BMB
Reports

Genetic variants and signatures of selective sweep of Hanwoo population (Korean native cattle)

Taeheon Lee[1], Seoae Cho[2], Kang Seok Seo[3], Jongsoo Chang[4], Heebal Kim[1,5,*] & Duhak Yoon[6,*]

[1]Department of Agricultural Biotechnology and Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul 151-742, [2]C&K Genomics, Seoul 151-919, [3]Department of Animal Science and Technology, Sunchon National University, Suncheon, 540-742, [4]Department of Agricultural Science, Korea National Open University, Seoul 110-810, [5]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, [6]Department of Animal Science, Kyungpook National University, Sangju 742-170, Korea

Although there have been many studies of native Korean cattle, Hanwoo, there have been no selective sweep studies in these animals. This study was performed to characterize genetic variation and identify selective signatures. We sequenced the genomes of 12 cattle, and identified 15125420 SNPs, 1768114 INDELs, and 3445 CNVs. The SNPs, INDELs, and CNVs were similarly distributed throughout the genome, and highly variable regions were shown to contain the *BoLA* family and *GPR180*, which are related to adaptive immunity. We also identified the domestication footprints of the Hanwoo population by searching for selective sweep signatures, which revealed the *RCN2* gene related to BPV resistance. The results of this study may contribute to genetic improvement of the Hanwoo population in Korea. [BMB Reports 2013; 46(7): 346-351]

INTRODUCTION

Identification of the genetic differences responsible for variations in phenotypic traits is one of the goals of livestock genomic research. Accordingly, it is important to characterize the genetic variation in livestock species. Domestication processes and breeding development have been imprinted in the genomes and provide detectable clues of selection within the cattle genome (1, 2). Selective sweep signatures may have contributed to the domestication process (3, 4). Despite various Hanwoo (Korean native cattle) mapping and diversity studies

*Corresponding authors. Heebal Kim, Tel: +82-2-880-4803; Fax: +82-2-883-8812; E-mail: heebal@snu.ac.kr; Duhak Yoon, Tel: +82-54-530-1227; Fax: +82-54-530-1229; E-mail: dhyoon@knu.ac.kr

(5, 6), the identification of mutations affecting some quantitative phenotypes (7, 8), and several expression studies (9-11), selective sweep signatures in these animals have not been identified. There have been several attempts to conduct such characterization at the population level of other cattle by detecting selective sweep signatures (1, 2). Hanwoo were used as a draft type until about 30 years ago, whereas dairy cattle have been artificially selected for milk production for a long time. However, Hanwoo have only been bred for meat for a short period, and it is important to evaluate its genetic features to improve performance as beef cattle. Therefore, the present study was performed to characterize the entire genome sequence of Hanwoo Korean native cattle at the population level.

Genetic variations, including SNPs, insertions and deletions (INDELs), and large structural variations, such as CNV, shape the genome architecture and provide signatures for identification of variations contributing to adaptation. Large structural variations can be attributed to inversions and translocations, duplications, and deletions. Although existing at lower rates than SNPs, CNVs and INDELs comprise another important type of genomic variation (11) that can be used to better understand genome features.

Adaptation from standing genetic variation implies that there are neutral or weakly deleterious variations that are maintained with a long history in populations and that they become advantageous with changes in the environment (12). Experimental evolution, which tests hypotheses or theories of evolution using an experimentally controlled population, has provided evidence that adaptation from standing variation is more repeatable than that from new mutations (12). Standing variation could contribute to rapid adaptation after a sudden environmental change (13); for example, Rowan et al. showed experimentally that sticklebacks have sufficient standing genetic variation to adapt to a significant change in climate in only three generations (14). Such adaptations are influenced by the type and quantity of the available genetic variation (15). Therefore, a comprehensive description of variations in a population will lead to a deeper understanding of its biology.

To investigate the features of the Hanwoo genome, a ge-

nomic scan of the Korean cattle population was performed for selective sweep signatures and highly variable regions were identified.

RESULTS

Identification of genetic variants

To construct high-quality Hanwoo re-sequencing data, we generated pair-end reads using an Illumina HiSeq2000. After removing possible PCR duplicates, 97.12% of all the reads, which together corresponds to mapping of 37603284033 bp, each individual was successfully aligned against the bovine reference genome UMD3.1. On average, a depth of 14.2× was achieved and the mapped reads covered an average of 98.96% of the reference genome. The depth was calculated for each individual using the DepthOfCoverage.jar of GATK (16), with the range of depth of 12.61×-15.53× (Table S1) and coverage was calculated using BEDTools (17).

