Journal of Animal Breeding and Genomics
pISSN:1226-5543 eISSN:2586-4297
JABG. 2021 December, 5(4):255-266
https://doi.org/10.12972/jabng.20210026
RESEARCH ARTICLE
OPEN ACCESS

An update of physical genomic information for commercial bovine SNP chip based on the *ARS-UCD1.2* reference genome

Jeongwoen Shin[1†], Dong Jae Lee[1†], Dongwon Seo[1], Soo Hyun Lee[2], You-Sam Kim[3], Myung Hum Park[3], Cedric Gondro[4], Yang-Mo Koo[5*] and Seung Hwan Lee[1*]

[1]Division of Animal and Dairy Science, Chungnam National University, Daejeon, 34134, Korea
[2]Division of Animal Breeding and Genetics, National Institute of Animal Science, RDA, Cheonan, 31000, Korea
[3]TNT Research. Co., Ltd., Anyang, 14059, Korea
[4]Department of Animal Science, Michigan State University, East Lansing, MI, 48824, USA
[5]Korea Animal Improvement Association (KAIA), Seoul, 06668, Korea

ABSTRACT

With rapid advances in next-generation sequencing technology, previously released reference genomes are being upgraded, and new genome assemblies for novel organisms are being published. In recent updates of the bovine reference genome, changes in single-nucleotide polymorphism (SNP) position and gene location occurred. The BovineSNP50 BeadChip series and Hanwoo SNP50K BeadChip are based on the older version of the bovine reference assembly *UMD3.1.* In a recent bovine reference genome of *ARS-UCD1.2*, SNP positions changed in the map file of a commercial SNP chip. Hence, this study assessed SNP re-positioning in the updated reference genome using the LiftoverVcf tool. We annotated SNP variants, which were mapped to both reference genomes, to compare SNP information such as position and adjacent genes between assemblies. The results showed that approximately 3.5~4.0% of the SNPs failed the lift-over procedure. Moreover, the number of SNPs successfully lifted over that affect protein expression, number of effects according to SNP type, and locations of Hanwoo economic trait-related candidate genes differed between reference genomes. In conclusion, the map file of SNP chip data based on *UMD3.1* needs to use the position information of *ARS-UCD1.2*, a new reference genome, for accurate comparison with other currently published studies.

Key words: Annotation, *ARS-UCD1.2*, Lift-over, SNP BeadChip, *UMD3.1*

INTRODUCTION

After the publication of a draft human genome sequence in 2003, genome projects for various species were carried out. As a result, reference genome sequences have been elucidated for a total of 49,286 living organisms to date (https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/). Furthermore, with the rapid development of next-generation sequencing (NGS) technology, the time and cost needed to produce reads and libraries for new species have been reduced dramatically (Kong et al., 2019). Furthermore, existing genome sequences are constantly being supplemented with new information.

UMD3.1, a previous version of the *Bos taurus* reference genome, was released by the Center for Bioinformatics and Computational Biology (NCBI) of the University of Maryland in December 2009. It was sequenced using the BAC-by-BAC hierarchical method and whole-genome shotgun methods and then aligned using Celera Assembler version 5.2. According to

[†]These authors contributed equally to this work.

***Corresponding author:** Seung Hwan Lee
Division of Animal & Dairy Science, Chungnam National University, Daejeon, 34134, Korea
Tel: 042-821-5878 Fax: 042-825-9754 E-mail: slee46@cnu.ac.kr

Received: 16 December, 2021 **Revised:** 29 December, 2021 **Accepted:** 29 December, 2021

the summary statistics for the references reported in Ensembl (Chen et al., 2010), in the newer *ARS-UCD1.2* reference genome, the total length and ungapped region length have increased by 45,447,833 (1.70%) and 66,157,558 (2.50%) bp, respectively, compared to the previous version (**Table 1**). By contrast, the number of scaffolds and contigs have decreased by 4,298 (66.03%) and 73,193 (96.57%) bp, respectively. However, the scaffold N50 and contig N50 increased 1.64- and 267.11-fold, respectively. In addition, the genome coverage in *UMD3.1* is 9.0x versus 80.0x in *ARS-UCD1.2*. Thus, the *ARS-UCD1.2* genome has an approximately nine-fold greater depth compared to the *UMD3.1* genome. Therefore, the most recent sequence appears to be more accurate, and we should expect to obtain better results using it in genome-wide association studies (GWAS), for example.

