addition, another massively parallel sequencing platform, the ABI SOLiD system, was applied successfully to compare two genomes that are representative of beef and dairy breeds: a Black Angus and a Holstein bull, and identified approximately 7 million SNPs and 790 putative copy number variations across the genomes (Stothard et al., 2011). Despite the increasing use of such technologies to dissect cattle genomes, to our knowledge, no genome sequencing studies have been published using Korean native cattle breeds, although an NGS study of a Hanwoo bull (Korea brown cattle) is underway for publication. Furthermore, there is a severe lack of genetic studies on native Korean cattle breeds such as the Chikso, which are threatened with extinction, while the more popular Hanwoo has been the subject of more extensive genetic investigations using their relatively more complete phenotype and pedigree records.

In this study, we describe the first whole-genome sequencing results for an endangered native Korean cattle breed, Chikso, using the Illumina HiSeq 2000 sequencing platform. The main objective of this work was to systematically identify genetic variations, including SNPs and insertion/deletions (InDels), throughout the genome to develop a catalog of genetic variation for breeding strategies using DNA marker-assisted selection or genomic selection.

MATERIALS AND METHODS

DNA sampling

We selected a 20-month-old Chikso bull with pedigree records and 11 test measurements recorded at 3-month intervals in the first year, which was raised in Gangwon Provincial Livestock Research Center. Whole blood from the bull was collected in an ethylenediaminetetraacetic-acid (EDTA) tube. Genomic DNA was isolated from the whole blood, specifically from leukocytes, using a PAXgene Blood DNA Kit, according to the manufacturer's instructions (PreAnalytiX GmbH, Hombrechtikon, Switzerland). The quality and quantity of the extracted DNA were assessed by calculating OD values with an Infinite F200 microplate reader (TECAN) and the concentration of double-stranded DNA was determined using a Quant-iT dsDNA BR Assay Kit for use in the Qubit fluorometer (Invitrogen, USA), according to the manufacturer's instructions. A further visual check of the status of the DNA was performed using 0.8% agarose gel electrophoresis.

Library construction and massively parallel sequencing

The purified genomic DNA was randomly sheared by a Covaris S2 (Covaris, USA) to yield DNA fragments in the target range

of 400-500 bp. The average fragment size was assessed by an Agilent Bioanalyzer 2,100 (Agilent Technologies, USA). Following the fragmentation, an Illumina TrueSeq End Repair Kit was used to convert the resulting overhangs to blunt ends prior to a cleanup step using AMPure XP Beads (Beckman Coulter Genomics, USA). To increase the success of ligation between the fragmented DNA and index adapters, as well as to reduce self-ligation of the blunt fragments, the 3' ends were adenylated. Immediately following adenylation, the index adapters were ligated to the freshly adenylated, fragmented genomic DNA, which was then purified using the AMPure XP Beads. The ligation products were then size-selected on a 2% agarose gel, extracted from the gel, and column purified. Successfully ligated DNA fragments that contained adapter sequences were enhanced via PCR using adapter-specific primers. The DNA was re-isolated using AMPure XP Beads (Beckman) and the average fragment sizes of the libraries were assessed by an Agilent Bioanalyzer 2,100 to check for a sharp peak in the expected 500-600 bp range. Each library was loaded onto the HiSeq2000 platform and subjected to high-throughput sequencing to ensure that each sample met the desired average sequencing depth. The Illumina pipeline with default settings performed the image analysis and base calling.

Mapping short reads, variation calling and annotation

To map the short reads, the bovine genome assembly UMD 3.1 (Zimin et al., 2009) was used as a reference assembly. In this study, sequence scaffolds assigned to unknown chromosomes were included and no repeat masker was applied to the assembly. Sequences passing through the standard Illumina Chastity filter were retained for further analysis. Furthermore, sequence reads were first trimmed to 90 bp, as there are normally more sequence errors at the very beginning or the end of the reads. Low quality reads were also removed. For short-read mapping, we used BWA ver. 0.5.9 (Li et al., 2009). After mapping, we discarded the reads with mapping quality = 0 and unmapped reads. To call SNPs and InDels, we used SAMtools (Li et al., 2009) and additional filters as follows: (1) SNPs and InDels with an overall quality less than 20 were removed; (2) variants with too low or too high read depths were removed. First, we calculated the mean and standard deviation read depth for all the variants. We then set the minimum as 10% of the mean and the maximum as the mean read depth + 3 times the standard deviation; (3) variants with less than one forward or reverse alternative allele were removed; (4) variants within 5 bp of each other were removed; (5) SNPs within 5 bp of an InDel were removed; (6) InDels within 10 bp of each other were removed; (7) variants with no sites in the reference genome were removed. After SNP and InDel calling, NGS-SNP (Grant et al., 2011) was used to assign a functional class to each variant and to provide several fields of information describing the affected transcript and protein, if applicable. The source databases used during the annotation included Ensembl release 68, Entrez Gene, NCBI and UniProt (Flicek et al., 2011; Sayers et al., 2012; Magrane and Consortium, 2011).

