G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2018;16(1):14-20
https://doi.org/10.5808/GI.2018.16.1.14

ORIGINAL ARTICLE

Estimation of the Genetic Substitution Rate of Hanwoo and Holstein Cattle Using Whole Genome Sequencing Data

Young-Sup Lee, Donghyun Shin*

Department of Animal Biotechnology, Chonbuk National University, Jeonju 54896, Korea

Despite the importance of mutation rate, some difficulties exist in estimating it. Next-generation sequencing (NGS) data yields large numbers of single-nucleotide polymorphisms, which can make it feasible to estimate substitution rates. The genetic substitution rates of Hanwoo and Holstein cattle were estimated using NGS data. Our main findings was to calculate the gene's substitution rates. Through estimation of genetic substitution rates, we found: diving region of altered substitution density exists. This region may indicate a boundary between protected and unprotected genes. The protected region is mainly associated with the gene ontology terms of regulatory genes. The genes that distinguish Hanwoo from Holstein in terms of substitution rate predominantly have gene ontology terms related to blood and circulatory system. This might imply that Hanwoo and Holstein evolved with dissimilar mutation rates and processes after domestication. The difference in meat quality between Hanwoo and Holstein could originate from differential evolution of the genes related to these blood and circulatory system ontology terms.

Keywords: genetic substitution rate, Hanwoo, hidden substitution factor, Holstein, whole genome sequencing

Introduction

Mutation rates include genetic changes such as single nucleotide polymorphisms (SNPs), copy number variations (CNVs), insertion/deletions (indels), microsatellites and minisatellites. Specifically, the subset of the total mutation rate caused by SNPs is called the substitution rate. Rate of point mutation can be determined indirectly by estimating neutral substitution rates [1]. In spite of the importance of substitution rates, there have been few investigations to date. We estimated the substitution rates of Hanwoo and Holstein cattle using the concept of the hidden substitution factor (HSF). The vast number of SNPs in next-generation sequencing (NGS) require the use of HSF to estimate the substitution rate. We defined the HSF as an independent unit for substitution rate in any lineage. For instance, the initial founder number is the HSF if the effective population size has been maintained as a constant. Otherwise, HSF is related to both founder number and effective population size. Using HSF, we estimated the genetic substitution rates of the genes of the Hanwoo and Holstein breeds in Korea.

Kumar and Subramanian [1] estimated the substitution rate in the mammalian genes using amino acid degenerate sites and evolutionary distance. Lee *et al*. [2] obtained the nonsynonymous SNPs, splice-site variants and coding indels in the Bovine genome including Hanwoo. They used whole-genome resequencing. Melka *et al*. [3] used Bovine SNP 50 beadchip to identify genomic differences between Hanwoo and Holstein. In our analysis, we used whole-genome sequencing to estimate substitution rates of Hanwoo and Holstein and identified genomic differences between Hanwoo and Holstein using gene ontology (GO) analysis. Our novel approach was to use intraspecies substitution rate estimation by inserting HSF into the evolutionary distance. HSF of Hanwoo is closely related to the initial found number that was caused by migration into Korean peninsula. We hypothesized that initial founder number was more than 1 [1-3].

Hanwoo designates the native, taurine type of Korean cattle. The breed originated approximately 4,000 years ago.

Received January 5, 2018; Revised February 14, 2018; Accepted February 14, 2018

***Corresponding author:** Tel: +82-63-279-4748, Fax: +82-63-270-2614, E-mail: sdh1214@gmail.com

The production of Hanwoo as the main beef cattle has occurred since the 1960s with the rapid growth of the Korean economy. The tasty beef from Hanwoo cattle is popular among Koreans and foreigners. Holstein-Friesian cattle represents the famous milk producing breeds. The holstein was introduced to Korea in 1902, and the Korean cattle industry has developed rapidly since 1960 [4-6]. Hanwoo and Holstein are very important in cattle industry in Korea.