Table 1. Assembly statistics about *UMD3.1* and *ARS-UCD1.2*

	UMD3.1	*ARS-UCD1.2*
Total length	2,670,405,961	2,715,853,794
Ungapped length	2,649,668,074	2,715,825,632
Chromosomes & Plasmid	30	31
Spanned gaps	69,281	386
Unspanned gaps	3,193	0
Scaffolds	6,509	2,211
Scaffold N50	6,3080,747	103,308,737
Contigs	75,790	2,597
Contig N50	96,951	25,896,116
Genome coverage	9.0x	80.0x

UMD3.1 is the previous version of bovine reference assembly, while *ARS-UCD1.2* is the latest. As the version upgraded, total length and ungapped length has been increased about 1.70% and 2.50%, respectively. In addition, scaffold N50 and contig N50 enlarged about 1.64 and 267.11 fold, respectively. And the genome coverage in *ARS-UCD1.2* has an approximately nine-fold greater than *UMD3.1*. By contrast, the number of scaffolds and contigs have decreased by 4,298 (66.03%) and 73,193 (96.57%) bp, respectively.

The BovineSNP50 BeadChip, developed for use with the Illumina system, is widely used for the prediction of genetic ability and in GWAS and genetic diversity studies based on genomic information for various cattle breeds. This single-nucleotide polymorphism (SNP) panel has been used for genomic prediction analysis in Nelore cattle (Boddhireddy, Prayaga, Barros, Lôbo, & DeNise, 2014), and in a GWAS of oleic acid and fatty acid composition in Japanese black cattle (Uemoto et al., 2011). In this SNP panel, more than half of the probes were designed using known genome-wide SNPs and Illumina sequencing of data from publicly available sources, such as bovine reference sequences, Btau, and the Bovine HapMap Consortium (https://www.illumina.com/documents/products/datasheets/datasheet_bovineLD.pdf). The chip has been updated from version 1 to 3; the reference sequence is also constantly being updated. In contrast to Btau 2.0, both the Bovine LD v2.0 Genotyping BeadChip and BovineSNP50 v3 BeadChip were developed based on six cattle types: Norwegian Red, Holstein, Brahman, Angus, Jersey, and Limousin. In 2019, the National Institute of Animal Science in South Korea developed the Hanwoo BeadChip v1, which is a SNP chip designed specifically for the Hanwoo breed (https://www.korea.kr/news/pressReleaseView.do?newsId=156250636). This SNP panel incorporates features of previous versions of the Illumina Bovine BeadChip array to allow use in Hanwoo and various cattle breeds. In addition, the panel contains additional SNP information from the whole-genome sequence of Hanwoo bulls and candidate SNPs related to 24 genetic diseases. Approximately one-third of the genetic information in the commercial chips, which was not relevant to the Hanwoo breed, was removed. Then, 4,762 population-specific, 4,383 carcass-related, and 44 genetic disease-related SNP markers were added to supplement the Hanwoo-related information.

In this study, SNPs in three panels were subjected to lift-over analysis to determine whether there had been any changes in location based on the latest reference genome, *ARS-UCD1.2*. Successfully remapped variants were annotated with reference to both *UMD3.1* and *ARS-UCD1.2* to obtain genetic information. We then compared the genomic information between *ARS-UCD1.2* and *UMD3.1* for each panel.

MATERIALS AND METHODS

1. Animals and genotypes

A total of 4,091 Hanwoo samples were used for genotyping in this study. SNP genotyping was conducted using the Bovine LD v2.0 Genotyping BeadChip (Bovine v2), BovineSNP50 v3 BeadChip (Bovine v3), and Hanwoo BeadChip v2 (Hanwoo v1). Bovine v2 contained an array of 54,609 SNPs and was used to analyze 341 of the 4,091 samples. Bovine v3 and Hanwoo v1 contained 53,218 and 53,866 variants and were used to genotype 100 and 3,650 samples, respectively. Reference sequence FASTA files and GTF annotation files for the *UMD3.1* and *ARS-UCD1.2* assemblies were downloaded from the Ensembl database (Chen et al., 2010).