Validation of the detected SNPs and InDels

To validate SNP calling from whole-genome resequencing (WGS) of the Chikso genome, we computed the genotype concordance between the WGS genotypes and SNP panel genotype data. The same sequenced genomic DNA was genotyped using Illumina's BovineSNP50 v2 BeadChip. The BovineSNP50 v2 BeadChip features 54,609 SNP probes that uniformly span the entire bovine reference genome. A small number of panel

Fig. 1. Morphological characteristics of Chikso cattle: (A) a picture of the Chikso used in this study, sampled at the Gangwon Provincial Livestock Research Center. (B) A picture of the front face of the Chikso

SNPs (1.3%) were excluded from the comparison because their locations on the genome were not known or they had alleles that were incompatible with those detected by sequencing. SNPs successfully genotyped using the BovineSNP50 Bead-Chip and that were not homozygous for the reference alleles were compared to SNPs derived from the sequencing. Genotype concordance was evaluated by two measures: genotype concordance at variant sites and non-reference sensitivity. Genotype concordance at variant sites is calculated by dividing the number of concordant non-reference genotypes (dark gray cells in Table 4) by all non-reference genotypes (dark and light gray cells in Table 4): $(13,506 + 12,741) / (13,506 + 23 + 21 + 12,741) * 100 = 99.9\%$. Non-reference sensitivity measures the rate at which non-reference sites in the genotyping panel data are recovered in the WGS data. It is computed by dividing the number of non-reference genotypes (dark and light gray cells in Table 4) by the number of WGS SNPs present on the chip (sum of A/B and B/B in the “WGS genotype” column of Table 4): $(13,506 + 23 + 21 + 12,741) / (13,565 + 12,782) * 100 = 99.8\%$.

Ten putative InDels ranging in length from 3 to 15 bp were validated by Sanger sequencing. Following the design of primer sets to amplify each candidate (Supplementary Table 1), PCR was performed in a 20- μ l volume containing 10 pmol of each primer, 0.25 mM of each dNTP, 2 μ l 10 $\times$ PCR buffer, 1.25 U DNA polymerase (Genet Bio., Korea), and 50 ng genomic DNA. The thermal cycling conditions included an initial denaturation for 10 min at 94°C; followed by 35 cycles of 30 s at 94°C, 30 s at 60°C or 64°C, and 1 min at 72°C; with a final 10-min extension at 72°C in a Veriti 96 well Thermal cycler (Applied Biosystems, USA). To detect differences in the nucleotide sequences, direct sequencing of the PCR products was performed using a Big Dye Terminator Cycle Sequencing Ready Reaction Kit V3.0 (Life Technologies Corp., USA) and an ABI PRISM® 3730 Genetic Analyzer (Life Technologies Corp.). The sequences were compared to find InDels using the SeqMan program (DNASTAR Inc., USA).

RESULTS

Massively parallel sequencing of the Chikso genome

The DNA extracted from the selected Chikso individual was determined as high quality (1.78 and 2.22 for the 260/280 and A260/230 nm ratio values, respectively), and was used to construct a paired-end library. The Illumina HiSeq 2000 sequenc-

ing platform was then used to massively parallel sequence the Chikso individual, generating 525,323,524 short reads of 100 bp. To detect reliable variations, strict quality checking of the 100 bp reads was performed to remove error-prone regions at both ends of each read. As a result, 2 bp and 8 bp trimming at the beginning and the end of each read, respectively, were applied to all the reads. The remaining 90 bp reads were further filtered by custom filtering steps, including removal of redundancy to generate higher quality reads for subsequent mapping. In total, 79.71% (418,730,058 of 90 bp paired-end reads) of the initial total reads were retained and 98.8% of them were successfully mapped to the *Bos taurus* reference sequence assembly (UMD 3.1) using BWA version 0.5.9 (Li et al., 2009). As a result, 98.81% of the reference genome sequence was covered, with an average mapping depth of 25.25-fold, which is sufficient to detect reliable SNPs and InDels.