We estimated the evolutionary distance of Holstein and Hanwoo using the number of SNPs, and HSF. Evolutionary distance refers to the cumulative change between DNA or protein sequences that were derived from a common ancestor. There were various methods to estimate it such as JC69 model, Kimura-2-parameter model (K2P), F81 model, HKY85 model. K2P used transition/transversion ratio (tr/tv ratio) to estimate the evolutionary distance. It is moderately complex model [7-10]. We assumed that the evolutionary distance had a linear relationship to HSF. The generation times of Hanwoo and Holstein were set to be 5 years. Then, the genetic substitution rates of the genes were calculated and those features were surveyed.

Materials and Methods

Data preparation

In this study, we used a whole-genome sequence data set consisting of 23 Hanwoo and 10 Holstein from NCBI Sequence Read Archive database (PRJNA210523, PRJNA210521, and PRJNA210519). We used fastQC software to perform a quality check on the raw sequence data [11]. Using Trimmomatic-0.32, we removed potential adapter sequences before sequence alignment [11]. Paired-end sequence reads were mapped to the reference genome (*UMD 3.1.75*) from the Ensemble database using Bowtie2 with default settings [12]. For downstream processing and variant-calling, we used open-source software packages: Picard tools (http:// broadinstitute.github.io/picard/), SAMtools, and Genome Analysis Toolkit (GATK) [13, 14]. The "CreateSequenceDictionary" and "MarkDuplicates" Picard command-line tools were used to read the reference FASTA sequence to write a bam file containing a sequence dictionary, and to filter potential PCR duplicates, respectively. Using SAMtools, we created index files for the reference and bam files. We then performed local realignment of sequence reads to correct misalignments due to the presence of small insertions and deletions using the GATK "Realigner-TargetCreator" and "IndelRealigner" arguments. Additionally, base quality score recalibration was performed to get accurate quality scores and to correct the variations in quality associated with machine cycle and sequence context. For calling variants, GATK "UnifiedGenotyper" and "SelectVariants" arguments were used with the following filtering criteria. All variants with (1) a Phred-scaled quality score of less than 30; (2) a read depth less than 5; (3) an MQ0 (total count across all samples of mapping quality zero reads) $>$ 4; or a (4) a Phred-scaled p-value using Fisher exact test of more than 200 were filtered out to reduce false-positive calls due to strand bias. We used the "vcf-merge" tools of VCFtools in order to merge all of the variant calling format files for the 33 samples [15].

The number of total SNPs was 37,484,886 and after using minor allele frequency; Holstein 0.1, Hanwoo 0.04) and Hardy-Weinberg equilibrium ($p < 0.0001$), the number of SNPs left in Hanwoo and Holstein were 12,626,097 and 8,636,673, respectively. The estimated transition/transversion ratio (tr/tv ratio) was calculated to be 2.24 using SnpEff [16].

The relationship between evolutionary distance and founder number

Evolutionary distance is linear to mutation rate and substitution rate [17, 18]. Evolutionary distance is proportional to the harmonic number of of sample size (a_1) because the sample size can increase the number of SNPs. Theta represents the population mutation parameter ($\theta = 4N\mu$). While Watterson's theta estimator ($\hat{\theta} = \frac{\#segregating\ sites}{a_1}$) have been used irrespective of HSF and general SNPchip did not require HSF, in the NGS data, substitution rate (μ) would be high without the concept of HSF. HSF concept assumes that the initial found number after migration would not be equal to 1 like SNPchips. Biologically, that initial migration number was not 1, is very natural. a_1 is the harmonic number of sample size n. We defined the evolutionary distance as the following:

$$d = f\, a_1\, \mu\, t \quad (1)$$

, where d is evolutionary distance, f is the HSF, a_1 is the harmonic number of sample size, μ is the substitution rate, and t is the divergence time between the species and reference species.

The evolutionary distance model was K2P [7, 19]. K2P assumes the transition/transversion ratio (tr/tv ratio). We set the tr/tv ratio to be 2.24, and the divergence time to be 4,000 years ago [4, 20, 21]. The increase in the number of SNPs could be caused by sample size and HSF.

HSF estimation via simulation

To estimate HSF, we simulated the average substitution rate of SNP chip data. We assumed the SNP number of general SNP chip data. The SNP chip data have 30,000–50,000 SNPs in total, and the founder number can be

assumed to be close to "1." We simulated using the following parameters: the number of SNPs would be 30,000–50,000 and the harmonic number of sample size would be 4–6. Using Eq. (1), the expected value of the substitution rate was 3×10^{-10} to 5×10^{-10} (/bp/y). Using the result, we set the average substitution rate to be 4×10^{-10}. This was a similar result to Roach *et al*. [22] who estimated the average mutation rate of human as 10^{-8} order per generation.