For the UMD3.1 reference assembly, a total of 15125420 SNPs were identified, and the change rate was 166 base pairs (Table S2). Of these SNPs, 27% were located in genic regions, while 73% were located in intergenic regions (Supplement Fig. 2). A total of 1768114 INDELs were identified, and the change rate was 1420 base pairs (Table S2). Of these INDELs, 30% were located in genic regions and 70% were located in intergenic regions (Supplement Fig. 2). The minimum quality of SNPs and INDELs was 30. The quality of variants and INDEL length are shown in Supplment Fig. 6 and 7.

In this study, CNVs were defined based only on deletion type using Genome STRiP (18). A total of 3445 CNVs (10.2 Mb), with an average length of 2.97 kb (range 0.23-902 kb) were identified, and the change rate was 729196 base pairs (Table S2). Overall, 17% of the CNVs were located in genic regions overlapping with at least one gene, if not covering an entire gene, while 83% were located in intergenic regions without overlapping any gene (Supplement Fig. 2). Evaluation of the length distribution of CNVs showed that the number of CNVs decreased with increasing size, except for the 1-3 kb region (Supplement Fig. 8). This exception corresponds to LINEs around 2,000 bp. This effect has also been identified in humans and cattle using the same method as used in this study and with other methods, respectively (18, 19). These findings indicated that the exception is not due to the analysis method or sample population.

Correlation among variants and repeat elements

We found that SNPs, INDELS, and CNVs had similar distributions throughout the genome, and that the number of variants was related to the length of the genome. We calculated the proportions of the three types of variants on each chromosome and found that chromosomal proportions of each type of variant were similar to those of chromosomal length (Supplement Fig. 1a) and that the number of variants had a strong linear relationship with chromosomal length (Supple-

ment Fig. 1b-d).

The number of variants at each 1 Mb bin and 10 Mb bin were similarly distributed throughout the genome (Fig. 1, Supplement Fig. 3). CNVs were relatively less similar to the other variants. As there was a small number of CNVs when using the large 10 Mb bin compared to the 1 Mb bin, the genomic distribution of the CNVs became similar to that of the other types of variants (Supplement Fig. 3).

Highly variable region

There are common regions in which many types of variants exist (Fig. 1). For example, the BoLA family is present in the first polymorphic common region, chr23:23-31 Mb, which is located near the centromere. The *BoLA* family is a type of cattle MHC, and has sufficient variation for immune defense. There are also many contigs around the 2nd highest peak of common region, chr12:69-77 Mb, which is near the centromere. Specifically, *GPC6, BIRC3, DCT, TGDS, GPR180, CLDN10, DZIP1, DNAJC3, HS6ST3, MBNL2,* and *RAP2A* are located in this region. The average nucleotide diversity was as high as 12.95 (Table S3), calculated by VCFtools (20). *BoLA* and *GPR180* are located in a highly variable region, containing sufficient variants, including synonymous SNPs, non-synonymous SNPs, and intronic SNPs (Supplement Fig. 4A, B). However, *RCN2* has only a small number of variants (Supplement Fig. 4C). SNPs and CNVs are enriched with MHC and a G protein-coupled receptor gene region, which has also been reported in other species (21).

A significant genome-wide correlation between the number of SNPs and INDELs was identified (Pearson's correlation r = 0.87, P < 0.05) (Supplement Fig. 5). SNPs and INDELs were also highly correlated with CNV, while low complexity and simple repeats were positively correlated with CNV, and the GC content was negatively correlated with CNV.