Identification of SNPs and InDel

The 5,874,026 SNPs and 551,363 InDels were identified across all 29 bovine autosomes and the X-chromosome using SAMtools (Li et al., 2009). Approximately 45% (2,630,162 SNPs) of the detected SNPs were novel. A higher proportion (approximately 75%) of the InDels was novel when compared against dbSNP build 133. Among the total SNPs and InDels, the homozygous and heterozygous ratios were 1:1.92 (2,014,115 versus 3,859,911 SNPs) and 1:1.27 (242,843 versus 308,520 InDels), respectively. Among the InDels, 270,665 are insertions in comparison with the bovine reference sequence. We also estimated the transition (TS) versus transversion (TV) ratio of all the detected SNPs as 2.24:1, which indicates the quality of our detected SNPs. The TS:TV ratio value is similar to the ratios (e.g. 2.1:1) reported elsewhere (Abecasis et al., 2012). All the SNPs and InDels detected in this study were submitted in variant calling format (VCF) to the dbSNP database under the handle name ‘AGL_CJW’. The SNPs from the Chikso bull were systematically compared with SNPs identified through WGS of individuals from diverse cattle breeds, such as Fleckvieh (approximately 2.4 million SNPs), Black Angus (approximately 3.2 million SNPs) and Kuchinoshima-Ushi (approximately 6.3 million SNPs) (Eck et al., 2009; Kawahara-Miki et al., 2011; Stothard et al., 2011). The overlapping SNPs between the Chikso and the other breeds were 1,239,222, 1,638,171, and 2,269,041 for the comparisons of Chikso vs. Fleckvieh, Chikso vs. Black Angus, and Chikso vs. Kuchinoshima-Ushi, respectively (Fig. 2).

Fig. 2. Venn diagram showing the number of shared SNPs between Chikso, Fleckvieh, Black Angus, and Kuchinoshima-Ushi cattle

Fig. 3. Length distribution of deletions and insertions detected in this study

Although there are substantially more overlapping SNPs between the Chikso and Kuchinoshima-Ushi, which is expected to be genetically closer to Chikso, we cannot rule out that this higher overlap is partly explained by the higher numbers of SNPs detected in Kuchinoshima-Ushi, as well as by differences in the sequencing platforms and filtering parameters applied in each study.

Validation of the putative SNPs and InDels

To evaluate the SNP calling from our high-throughput genome sequencing data, concordance analysis was used to compare SNPs obtained from whole-genome resequencing (WGS) and from a SNP genotyping panel. The same genomic DNA from the Chikso bull used for the deep resequencing was genotyped for 54,609 SNPs using the BovineSNP50 BeadChip (Illumina). All probe sequences were mapped against the UMD 3.1 reference genome, and 53,872 sites (98.6%) were identified as valid SNPs. Among the SNP chip probes, we excluded 553 SNPs with unknown locations in the reference genome and 184 SNPs with alleles incompatible with the WGS SNPs. The call rate on

the chip was 99.8% for all valid SNPs. Only 115 probes failed to yield a genotype on the chip (Table 4). In total, 12,741 (99.7%) of 12,782 homozygous variant genotypes (B/B) called by WGS SNPs were identified as homozygous variants by chip SNPs, and 13,506 (99.6%) of 13,565 heterozygous genotypes (A/B) called by the WGS SNPs were identified as heterozygous genotypes by chip SNPs (Table 4). We evaluated the genotype concordance using two measures: genotype concordance at variant sites and non-reference sensitivity (see “Materials and Methods”). Genotype concordance at variant sites measures the overall accuracy of variant genotype calls, and was found to be 99.9% in this study. Non-reference sensitivity is the rate at which non-reference sites in the genotyping panel data are recovered in the WGS data. The non-reference sensitivity was 99.8%. Thus, we conclude that almost all variants were correctly called by WGS genotyping. Such high concordance of WGS SNP genotyping and chip SNP genotyping suggested that the WGS-based SNP genotypes used in this study contain few genotyping errors. For InDel validation, we selected 10 candidate InDels for characterization by capillary sequencing. The

Table 1. Functional class and the novelty status of the identified SNPs and InDels

SNP		InDel	
3 prime UTR variant	11,785	3 prime UTR variant	1,348
5 prime UTR variant	1,906	5 prime UTR variant	116
Coding sequence variant	28	INTERGENIC	366,341
Downstream gene variant	173,885	Coding sequence variant	66
Initiator codon variant	32	Downstream gene variant	17,539
Intergenic variant	3,934,208	Frameshift variant	514
Intron variant	1,511,327	Inframe deletion	86
Mature miRNA variant	32	Inframe insertion	70
Missense variant	16,273	Intron variant	145,767
Cc transcript variant	5	Mature miRNA variant	10
Non coding exon variant	1,810	Missense variant	14
Splice acceptor variant	94	Nc transcript variant	5
Splice donor variant	97	Non coding exon variant	111
Splice region variant	3,630	Splice acceptor variant	32
Stop gained	156	Splice donor variant	34
Stop lost	12	Splice region variant	354
Stop retained variant	10	Stop gained	1
Synonymous variant	22,086	Upstream gene variant	18,955
Upstream gene variant	196,650		
Fully known	3,243,864	Fully known	142,562
Novel	2,630,162	Novel	534,510
Partially known	0	Partially known	0
Total	5,874,026	Total	551,363

expected length, based on WGS, ranged from 3 to 15 bp. Seven of the 10 InDels gave Sanger sequencing results that were consistent with the alleles reported by SAMtools.