2. Lift-over

PLINK binary files for Hanwoo v1, Bovine v2, and Bovine v3 were converted into variant calling format (VCF) using the recode option in PLINK 1.9 software (Purcell et al., 2007) for lift-over analysis. As the SNP chips were developed based on the *UMD3.1* assembly, the positions of variants should be updated to match those in the latest version (*ARS-UCD1.2*). To search for SNPs in the new reference genome, we used the Picard LiftoverVcf tool (https://broadinstitute.github.io/picard/). To assess shifts in the locations of variants, a chain file was downloaded from the UCSC Genome Browser website (Kent et al., 2002). Its pairwise alignment allows gaps in both sequences, *ARS-UCD1.2* and *UMD3.1*, at the same time. Each set of chain alignments starts with a header line, includes one or more alignment data lines, and ends with a blank line; thus, the format is relatively dense.

3. Annotations

To compare the variants common to *UMD3.1* and *ARS-UCD1.2*, variants that had successfully passed the lift-over procedure were extracted from the original sequence (*UMD3.1*) using R software and PLINK 1.9. Next, we functionally annotated only autosomes from the lifted-over SNPs using the SnpEff program (Cingolani et al., 2012). Data on these variants and the functional annotations were stored in HTML summary documents and VCF files. The SNP panels were then compared based on this information using R software.

RESULTS

1. Comparison of the number of variants in each SNP panel before and after lift-over

In total, 47,450 SNPs were identified as common among the three panels (Hanwoo v1, Bovine v2, and Bovine v3). The number of SNPs not shared between each pair of panels ranged from 64 to 3,458. Hanwoo v1 and Bovine v2 do not share 4,042 and 3,637 SNPs, respectively, with the other panels. Bovine v3 contains all SNPs found in all panels except for 64 variants, which are only included in Hanwoo v1 and Bovine v2 (**Figure 1**).

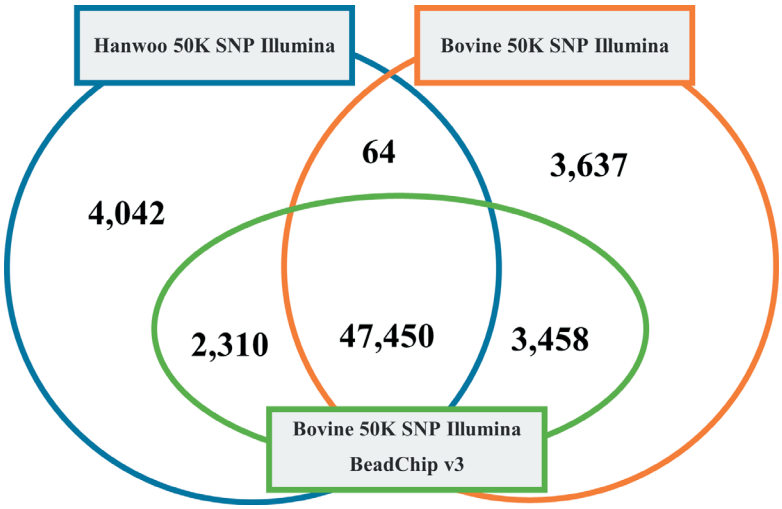

Figure 1. Structure of Hanwoo 50K SNP Illumina BeadChip v1, Bovine 50K SNP Illumina BeadChip v2 and Bovine 50K SNP Illumina BeadChip v3.
47,450 SNPs were identified as common among the three panels (Hanwoo v1, Bovine v2, and Bovine v3). Each pair of panels shares SNPs ranged from 64 to 3,458. Hanwoo v1 and Bovine v2 do not shared 4,042 and 3,637, respectively, with the other panels.