The HSF of Hanwoo and Holstein was estimated using Eq. (1). The evolutionary distances of the Hanwoo and Holstein population across chromosomes were calculated and then average HSF of Hanwoo (23 individuals) and Holstein (10 individuals) were estimated using the substitution rates of each chromosome.

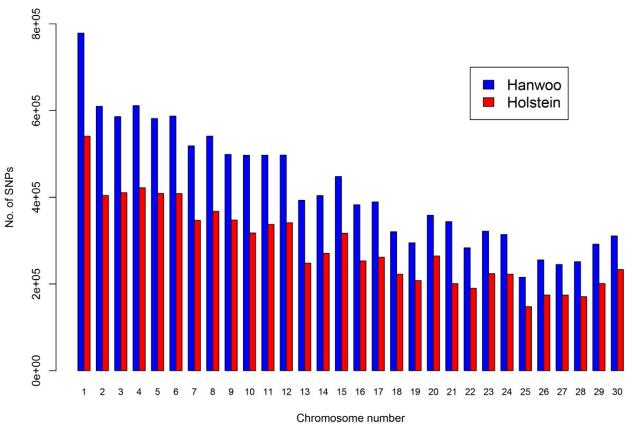

Fig. 1. The number of single nucleotide polymorphisms (SNPs) across chromosomes in Hanwoo and Holstein. The number of individuals of Hanwoo and Holstein was 23 and 10, respectively. Because the number of inidividuals in Hanwoo and Holstein differed, this barplot just showed the patterns of the number of SNPs in Hanwoo and Holstein.

Genetic substitution rate estimation of genes

After estimating HSF, we calculated each gene's substitution rate in Hanwoo and Holstein populations. The bovine gene catalog was retrieved from the ensemble website (http://www.ensembl.org). The novel calculation was based on the number of SNPs inside each gene and used K2P model. We surveyed the genes that had the highest and lowest substitution rates. We compared the genetic substitution rates of Hanwoo with those of Holstein.

Results

HSF estimation

We estimated HSF of Hanwoo and Holstein. The HSF of Hanwoo and Holstein were 351 ± 46 (mean ± SD) and 279 ± 39 (mean ± SD), respectively. The Holstein HSF was not greater than that of Hanwoo because the Holstein HSF might pertain only to the history of Korean Holsteins. Fig. 1 shows the number of SNPs across chromosomes in Hanwoo and Holstein. The difference in the number of SNPs across chromosomes might be mainly due to the number of individuals i.e., 23 and 10 in Hanwoo and Holstein, respectively.