Annotation of SNPs and potential implication with traits of interest in cattle

To assign potential functional roles to the putative variations, further extensive annotation was performed on each of the detected SNPs and InDels (Table 1). The functional class terms used in the annotation are a subset of the variation terms used by Ensembl 68 (Flicek et al., 2012). The overlapping functional class terms ascribed to both SNPs and InDels were 3' UTR, 5' UTR, coding sequence, downstream gene, intergenic, intron, mature miRNA, missense, non-coding exon, splice acceptor, splice donor, splice regions, stop gained and upstream gene. Annotated functional classes unique to SNPs were initiator codon, stop lost, stop retained, and synonymous, while frameshift, inframe deletion, and inframe insertion were only assigned to InDels. We identified substantial numbers of SNPs and InDels across all 29 autosomes and the X-chromosome. Of SNPs, 92.7% were located in intergenic and intronic regions (3,934,208 intergenic and 1,511,327 intronic) and 92.9% of InDels were located in intergenic and intronic regions (366,341 intergenic and 145,767 intronic). Many non-synonymous SNPs, such as missense and stop gained mutations, were detected in this study: 16,273 SNPs (in 7,111 Ensembl genes) were found to be missense mutations, a few of which may influence phenotypic variation in economically important traits in cattle.

DISCUSSION

In this study, we performed whole-genome sequencing using the Illumina HiSeq 2000 sequencing platform on a Korean native cattle breed, Chikso, which is threatened with extinction in the Korea peninsula. Chikso has suffered from a limited population size; therefore, selecting an individual Chikso was carefully performed such that an animal was sequenced that properly represented the breed. An individual Chikso bull that was bred and protected in Gangwon Provincial Livestock Research Center was chosen because it had a proper pedigree and phenotypic records and an influential animal as a sire to be used for artificial insemination throughout the population.

Following the generation of short reads by the sequencing reaction, strict custom filtering criteria for better quality reads were applied, leading to higher mappability (98.80% mapped and 94.80% properly paired), suggesting that reliable variations could be identified using this approach. Despite partial loss of coverage depth caused by the strict criteria, 837,460,116 filtered reads were obtained, corresponding to ~25.3-fold coverage, which would be sufficient to call reliable putative genetic variations in the genome. SNP validation using the BovineSNP50 BeadChip showed a high genotype concordance rate (above 99.8%), while the InDel validation by capillary sequencing demonstrated a 70% concordance rate, which is similar to that calculated in a previous study (Levy et al., 2007). From all 29 autosomes and the X-chromosome, we identified more than 5.8 million SNPs and 0.55 million InDels (Tables 2 and 3), of which approximately 45% and 75% were novel. The larger

Table 2. Summary of the putative SNPs detected in this study grouped by chromosome

BTA	3' UTR	5' UTR	Cod-ing	Down	Ini.	Intergen	Intron	Mi RNA	Missen	Nc Trans	Non Code	SA	SD	SR	SG	SL	SR2	Syn	Up
1	532	84	1	8394	0	273513	90325	2	578	1	83	5	6	157	5	1	0	855	9121
2	551	74	3	6311	2	197328	75153	3	577	0	72	3	1	163	5	0	0	1003	7244
3	758	117	1	9018	3	169648	72482	1	866	0	60	1	3	176	6	0	0	1037	10734
4	523	77	2	7089	0	172699	88690	1	618	88	5	1	136	8	1	1	0	808	7998
5	678	98	1	8334	1	165394	76599	0	813	0	87	3	4	208	12	1	0	1127	10120
6	426	62	0	5595	1	196520	66785	3	405	0	71	6	6	98	5	0	0	597	5592
7	547	101	1	9776	0	182814	57025	0	963	0	78	4	6	173	16	0	0	1278	11591
8	383	73	0	6430	1	175890	58505	0	509	1	63	2	3	100	5	1	0	658	6435
9	183	37	0	4200	0	164180	53373	0	412	74	1	2	92	6	1	1	0	484	4160
10	423	70	2	7867	0	146985	69368	0	777	0	79	6	4	157	6	1	0	967	9247
11	592	93	2	7183	1	150153	65288	1	616	1	81	2	6	197	1	0	0	963	8096
12	176	31	1	3143	1	176535	41419	0	234	0	37	1	3	67	3	0	0	398	3235
13	597	57	1	6865	0	121972	56444	0	592	62	5	2	145	10	1	1	0	756	6771
14	300	51	1	3960	0	131745	45580	0	283	0	42	1	2	74	2	0	1	467	4364
15	434	82	1	14034	2	140453	46879	2	1350	0	239	3	6	141	9	1	0	1470	16014
16	283	50	0	4144	1	116974	48678	2	461	0	29	2	5	88	2	0	0	595	4634
17	395	50	0	4018	2	128514	38953	0	361	0	49	3	3	103	5	0	0	599	4490
18	455	110	3	7019	5	85518	40661	0	887	0	60	8	7	169	13	1	0	1084	8666
19	720	100	1	8118	2	68749	53192	1	871	0	44	2	5	259	7	0	0	1342	9952
20	182	25	0	3066	0	129768	34752	2	217	0	48	3	2	61	2	0	2	343	3021
21	279	46	0	4400	1	116690	38065	3	405	0	41	4	5	96	5	3	0	625	5429
22	293	43	3	3591	3	76314	47949	1	345	1	45	1	3	86	1	0	1	550	3907
23	565	149	0	8622	2	85154	41482	2	951	0	42	3	2	178	5	0	2	1164	10354
24	242	7	1	3070	0	109065	35085	2	241	0	54	1	2	57	1	0	0	309	3042
25	371	54	1	3693	1	52827	31554	1	449	0	13	2	0	136	3	0	0	655	4621
26	179	22	0	2454	0	74471	33074	0	261	0	23	4	2	68	1	0	1	344	2970
27	160	43	0	1954	1	80186	21005	1	165	0	28	0	1	41	2	0	0	293	2386
28	168	18	0	3161	0	70174	37239	1	221	1	37	4	0	68	2	0	0	358	3258
29	257	59	1	5318	2	90272	29541	0	585	0	39	7	4	96	6	0	0	709	6350
X	133	23	1	3058	0	83703	16182	3	260	0	42	2	1	40	2	0	0	248	2848
All	11785	1906	28	173885	32	3934208	1511327	32	16273	229	1597	88	465	3281	135	12	7	22086	196650