Selective sweep region

We estimated a folded SFS because no ancestral allele information was available. The SFS of the SNPs and INDELs were normal, while those of CNV showed balancing selection (Supplement Fig. 11). Linkage disequilibrium (LD)-based ω statistics and SFS-based Λ statistics available within the population were used to identify selective sweep regions (Fig. 2). As expected, SweepFinder and OmegaPlus detected different regions and showed strong signals. The highest significance regions of Λ statistics were chr2:72-72.5 Mb and chr10:35.5-36 Mb. The chr21:32-33 Mb, chr18:5.5-10.5 kb, and 15.9 Mb regions showed clear ω statistics signatures. The first significant Λ statistics region (chr2:72-72.5 Mb) included the genes *PTPN4, EPB41L5, TMEM185B, RALB,* and *INHBB,* while the genes *FSIP1, GPR176, SRP14, BMF,* and *PAK6* were located in the second significant Λ statistics region (chr10:35.5-36 Mb) (Table 1). The first significant ω statistics region (chr21:32-33 Mb) included the genes *ETFA, ISL2, RCN2, PSTPIP1,* and *TSPAN3,* while the *ITFG1* gene was located in the second sig-

nificant ω statistics region (chr18:15.9 Mb) (Table 1).

Combination statistics were used to identify common signals. The *RCN2* gene was located in the highest combination statistics region (chr21:32.5 Mb) between the ω statistics and Λ statistics region. This gene produces a calcium-binding protein located in the lumen of the ER that contains six conserved regions with similarity to a high affinity Ca^{+2}-binding motif, the EF-hand.

DISCUSSION

Highly variable region

The three types of variants were similarly distributed throughout the genome. Even if a type of variant was different, the quantity of the variant was affected by common effects, such as the mutation and recombination rates. These findings were supported by the results of a previous study in which SNP detection accuracy was shown to be affected by CNVs (19, 21). A significant genome-wide correlation between SNP and INDEL density was identified in a separate study (19).

We identified a hypervariable region near the centromere and showed that the mutations were related to adaptive immunity. Specifically, the bovine MHC class II region lies near the centromere of BTA23 (Fig. 1). Variation in the high recombination rate region may be due to the existence of a polymorphic recombination hotspot (22-24), and regions with high recombination rates were significantly closer to the centromere (25). Genetic standing variants facilitate the emergence of adaptation. The MHC region contains a diverse array of genes that are crucial for the initiation of adaptive immune

responses. All types of variants were enriched with the *BoLA* (bovine MHC) family in the highest peak region and *GPR* genes involved in the interaction with extracellular molecules in the second highest peak region (Fig. 1). Similar findings have been reported in other animals, e.g., the three-spined stickleback has many SNP and CNV variants in the MHC region (21), and genetic variations at the MHC loci are known to be involved in pathogen resistance (26). These regions are comprised of many contigs (Supplement Fig. 10) and the region of BTA23 has been confirmed experimentally (22-24). The cattle genome build UMD 3.1 has a low amount of erroneous duplication and error (27, 28); therefore, this was not caused by UMD 3.1 genome assembly.

Selective sweep genes

Generally, cattle have been marked by selection during domestication, breed formation, and ongoing selection to enhance performance and productivity. We utilized two methods to detect genomic selection in cattle: (i) the ω statistic (29, 30), which detects specific LD patterns caused by genetic hitchhiking, and (ii) the composite likelihood ratio (CLR) Λ (31) using the SFS, which describes the frequency of allelic variants and shifts from neutral expectation toward rare and high frequency derived variants.

We found evidence of selective sweeps on chromosomes 2, 10, 18, and 21 (Table 1 and Fig. 2). Among these regions, we identified commonly selected regions near *RCN2* (*E6BP*) by considering the ω and Λ values together, and this region had the highest correlation (r = 0.9983). This gene interacted with cancer-associated HPV E6 and with BPV-1 E6. The transforming
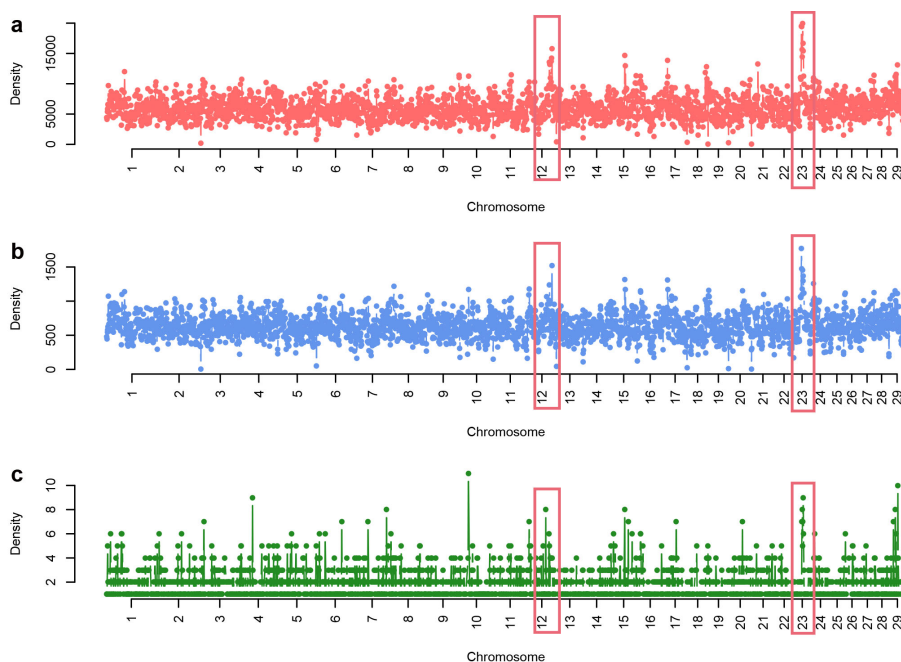