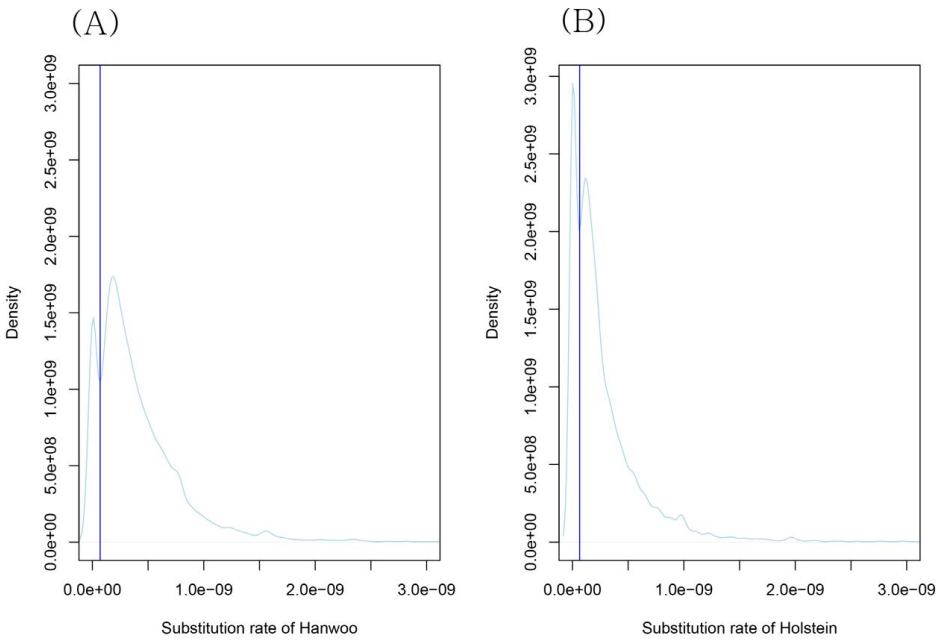

Fig. 2. Density plot of substitution rates. It indicates that diving region exists in both Hanwoo (A) and Holstein (B). The protected region (low substitution rates) and the unprotected region (high substitution rates) can be differentiated by the diving region. The dividing region's substitution rate (blue vertical line) was 7×10^{-11} and 6.5×10^{-11} in Hanwoo and Holstein, respectively.

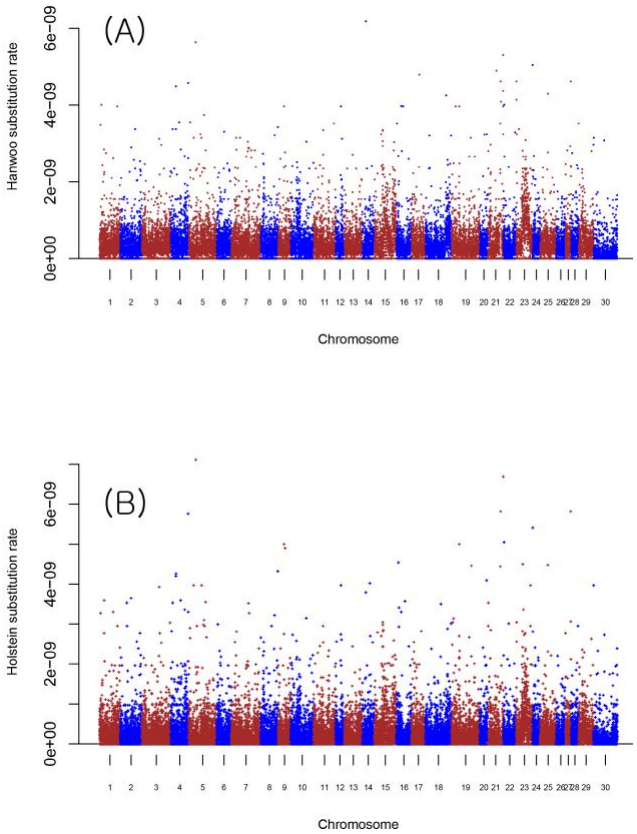

Fig. 3. Manhattan plot of Hanwoo (A) and Holstein (B) across chromosomes.

Substitution rate

Fig. 2 shows the density of Hanwoo (Fig. 2A) and Holstein (Fig. 2B) substitution rates. The density of substitution rates showed the presence of a region in which the density showed significant decrease and increase. We might consider that this diving region differentiated the protected regions from unprotected regions. The substitution rates of the boundary regions in Hanwoo and Holstein were 7×10^{-11} and 6.5×10^{-11}, respectively (/bp/y).

Fig. 3 shows Manhattan plot of substitution rates. It shows that the distribution of substitution rates between Hanwoo and Holstein was somewhat different across the genome. Fig. 4 shows the genetic substitution rate of Holstein against that of Hanwoo.

GO in Hanwoo and Holstein

Table 1 shows the results of a DAVID analysis of GO terms for genes with low substitution rates (Hanwoo $< 7 \times 10^{-11}$; Holstein $< 6.5 \times 10^{-11}$). The criteria was substitution rates below diving regions. Most terms were regulation. It is obvious that regulatory genes would have low substitution rates. Table 2 shows GO terms of genes with different substitution rates between Hanwoo and Holstein. The major GO terms were related to blood and circulatory system. The criteria of genes was an over 20-fold difference in substitution rate between Hanwoo and Holstein (Fig. 4). Supplementary Table 1 shows the total genetic substitution rates and those ratios between Hanwoo and Holstein for protein-coding genes. The averages of gene substitution rates in Hanwoo