Abbreviations in column titles are: BTA, *Bos taurus* autosome; 3'UTR, variants in the 3' UTR; 5'UTR, variants in the 5' UTR; Coding, variants in the coding sequence; Down, variants within 5 kb downstream of the 3' end of a transcript; Ini., variants in the initiator codon; Intergen, variants in the intergenic region; Intron, variants in an intron; miRNA, variants in a mature miRNA sequence; Missen, missense variants; Nc Trans, variants in a non-coding transcript; Non Code, variants in a non-coding exon; SA, variants in a splice acceptor; SD, variants in a splice donor; SR, variants in a splice region; SG, variants creating a stop codon; SL, variants abolishing a stop codon; SR2, synonymous variants in a stop codon; Syn, synonymous variants; Up, variants within 5 kb upstream of the 5' end of a transcript.

proportion of novel InDels could in part be because most of the recent genome sequencing studies using NGS in cattle reported SNPs rather than InDels. The proportion of InDels detected in this study only accounts for approximately 8.6% of all events, including SNPs. However, the variant bases in the InDels involve approximately 19.1% of all variant bases, suggesting that InDels may be an important source of both genomic and phenotypic diversity. The lengths of the InDels ranged from 30 (insertion) to -48 (deletion); however, most InDels were short: approximately 73% of insertions and 70% of deletions were less than 3 bp (Fig. 3), which is similar to previous results (Kawahara-Miki et al., 2011).

The proportion of novel SNPs is lower (~45%) than previous studies, such as 82%, 81%, and 87% from sequencing bulls from the Fleckvieh, Holstein, Black Angus, and Kuchinoshima-Ushi breeds, respectively (Eck et al., 2009; Kawahara-Miki et al., 2011; Stothard et al., 2011). The lower proportion may be largely accounted for by recent SNP depositions from these and other previous studies of diverse cattle breeds. However, despite the lower proportion of novel SNPs, this result clearly suggests that large numbers of SNPs remain to be discovered by sequencing multiple individuals and more diverse cattle breeds. Furthermore, extensive comparisons of the SNPs in this study were made against SNPs obtained from European