Fig. 1. Distribution of each variant. The x-axis of the plot shows the genomic position, and the y-axis represents the quantity of each variant within the 1 Mb bin. Panel (A) SNPs, panel (B) INDELs, and panel (C) CNVs.
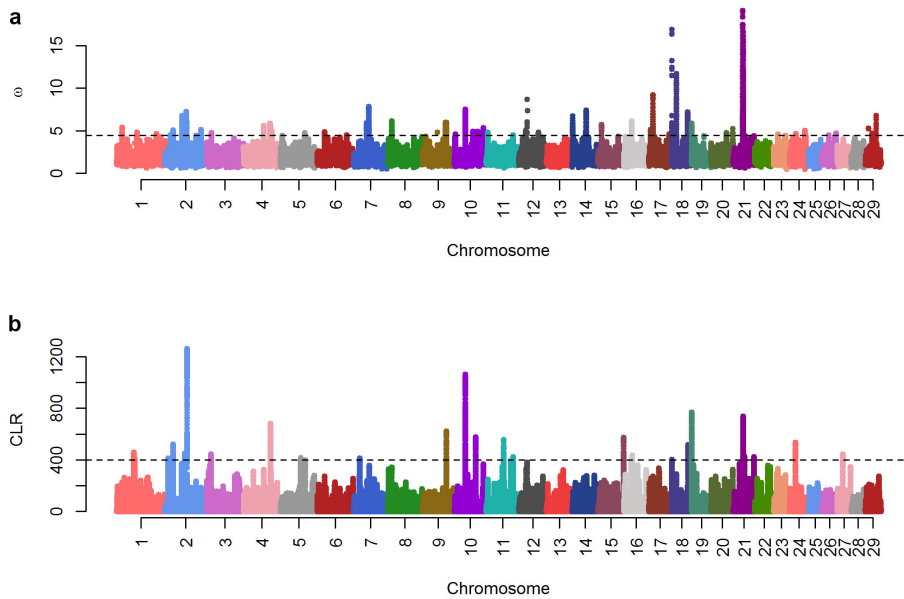
a

b

Fig. 2. Manhattan plots of selective sweep signature of the Hanwoo population: panel (A) ω statistics from Omega Plus; panel (B) Λ statistics (CLR) from SweepFinder. The x-axis of the Manhattan plot shows the genomic position, and the y-axis represents each statistic. The cutoff for the ω statistics and the Λ statistics was the 99.9% quartile of empirical distribution of each statistic.

Table 1. Selective sweep genes of CLR ratio and Omega statistics (significance level: empirical P value 0.001)

Symbol	Location (Mb)	CLR ratio
PTPN4	2:72.05-72.15	1106.470
EPB41L5	2:72.17-72.33	1190.173
RALB	2:72.36-72.43	709.583
INHBB	2:72.52	540.436
FSIP1	10:35.35-35.54	551.626
GPR176	10:35.56-35.69	851.561
SRP14	10:35.80	456.310
BMF	10:35.85-35.87	442.939
PAK6	10:35.98-36.01	537.912

Symbol	Location (Mb)	Omega
ITFG1	18:15.6-15.9	4.519
ETFA	21:31.9-32	9.607
ISL2	21:32-32.1	5.7917
RCN2	21:32.57-32.59	5.291
PSTPIP1	21:32.64-32.69	8.454
TSPAN3	21:32.7-32.72	12.280

activity of BPV-1 E6 mutants was correlated with their E6BP-binding ability (32). Calcium is required for entry into mitosis, and E6 may play a role in this stage of cell growth, indicating that *RCN2* is also important. The *RCN* gene has undergone a selective sweep, which may suggest that Korean cattle are resistant to BPV (33). In support of this suggestion, experimental evidence has been reported indicating that Korean native cattle have greater resistance to BPV than Holsteins (34).