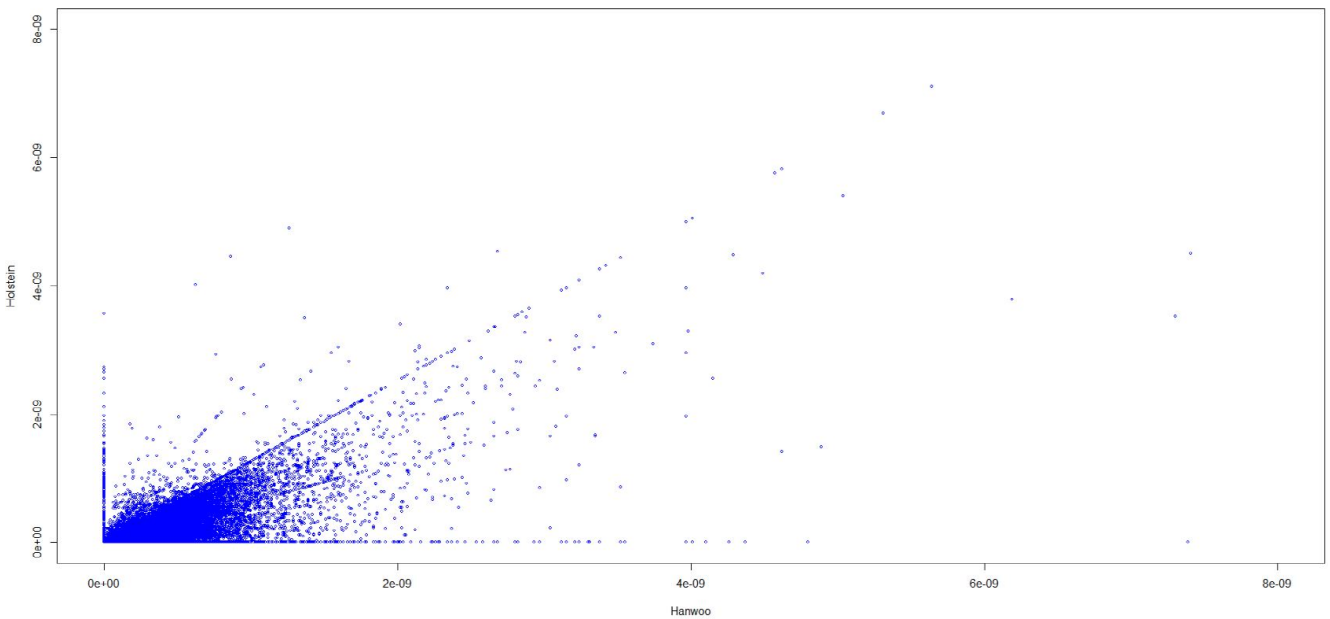

Fig. 4. The substitution rates of Holstein against Hanwoo. It showed the nonlinearlity of substitution rates between Hanwoo and Holstein.

Table 1. Gene ontology terms (GO terms) with low substitution rates of 7×10^{-11} and 6.5×10^{-11} in Hanwoo and Holstein, respectively

Term	Count	p-value
GO:0009952; anterior/posterior pattern specification	10	4.79E-05
GO:0006357; regulation of transcription from RNA polymerase II promoter	18	1.23E-04
GO:0000122; negative regulation of transcription from RNA polymerase II promoter	24	3.22E-04
GO:0006351; transcription, DNA-templated	29	0.001
GO:0045944; positive regulation of transcription from RNA polymerase II promoter	27	0.002
GO:0043518; negative regulation of DNA damage response, signal transduction by p53 class mediator	4	0.003
GO:0060828; regulation of canonical Wnt signaling pathway	4	0.004
GO:0045668; negative regulation of osteoblast differentiation	5	0.008
GO:0007264; small GTPase mediated signal transduction	12	0.0106
GO:0032525; somite rostral/caudal axis specification	3	0.015
GO:0001764; neuron migration	7	0.017
GO:0048701; embryonic cranial skeleton morphogenesis	4	0.031
GO:0007219; Notch signaling pathway	6	0.036
GO:0007519; skeletal muscle tissue development	4	0.037
GO:0007413; axonal fasciculation	3	0.039
GO:0045671; negative regulation of osteoclast differentiation	3	0.039
GO:0048704; embryonic skeletal system morphogenesis	4	0.043
GO:0021722; superior olivary nucleus maturation	2	0.048
GO:0050725; positive regulation of interleukin-1 beta biosynthetic process	2	0.048
GO:0032078; negative regulation of endodeoxyribonuclease activity	2	0.048
GO:0021568; rhombomere 2 development	2	0.048
GO:0061104; adrenal chromaffin cell differentiation	2	0.048

The analysis shows that the regulatory genes have low substitution rates.