Table 3. Summary of putative InDels detected in this study grouped by chromosome

BTA	3' UTR	5' UTR	Intergen	Coding	Down	FS	ID	II	Intron	miRNA	Missen	Nc Tran	Non Code	SA	SD	SR	SG	Up
1	63	5	26143	4	894	20	5	1	8990	1	1	0	10	1	2	12	0	863
2	58	3	18839	2	704	25	1	3	7634	2	0	0	1	4	1	17	0	726
3	85	7	15365	4	888	23	3	5	7244	1	1	0	3	0	1	17	1	1035
4	54	7	16652	4	744	15	3	4	8717	0	1	0	5	1	1	15	0	729
5	92	8	15202	4	878	31	3	2	7571	0	1	0	0	0	2	14	0	1001
6	65	8	18911	0	591	12	3	5	6633	0	1	0	4	4	1	26	0	601
7	58	3	16940	2	1023	33	4	4	5612	1	0	0	7	1	2	12	0	1154
8	53	6	16515	2	629	12	2	1	5711	0	0	1	4	3	0	11	0	604
9	21	4	16005	2	422	13	0	3	5252	0	1	0	2	2	1	10	0	443
10	49	8	13424	2	810	20	5	5	6955	0	1	0	4	3	0	17	0	924
11	59	4	13687	2	702	21	2	1	6235	1	0	0	5	1	1	27	0	708
12	19	2	16568	0	339	6	3	0	3875	0	1	0	1	0	0	3	0	341
13	54	4	10678	1	703	17	4	3	5167	0	1	1	5	0	1	15	0	652
14	46	5	12331	2	467	8	3	1	4439	1	0	0	4	1	0	9	0	452
15	61	4	12854	2	1251	36	5	3	4504	1	3	0	14	0	3	12	0	1462
16	47	7	10594	3	455	22	4	4	4826	0	1	0	2	0	0	13	0	442
17	30	1	12208	4	438	9	4	1	3477	0	0	0	4	1	0	6	0	446
18	54	2	7600	3	688	30	5	2	3742	1	0	0	8	4	2	17	0	764
19	64	8	5883	5	755	24	7	5	4901	0	0	0	4	0	2	16	0	922
20	23	3	12227	2	352	5	0	0	3326	1	0	0	2	1	3	7	0	327
21	25	3	10565	3	416	11	1	2	3419	0	0	0	3	2	0	4	0	474
22	36	3	7022	2	344	13	2	2	4416	0	0	0	2	1	0	8	0	376
23	60	5	7494	5	796	18	6	2	3919	0	0	0	5	0	1	18	0	1019
24	24	2	9919	0	311	7	1	1	3215	0	0	1	2	0	0	4	0	287
25	37	0	4483	0	322	18	5	2	2596	0	1	0	0	0	5	11	0	403
26	25	1	6749	0	240	10	1	3	3217	0	0	0	3	0	1	5	0	313
27	22	1	7550	0	217	5	1	0	2100	0	0	0	0	0	0	5	0	242
28	21	0	6436	3	326	9	1	2	3635	0	0	2	2	1	4	5	0	306
29	21	1	7937	2	502	27	0	1	2547	0	0	0	3	1	0	12	0	592
X	22	1	9560	1	332	14	2	2	1892	0	0	0	2	0	0	6	0	347
All	1348	116	366341	66	17539	514	86	70	145767	10	14	5	111	32	34	354	1	18955

Abbreviations are the same as in Table 2 except: FS, variants causing a frameshift; ID, variants causing a deletion; II, variants causing an insertion.

and Asian *Bos taurus* cattle breeds. The results showed a higher number of overlapping SNPs, particularly in the comparison with Kuchinoshima-Ushi. This phenomenon may reflect the fact that Kuchinoshima-Ushi is a Japanese indigenous breed that is geographically closer to the Korea peninsula. However, we must be cautious in concluding that our results imply a closer genetic relationship between the Chikso and the Japanese native breed, because the SNPs used for the between-breed comparisons were identified with different sequencing platforms, sequencing coverage, and parameters applied to call variants, leading to different numbers of total SNPs. Thus further investigations, preferably using similar experimental methods, will be required to clearly dissect the genetic relationships between these diverse cattle breeds.

Throughout all 29 autosomes, the numbers of detected variations within each chromosome were proportional to the chromosome length (Tables 2 and 3), with a range of 0.21–0.26%

for SNPs and 0.018–0.025% for InDels. However, there was considerably less variation observed for the X-chromosome, with 0.07% variation in SNPs and 0.008% variation in InDels, compared with the autosomes. These results are in line with our expectation, which is supported by previous studies showing a smaller population size and lower mutation rate on the X-chromosome compared with autosomes (Li et al., 2002; Makova et al., 2002). As for the homozygous and heterozygous ratio of the detected SNPs, we did not observe a distinctly lower homozygous and heterozygous ratio (1:1.92) for the Chikso animal. This result is somewhat surprising, because Chikso has been regarded as an endangered cattle breed in Korea, and a small population is expected to show more homozygosity caused by potentially higher rates of inbreeding. We additionally determined the homozygous and heterozygous ratio for a recently sequenced Japanese native breed, Kuchinoshima-Ushi, using its complete SNPs retrieved from dbSNP. The re-

Table 4. Genotype concordance between whole-genome resequencing and the BovineSNP50 SNP chip

Chip genotype	No. of chip SNPs	WGS genotype			
		A/B		B/B	
A/A	25,866	19	(0.1%)	2	(0.0%)
A/B	14,571	13,506	(99.6%)	23	(0.2%)
B/B	13,320	21	(0.2%)	12,741	(99.7%)
. / .	115	19	(0.1%)	16	(0.1%)
Total	53,872	13,565		12,782	