Table 2. Number of SNPs that succeed Liftover in each panel

	Hanwoo v1	Bovine v2	Bovine v3
Total	53,866	54,609	53,218
Without indels	53,860	54,609	53,212
Liftover	51,978	52,696	51,083
Rejected	1,882	1,913	2,129
Liftover rate (%)	96.5	96.5	96.0

Hanwoo v1, Bovine v2, and Bovine v3 were developed according to the UMD3.1 genome. They have 53,866, 54,609, and 53,218 SNPs respectively. After lift-over them to the ARS-UCD1.2 genome, the results showed that 51,978, 52,696, and 51,083 could be located in the updated reference assembly. The locations of the rest of the SNPs in the current reference were not identified as incompatibility between the previous and current references.

The variants in each panel were subjected to lift-over analysis using the Picard LiftoverVcf tool. The results showed that 51,978 of 53,866 variants in Hanwoo v1, 52,696 of 54,609 variants in Bovine v2, and 51,083 of 53,218 variants in Bovine v3 could be located in the *ARS-UCD1.2* assembly (**Table 2**). The locations of the rest of the SNPs in the current reference were not identified due to incompatibility between the previous and current references.

2. Summary of annotation results for each panel according to the references

For Hanwoo v1 in *ARS-UCD1.2*, 50,955 of 51,734 variants passed the SnpEff filter and were annotated. The total number of functional effects increased by 27,423 (46.95%) when *ARS-UCD1.2* (n = 85,830) was used as the reference genome rather than *UMD3.1* (n = 58,407). In addition, in *UMD3.1*, genetic variants were identified every 49,088 bp in the entire genome, while in *ARS-UCD1.2* SNPs were found every 48,854 bp (**Table 3-A**). For Bovine v2, 45,972 of 52,495 variants were annotated. When *ARS-UCD1.2* was used instead of *UMD3.1*, the number of functional effects increased by 25,140 (47.65%; n = 77,896 and 52,756, respectively). Also, the SNP frequency changed from 1 every 54,432 bp to 1 every 54,150 bp (**Table 3-B**). For Bovine v3, 45,136 of 50,864 variants passed the filter and were annotated (**Table 3-C**). The number of predicted effects increased by 24,254 (46.83%) when *ARS-UCD1.2* (n = 76,050) was used instead of *UMD3.1* (n = 51,796).

Table 3. Summary of annotation results of three panels for *ARS-UCD1.2* and *UMD3.1* references

(A) Hanwoo v1

	UMD3.1.94	*ARS-UCD1.2.98*
Number of variants (Before filter)	51,972	51,734
Number of variants processed	51,175	50,955
Number of effects	58,407	85,830
Genome total length	2,670,422,299	2,715,853,792
Genome effective length	2,512,082,506	2,489,385,779
Variant rate	1 variant every 49,088 bases	1 variant every 48,854 bases

(B) Bovine v2

	UMD3.1.94	*ARS-UCD1.2.98*
Number of variants (Before filter)	52,696	52,495
Number of variants processed	46,150	45,972
Number of effects	52,756	77,896
Genome total length	2,670,422,299	2,715,853,792
Genome effective length	2,512,082,506	2,489,385,779
Variant rate	1 variant every 54,432 bases	1 variant every 54,150 bases

(C) Bovine v3

	UMD3.1.94	*ARS-UCD1.2.98*
Number of variants (Before filter)	51,077	50,864
Number of variants processed	45,320	45,136
Number of effects	51,796	76,050
Genome total length	2,670,422,299	2,715,853,792
Genome effective length	2,512,082,506	2,489,385,779
Variant rate	1 variant every 55,429 bases	1 variant every 55,152 bases

Number of effects predicted by SnpEff program, and the genome total length increased when using the *ARS-UCD1.2* genome in all SNP chips. In Hanwoo v1 and Bovine v2 results, the SNP frequency and the genome effective length increased while decreased in Bovine v3.

3. Variants that affect protein expression

In the SnpEff results file, effects were classified according to their impact on proteins. Variants were classified according to whether they had large, small, moderate, or modifying effects. Large effects cause protein modification, whereas small effects are harmless or cause little protein deformation. Moderate-effect variants are non-disruptive but cause protein transformation, whereas modifying variants affect non-coding variants and genes. For Hanwoo v1, when *ARS-UCD1.2* was used, the number of large, small, moderate, and modifying effects increased by 43 (119.44%), 515 (81.10%), 382 (93.86%), and 26,483 (46.19%), respectively. For Bovine v2, the number of large, small, moderate, and modifying effects increased by 33 (143.48%), 476 (81.65%), 313 (94.28%), and 24,318 (46.93%), respectively. For Bovine v3, the number of large, small, moderate, and modifying effects increased by 29 (131.82%), 459 (79.41%), 312 (95.71%), and 23,454 (46.11%), respectively. The overall number of effects mapped to the *ARS-UCD1.2* reference was higher than *UMD3.1*. In all arrays, most of the variants mapped to both genome assemblies have modifying effects, followed by those with small, moderate, and large effects (**Table 4**).