Table 2. GO of the genes based on Hanwoo/Holstein substitution rate ratio

Term	Count	p-value
GO:0043933; macromolecular complex subunit organization	18	0.002
GO:0065003; macromolecular complex assembly	17	0.003
GO:0006461; protein complex assembly	12	0.006
GO:0070271; protein complex biogenesis	12	0.006
GO:0050880; regulation of blood vessel size	5	0.007
GO:0035150; regulation of tube size	5	0.007
GO:0003018; vascular process in circulatory system	5	0.007
GO:0003013; circulatory system process	7	0.009
GO:0008015; blood circulation	7	0.009
GO:0009303; rRNA transcription	3	0.012
GO:0008217; regulation of blood pressure	5	0.02
GO:0006351; transcription, DNA-dependent	6	0.03
GO:0006940; regulation of smooth muscle contraction	3	0.03
GO:0003044; regulation of systemic arterial blood pressure mediated by a chemical signal	3	0.04
GO:0048754; branching morphogenesis of a tube	4	0.04
GO:0032774; RNA biosynthetic process	6	0.04
GO:0006813; potassium ion transport	7	0.04
GO:0002035; brain renin-angiotensin system	2	0.05
GO:0002016; regulation of blood volume by renin-angiotensin	2	0.05
GO:0003072; renal control of peripheral vascular resistance involved in regulation of systemic arterial blood pressure	2	0.05

The ratio was above 20-fold in Hanwoo compared to Holstein. It is noteworthy that blood and circulatory system gene ontology (GO) terms were frequent. We suggest that these categories can differentiate the evolution of Hanwoo from that of Holstein.

and Holstein were 4.1×10^{-10} and 2.9×10^{-10}. This was concordant with the simulation study of substitution rates (see Materials and Methods). We used the pseudocount 10^{-13} where the substitution rate in Holstein was estimated at zero to calculate the ratios.

Discussion

HSF estimation

The HSF can be viewed as the number of individuals requiring to explain the substitution rates present in a population. In SNP chip data, the HSF can be as close as 1. Founder effect can decrease nucleotide diversity and number of SNPs [23, 24]. Thus, smaller founder number can decrease the number of SNPs. Additionally, a population expansion after bottleneck events can increase the number of SNPs. Thus HSF should be estimated carefully, taking care of population history such as bottleneck event and population expansion.

The genes in Hanwoo and Holstein have average substitution rates on the order of 10^{-10}. CNVs, microsatellites and minisatellites can also cause mutations. Therefore, it is very difficult to estimate the mutation rates with great accuracy. Instead, we focused on substitution rates which could be estimated using SNPs. The substitution rates on the order of 10^{-10} in Hanwoo and Holstein were similar to the human mutation rates [22].

The evolutionary distance can be assumed to be linear with HSF and the harmonic number of sample size (a_1). Like Watterson's theta estimator, the number of SNPs increases with sample size and thus we assumed that the evolutionary distance was linear with the harmonic mean of sample size (a_1). Because HSF represents the independent substitution unit in the population, thus is linear with the number of SNPs in the population as Eq. (1).

Substitution diving region

The gene ontology (GO terms) analysis for genes with the lowest substitution rates in Hanwoo and Holstein showed that regulatory genes belong to the low substitution rate gene set. Regulatory genes might be protected genes. The existence of boundaries between the highest and lowest substitution rates showed that there were gaps in substitution rate between protected genes (lowest substitution rates) and unprotected genes (highest substitution rates). Such diving regions might not be the special case in Hanwoo or Holstein. This could provide evidence of differentiation between protected genes and unprotected genes.