A, reference allele; B, non-reference (alternative) allele; and '.', no call

Dark gray cells indicate the concordant non-reference genotypes. Light gray cells indicate the discordant non-reference genotypes.

sult shows a ratio (1:1.2) which is lower than Chikso in this study. This difference could reflect the fact that Kuchinoshima-Ushi has long been isolated on a small Kuchinoshima Island, and still in the highly inbred condition, potentially leading to a higher degree of homozygosity (Kawahara-Miki et al., 2011). In addition, Dadi et al. (2012) recently showed that Chikso has a similar genetic diversity to Hanwoo (Korean brown cattle), based on an analysis of mitochondrial DNA. The population size of Chikso has been reduced partly by the policy to unify coat colors since the beginning of the 20th century in Korea. Thus, despite recent decreases in the population size, we may postulate that Chikso has maintained a similar genetic diversity to cattle breeds with larger population size, such as Hanwoo. This idea will need to be interrogated by further studies at the population level preferentially including multiple cattle breeds.

To evaluate the potential functional roles of the detected variations, they were extensively annotated. A large number of missense SNPs (16,273 SNPs in 7,111 genes), were identified, some of which may affect phenotypic variation in cattle or account for some of the notable characteristics of the Chikso breed. For example, some of the SNPs were detected in pigmentation-related genes, such as tyrosinase (*TYR*), tyrosinase-related protein 1 (*TYRP1*) and dopachrome tautomerase (*DCT*); however, no SNPs were detected in the melanocortin 1 receptor (*MC1R*) gene in this work (nucleotide positions 6461851 in *Bos taurus* autosome (BTA) 29 as G > A, 31717680 in BTA8 as T > C, 69544299 in BTA12 as G > C for TYR, TYRP1, and DCT respectively). Coat color depends on the relative amount of pheomelanin and eumelanin, and the bridling coat pattern found in Chikso requires at least one wild-type *MC1R* without any dominant allele to the wild-type (Klungland et al., 1995; Seo et al., 2007). Coat color and its pattern are polygenic traits whose underlying genetic mechanisms remain to be determined; therefore, further research is warranted to dissect the genetics of coat color and pattern by comparing multiple individuals in diverse cattle breeds. As another example, candidate SNPs were detected in the fatty acid synthase (*FASN*) and acetyl-CoA carboxylase alpha (*ACACA*) genes on BTA19 (nucleotide positions 51394090 as A > G and 51402032 as G > A for *FASN*, and 13915963 as G > C for *ACACA*), which are thought to be associated with fatty acid compositions (de Souza et al., 2012; Zhang et al., 2010). Recently, *FASN* was reported to be significantly associated with fatty acid composition in Hanwoo steers (Yeon et al., 2013). Hanwoo exhibit a higher ratio of monounsaturated fatty acids in their intramuscular fat than other breeds (Kim et al., 2005; Smith et al., 2009). Although it is beyond the scope of this study to conclude that Chikso also have a genetic potential to show different monosaturated fatty acid ratios, the candidate SNPs provided in this study could be

a valuable resource to further dissect the genetic dynamics associated with traits of interest in cattle.

In this study, we massively parallel sequenced a Korean native cattle breed, Chikso, and successfully identified substantial numbers of SNPs and InDels throughout the genome. The potential functional roles of each of the detected variations were assessed by extensive annotations. We are aware that only an individual animal has been sequenced in this study; therefore, further studies will be required to clarify how genetic variations are associated with traits of interest using multiple individuals, and to more completely characterize the variation present in this breed. However, despite this limitation, our findings provide valuable genomic information to further develop more accurate genomic tools to dissect the genetic mechanisms underlying phenotypic differences in cattle.

Note: Supplementary information is available on the Molecules and Cells website (www.molcells.org).

ACKNOWLEDGMENTS

The authors thank Dr. Stephen Miller for his critical reading of the manuscript. This work was supported by a grant from the Next-Generation BioGreen 21 Program, Rural Development Administration, Korea (grant#: PJ008196, PJ008028); Xiaoping Liao is funded by the Genome Canada project entitled "Whole Genome Selection Through Genome Wide Imputation in Beef Cattle."

REFERENCES

- Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56-65.
- Bovine Genome Sequencing Analysis Consortium, Elsik, C.G., Tellam, R.L., Worley, K.C., Gibbs, R.A., Muzny, D.M., Weinstock, G.M., Adelson, D.L., Eichler, E.E., Elntski, L., et al. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 324, 522-528.
- Bovine HapMap Consortium, Gibbs, R.A., Taylor, J.F., Van Tassel, C.P., Barendse, W., Eversole, K.A., Gill, C.A., Green, R.D., Hamernik, D.L., Kappes, S.M., et al. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324, 528-532.
- Choi, T.J. (2009). Establishment of phylogenomic characteristics for Korean traditional cattle breeds (Hanwoo, Korean brindle and black). Doctoral Thesis. Jeon-buk National University, Republic of Korea.
- Dadi, H., Lee, S.H., Jung, K.S., Choi, J.W., Ko, M.S., Han, Y.J., Kim, J.J., and Kim, K.S. (2012). Effect of population reduction on mtDNA diversity and demographic history of Korean Cattle populations. *AJAS* 25, 1223-1228.