Using the ω and Λ values, several genes were detected and we investigated whether this region was related with the cattle quantitative trait locus (QTL) except dairy cattle QTL information. Among the QTLs identified in the selective sweep region, the chr10:35.5-36 Mb region is related to the longissimus muscle area and the chr18:15.9 Mb region is related to carcass weight, longissimus muscle area, and social separation vocalization. Hanwoo was used as a draft type of cattle before 1980, but is now used for beef production; therefore, the QTLs of the selective sweep region are reasonable.

The results of the present study indicated that the distributions of SNPs, INDELs, and CNVs have a correlated pattern. We found that the selective sweep signatures of the Hanwoo genome and the highly variable region were related to adaptive immunity. We hope that these characteristics will contribute to genetic improvement and breeding of this strain. Future studies should include larger samples and various breeds and phenotypes to obtain better understanding of cattle genomes.

MATERIALS AND METHODS

Sample preparation and re-sequencing
Whole-blood samples were collected from 12 Hanwoo bulls from Kyungpook National University, Korea. Blood (10 ml) was drawn from the carotid artery and treated with heparin to prevent clotting. DNA was isolated from whole blood using a G-DEXTMIIb Genomic DNA Extraction Kit (iNtRoN Biotechnology, Seoul, Korea) according to the manufacturer's protocol. We randomly sheared 3 μg of genomic DNA using the Covaris System to generate inserts of ~300 bp. The fragments of sheared DNA were end-repaired, A-tailed, adaptor ligated, and amplified using a TruSeq DNA Sample Prep. Kit (Illumina, San Diego, CA). Paired-end sequencing was conducted by NICEM

(National Instrumentation Center for Environmental Management, Seoul, Korea) using the Illumina HiSeq2000 platform with TruSeq SBS Kit v3-HS (Illumina). Finally, sequence data were generated using the Illumina HiSeq system.

Sequence alignment and genotype calling

Pair-end sequence reads were mapped to the reference bovine genome (UMD 3.1) using Bowtie 2 with the default settings (35). Four open-source packages were used for downstream processing and variant calling for SNPs and INDELs: Picard tools (http://picard.sourceforge.net), SAMtools 0.1.18 (36), VCFtools 4.0 (20), and the Genome Analysis Toolkit 1.4 (16). Read Group was added and duplicate reads were removed using MarkDuplicates of Picard tools. SAMtools was used to index the resulting bam format files and to calculate the mapped read length with the flagstat option (36). Realignment and variant calling were performed using GATK (16) with Count-Covariates, RealignerTargetCreator, IndelRealigner, SelectVariant, VariantFiltration, and UnifiedGenotyper. VCFtools was used for handling the vcf file format (20).

Substitution calls were made with GATK UnifiedGenotyper (16). SNPs and INDELs called with a Phred-scaled quality score of less than 30 were filtered out. Variants were removed based on MQ0 (median quality score zero) >4, (MQ0/read depth) >10%, quality depth 5, and FS (Phred-scaled P value using Fisher's exact test to detect strand bias) >200. After filtering, the SNPs were filtered again by removing those within 10 bp of INDELs. As ω and Λ statistics require haplotype information of each chromosome, we used BEAGLE (37) to infer the haplotype phase and impute missing alleles for the entire set of cattle populations simultaneously.

Genome STRiP (18) was used for deletion discovery and genotyping of structural variants in the population using the repeat masked genome. Initial genotype likelihoods were derived with a Bayesian model. Annotation information was obtained from Ensembl 68 (UMD_3.1) (38) and SnpEff 3.0 (39).