GO results differentiating Hanwoo from Holstein

The differentiation of Hanwoo from Holstein was identified by comparing the genetic substitution rate at the gene level. In GO analysis of differentially substituted genes between Hanwoo and Holstein, the blood and circulatory system terms were most common. The disparate evolutionary patterns after domestication between Hanwoo and Holstein could be caused by these genes. Hanwoo, which mainly provides meat, and Holstein, which mainly provides milk, might evolve differently, and blood and circulatory system-related GO genes could play some role in the evolution of Hanwoo and Holstein.

ORCID: Young-Sup Lee: http://orcid.org/0000-0002-0819-0553; Donghyun Shin: http://orcid.org/0000-0001-9267-0850

Authors' contribution

Conceptualization: YSL
Data curation: DS
Formal analysis: YSL
Methodology: YSL
Writing: YSL
Writing – review & editing: YSL, DS

Acknowledgments

This work was supported by a grant from the Next Generation BioGreen21 project (No. PJ011044), Rural Development Administration, Republic of Korea. In addition, this research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No.2017R 1A6A3A11033784).

Supplementary material

Supplementary data including one table can be found with this article online at http://www.genominfo.org/src/sm/gni-16-14-s001.pdf.

References

1. Kumar S, Subramanian S. Mutation rates in mammalian genomes. *Proc Natl Acad Sci U S A* 2002;99:803-808.
2. Lee KT, Chung WH, Lee SY, Choi JW, Kim J, Lim D, *et al*. Whole-genome resequencing of Hanwoo (Korean cattle) and insight into regions of homozygosity. *BMC Genomics* 2013;14:519.
3. Melka HD, Jeon EK, Kim SW, Han JB, Yoon D, Kim KS. Identification of genomic differences between Hanwoo and Holstein breeds using the Illumina Bovine SNP50 BeadChip. *Genomics Inform* 2011;9:69-73.

4. Lee HJ, Kim J, Lee T, Son JK, Yoon HB, Baek KS, *et al*. Deciphering the genetic blueprint behind Holstein milk proteins and production. *Genome Biol Evol* 2014;6:1366-1374.
5. Rhee MS, Ryu YC, Imm JY, Kim BC. Combination of low voltage electrical stimulation and early postmortem temperature conditioning on degradation of myofibrillar proteins in Korean native cattle (Hanwoo). *Meat Sci* 2000;55:391-396.
6. Lee SH, Park BH, Sharma A, Dang CG, Lee SS, Choi TJ, *et al*. Hanwoo cattle: origin, domestication, breeding strategies and genomic selection. *J Anim Sci Technol* 2014;56:2.
7. Srivathsan A, Meier R. On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. *Cladistics* 2012;28:190-194.
8. Yang Z. Estimating the pattern of nucleotide substitution. *J Mol Evol* 1994;39:105-111.
9. Yang Z. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst Biol* 1994;43:329-342.
10. Steel MA, Fu YX. Classifying and counting linear phylogenetic invariants for the Jukes-Cantor model. *J Comput Biol* 1995;2: 39-47.
11. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114-2120.
12. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357-359.
13. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297-1303.
14. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al*. The sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078-2079.
15. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al*. The variant call format and VCFtools. *Bioinformatics* 2011;27:2156-2158.
16. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, *et al*. 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* (*Austin*) 2012;6:80-92.
17. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 1980;16:111-120.
18. Tajima F, Nei M. Estimation of evolutionary distance between nucleotide sequences. *Mol Biol Evol* 1984;1:269-285.
19. Casanellas M, Fernández-Sánchez J. Geometry of the Kimura 3-parameter model. *Adv Appl Math* 2008;41:265-292.
20. Murray C, Huerta-Sanchez E, Casey F, Bradley DG. Cattle demographic history modelled from autosomal sequence variation. *Philos Trans R Soc Lond B Biol Sci* 2010;365:2531-2539.
21. Bovine HapMap Consortium, Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, *et al*. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 2009;324:528-532.
22. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, *et al*. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 2010;328:636-639.
23. Nei M, Maruyama T, Chakraborty R. The bottleneck effect and genetic variability in populations. *Evolution* 1975;29:1-10.
24. Ladizinsky G. Founder effect in crop-plant evolution. *Econ Bot* 1985;39:191-199.