- de Souza, F.R., Chiquitelli, M.G., da Fonseca, L.F., Cardoso, D.F., da Silva Fonseca, P.D., de Camargo, G.M., Gil, F.M., Boligon, A. A., Tonhati, H., Mercadante, M.E., et al. (2012). Associations of FASN gene polymorphisms with economical traits in Nellore cattle (*Bos primigenius indicus*). *Mol. Biol. Rep.* **39**, 10097-10104.
- Eck, S.H., Benet-Pages, A., Flisikowski, K., Meitinger, T., Fries, R., and Strom, T.M. (2009). Whole genome sequencing of a single *Bos taurus* animal for single nucleotide polymorphism discovery. *Genome Biol.* **10**, R82.
- FAO (Food and Agriculture Organization). (2012). Domestic Animal Diversity Information Service (DAD-IS). <http://dad.fao.org/> Accessed December 20, 2012.
- Fliscek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2012). Ensembl 2012. *Nucleic Acids Res.* **40**, D84-90.
- Grant, J.R., Arantes, A.S., Liao, X., and Stothard, P. (2011). In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. *Bioinformatics* **27**, 2300-2301.
- Jo, C., Cho, S.H., Chang, J., and Nam, K.C. (2012). Keys to production and processing of Hanwoo beef: a perspective of tradition and science. *Animal Frontiers* **2**, 32-38.
- Kawahara-Miki, R., Tsuda, K., Shiwa, Y., Arai-Kichise, Y., Matsumoto, T., Kanesaki, Y., Oda, S., Ebihara, S., Yajima, S., Yoshikawa, H., et al. (2011). Whole-genome resequencing shows numerous genes with nonsynonymous SNPs in the Japanese native cattle Kuchinoshima-Ushi. *BMC Genomics* **12**, 103.
- Kim, K.H., Lee, J.H., Lee, S.C., Park, W.Y., Oh, Y.G., Kang, S.W., and Ko, Y.D. (2005). The optimal TDN levels of concentrates and slaughter age in Hanwoo steers. *J. Anim. Sci. Technol.* **47**, 731-744.
- Klungland, H., Vage, D.I., Gomez-Raya, L., Adalsteinsson, S., and Lien, S. (1995). The role of melanocyte-stimulating hormone (MSH) receptor in bovine coat color determination. *Mamm. Genome* **6**, 636-639.
- Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. (2007). The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254.
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760.
- Li, W.H., Yi, S., and Makova, K. (2002). Male-driven evolution. *Curr. Opin. Genet. Dev.* **12**, 650-656.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079.
- Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database* **2011**, bar009.
- Makova, K.D., and Li, W.H. (2002). Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**, 624-626.
- NIAS (National Institute of Animal Science). (2012). The status of local livestock breeds in Korea, registered in DAD-IS. <http://www.nias.go.kr/> Accessed December 20, 2012.
- Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetverin, V., Church, D.M., Dicuccio, M., Federhen, S., et al. (2012). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **40**, D13-25.
- Seo, K., Mohanty, T.R., Choi, T., and Hwang, I. (2007). Biology of epidermal and hair pigmentation in cattle: a mini-review. *Vet. Dermatol.* **18**, 392-400.
- Smith, S.B., Gill, C.A., Lunt, D.K., and Brooks, M.A. (2009). Regulation of fat and fatty acid composition in beef cattle. *AJAS* **22**, 1225-1233.
- Stothard, P., Choi, J.W., Basu, U., Sumner-Thomson, J.M., Meng, Y., Liao, X., and Moore, S.S. (2011). Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. *BMC Genomics* **12**, 559.
- Yeon, S.H., Lee, S.H., Choi, B.H., Lee, H.J., Jang, G.W., Lee, K.T., Kim, K.H., Lee, J.H., and Chung, H.Y. (2013). Genetic variation of FASN is associated with fatty acid composition of Hanwoo. *Meat Sci.* **94**, 133-138.
- Zhang, S., Knight, T.J., Reecy, J.M., Wheeler, T.L., Shackelford, S.D., Cundiff, L.V., and Beitz, D.C. (2010). Associations of polymorphisms in the promoter I of bovine acetyl-CoA carboxylase-alpha gene with beef fatty acid composition. *Anim. Genet.* **41**, 417-420.
- Zimin, A.V., Delcher, A.L., Florea, L., Kelley, D.R., Schatz, M.C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C.P., Sonstegard, T.S., et al. (2009). A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol.* **10**, R42.