Statistical analysis

Cattle genomes were divided into bins of 1 Mb, and footprints of positive selection were calculated for each bin with a grid size of 1000 using LD-based statistics with OmegaPlus (29, 30) and SFS-based statistics with SweepFinder (31). The cutoff for ω statistics from OmegaPlus and Λ statistics from SweepFinder was the 99.9% quartile of empirical distribution of each statistic. Combination statistics using Λ statistics and ω statistics were calculated . To check the correlation between distribution of variants and that of repeat elements, repeat element information was obtained from UCSC with RepeatMasker Open-3.0 (40) using bovine genome UMD 3.1. The number of variants and each of the repeat elements were counted in each 1 Mb bin region. The correlations were calculated among them, and they were drawn using the *corrplot R* package (41).

REFERENCES

1. Qanbari, S., Gianola, D., Hayes, B., Schenkel, F., Miller, S., Moore, S., Thaller, G. and Simianer, H. (2011) Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. *BMC Genomics* **12**, 318.
2. Larkin, D. M., Daetwyler, H. D., Hernandez, A. G., Wright, C. L., Hetrick, L. A., Boucek, L., Bachman, S. L., Band, M. R., Akraiko, T. V. and Cohen-Zinder, M. (2012) Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 7693-7698.
3. Xia, Q., Guo, Y., Zhang, Z., Li, D., Xuan, Z., Li, Z., Dai, F., Li, Y., Cheng, D. and Li, R. (2009) Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). *Science* **326**, 433-436.
4. Rubin, C. J., Zody, M. C., Eriksson, J., Meadows, J. R. S., Sherwood, E., Webster, M. T., Jiang, L., Ingman, M., Sharpe, T. and Ka, S. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587-591.
5. Kim, J., Park, S. and Yeo, J. (2003) Linkage mapping and QTL on chromosome 6 in Hanwoo (Korean Cattle). *Asian-Aust. J. Anim. Sci.* **16**, 1402-1405.
6. Lee, Y., Lee, J., Lee, J., Kim, J., Park, H. and Yeo, J. (2008) Identification of candidate SNP (single nucleotide polymorphism) for growth and carcass traits related to QTL on chromosome 6 in Hanwoo (Korean cattle). *Asian-Aust. J. Anim. Sci.* **21**, 1703-1709.
7. Lee, S. H., Gondro, C., van der Werf, J., Kim, N. K., Lim, D., Park, E. W., Oh, S. J., Gibson, J. and Thompson, J. (2010) Use of a bovine genome array to identify new biological pathways for beef marbling in Hanwoo (Korean Cattle). *BMC Genomics* **11**, 623.
8. Lee, S., Van Der Werf, J., Park, E., Oh, S., Gibson, J. and Thompson, J. (2010) Genetic polymorphisms of the bovine Fatty acid binding protein 4 gene are significantly associated with marbling and carcass weight in Hanwoo (Korean Cattle). *Anim. Genet.* **41**, 442-444.
9. Lim, D., Lee, S. H., Cho, Y. M., Yoon, D., Shin, Y., Kim, K. W., Park, H. S. and Kim, H. (2010) Transcript profiling of expressed sequence tags from intramuscular fat, longissimus dorsi muscle and liver in Korean cattle (Hanwoo). *BMB Rep.* **43**, 151-121.
10. Lee, S. H., Park, E. W., Cho, Y. M., Kim, S. K., Lee, J. H., Jeon, J. T., Lee, C. S., Im, S. K., Oh, S. J. and Thompson, J. (2007) Identification of differentially expressed genes related to intramuscular fat development in the early and late fattening stages of hanwoo steers. *J. Biochemical. Mol. Biol.* **40**, 757-764.
11. Yang, B. C., Hwang, S. S., Im, G. S., Lee, D. K., Jeon, I. S.

and Park, S. B. (2012) Phenotypic characterization of Hanwoo (native Korean cattle) cloned from somatic cells of a single adult. *BMB Rep.* **45**, 38-43.

12. Barrett, R. D. H. and Schluter, D. (2008) Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**, 38-44.

13. Eizaguirre, C., Lenz, T. L., Kalbe, M. and Milinski, M. (2012) Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. *Nat. Commun.* **3**, 621.

14. Barrett, R. D. H., Paccard, A., Healy, T. M., Bergek, S., Schulte, P. M., Schluter, D. and Rogers, S. M. (2011) Rapid evolution of cold tolerance in stickleback. *P. Roy. Soc. B-Biol Sci.* **278**, 233-238.

15. Hermisson, J. and Pennings, P. S. (2005) Soft sweeps molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**, 2335-2352.

16. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M. A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297-1303.

17. Quinlan, A. R. and Hall, I. M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842.

18. Handsaker, R. E., Korn, J. M., Nemesh, J. and McCarroll, S. A. (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269-276.

19. Zhan, B., Fadista, J., Thomsen, B., Hedegaard, J., Panitz, F. and Bendixen, C. (2011) Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. *BMC Genomics.* **12**, 557.

20. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T. and Sherry, S. T. (2011) The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158.

21. Feulner, P. G. D., Chain, F. J. J., Panchal, M., Eizaguirre, C., Kalbe, M., Lenz, T. L., Mundry, M., Samonte, I. E., Stoll, M. and Milinski, M. (2012) Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. *Mol. Ecol.* **22**, 635-649.

22. Jarrell, V. L., Lewin, H. A., Da, Y. and Wheeler, M. B. (1995) Gene-centromere mapping of bovine DYA, DRB3, and PRL using secondary oocytes and first polar bodies: evidence for four-strand double crossovers between DYA and DRB3. *Genomics* **27**, 33-39.

23. Andersson, L., Lundén, A., Sigurdardottir, S., Davies, C. J. and Rask, L. (1988) Linkage relationships in the bovine MHC region. High recombination frequency between class II subregions. *Immunogenetics* **27**, 273-280.

24. Park, C., Russ, I., Da, Y. and Lewin, H. A. (1995) Genetic mapping of F13A to BTA23 by sperm typing: difference in recombination rate between bulls in the DYA-PRL interval. *Genomics* **27**, 113-118.

25. Paape, T., Zhou, P., Branca, A., Briskine, R., Young, N. and Tiffin, P. (2012) Fine-scale population recombination rates, hotspots, and correlates of recombination in the

medicago truncatula genome. *Genome Biol. Evol.* **4**, 726-737.

26. Hedrick, P. W. (2002) Pathogen resistance and genetic variation at MHC loci. *Evolution* **56**, 1902-1908.

27. Zimin, A. V., Kelley, D. R., Roberts, M., Marçais, G., Salzberg, S. L. and Yorke, J. A. (2012) Mis-assembled "segmental duplications" in two versions of the bos taurus genome. *PloS One* **7**, e42680.

28. Partipilo, G., D'Addabbo, P., Lacalandra, G. M., Liu, G. E. and Rocchi, M. (2011) Refinement of Bos taurus sequence assembly based on BAC-FISH experiments. *BMC Genomics* **12**, 639.

29. Alachiotis, N., Stamatakis, A. and Pavlidis, P. (2012) OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics* **28**, 2274-2275.

30. Kim, Y. and Nielsen, R. (2004) Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**, 1513-1524.

31. Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G. and Bustamante, C. (2005) Genomic scans for selective sweeps using SNP data. *Genome Res.* **15**, 1566-1575.

32. Chen, J. J., Reid, C. E., Band, V. and Androphy, E. J. (1995) Interaction of papillomavirus E6 oncoproteins with a putative calcium-binding protein. *Science* **269**, 529-531.

33. De Groot, N. G., Otting, N., Doxiadis, G. G. M., Balla-Jhagjhoorsingh, S. S., Heeney, J. L., Van Rood, J. J., Gagneux, P. and Bontrop, R. E. (2002) Evidence for an ancient selective sweep in the MHC class I gene repertoire of chimpanzees. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11748-11753.

34. Bae, Y., Lee, C., Kang, M., Yoon, S., Park, J. and Jean, Y. (2005) Bovine papillomavirus detection from bovine teats using immunohistochemistry and electronmicroscopy. *Korean J. Vet. Res.* **45**, 233-238.

35. Langmead, B. and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* **9**, 357-359.

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

37. Browning, S. R. and Browning, B. L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Gen.* **81**, 1084-1097.

38. Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S. and Fitzgerald, S. (2012) Ensembl 2012. *Nucl Acids Res.* **40**, D84-D90.

39. Cingolani, P., Platts, A., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. and Ruden, D. M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80-92.

40. Smit, A., Hubley, R. and Green, P. (1996) Repeat Masker Open-3.0. http://www.repeatmasker.org.

41. Friendly, M. (2002) Corrgrams. *Am. Stat.* **56**, 316-324.