The effects could also be classified by function, i.e., missense, nonsense, or silent effects. Missense effects cause amino acid changes, while nonsense effects result in the formation of the stop codon, and silent effects have no effect on amino acid synthesis. When variants in Hanwoo v1 were mapped to the current reference, the number of missense, nonsense and silent effects increased by 384 (93.43%), 37 (132.14%), and 253 (63.73%), respectively. The missense/silent ratio increased from 1.03% to 1.22%. For Bovine v2, the number of missense, nonsense, and silent effects increased by

314 (93.73%), 30 (150.00%), and 222 (59.84%), respectively. The missense/silent ratio increased from 0.90% to 1.09%. As with Bovine v2, nonsense effects were found for Bovine v3 after mapping, and the number of missense and silent effects increased by approximately 95.14% and 62.18%, respectively. Similarly, the missense/silent ratio increased from 0.92 to1.11. Overall, the missense/silent ratio increased slightly, and the number of effects for each functional class also increased (**Table 5**).

Table 4. A number of effects by impact

	Hanwoo v1		Bovine v2		Bovine v3	
	UMD3.1	*ARS-UCD1.2*	*UMD3.1*	*ARS-UCD1.2*	*UMD3.1*	*ARS-UCD1.2*
Large[w]	36	79	23	56	22	51
Small[x]	635	1,150	583	1,059	578	1,037
Moderate[y]	407	789	332	645	326	638
Modifying[z]	57,329	83,812	51,818	76,136	50,870	74,324

[w]Mutation likely to cause protein deformation;
[x]Mutation that is unlikely to cause a protein mutation;
[y]Non-disruptive mutation but causes a protein mutation;
[z]Non-coding variant or affecting non-coding genes

Table 5. The number of effects by functional class

	Hanwoo v1		Bovine v2		Bovine v3	
	UMD3.1	*ARS-UCD1.2*	*UMD3.1*	*ARS-UCD1.2*	*UMD3.1*	*ARS-UCD1.2*
Missense[w]	411	795	335	649	329	642
Nonsense[x]	28	65	20	50	18	45
Silent[y]	397	650	371	593	357	579
Missense/Silent[z]	1.03	1.22	0.90	1.09	0.92	1.11

[w]Effects which cause amino acid change;
[x]Effects that form the stop codon;
[y]Effects that have no effect on amino acid synthesis;
[z]Ratio of the number of missense effects divided into the number of silent effects

4. Effects by SNP type

For Hanwoo v1, 34,777 (59.46%) effects were found in intergenic regions, and 16,912 (28.91%) effects were found in intron regions, when the *UMD3.1* reference was used, compared to 30,506 (35.47%) and 43,567 (50.66%) effects, respectively, for *ARS-UCD1.2*. Thus, the number of variants of *ARS-UCD1.2* in intergenic regions decreased by 4,271, whereas that in intron regions increased by 26,655. For Bovine v2, 31,345 (59.34%) SNPs were found in intergenic regions, and 15,320 (29.00%) in intron regions, with the use of *UMD3.1*. Using *ARS-UCD1.2* as the reference, the respective number of SNPs was 27,462 (35.19%) and 39,752 (50.93%). Therefore, the number of SNPs in intergenic regions decreased by 3,883, and that in intron regions increased by 24,432. For Bovine v3, 30,791 (59.37%) and 15,013 (28.95%) mutations were found in the intergenic and intron regions, respectively, with *UMD3.1*, whereas 27,053 (35.50%) and 38,518 (50.55%) mutations were found, respectively, with *ARS-UCD1.2*. With the latter reference, the number of mutations in intergenic regions decreased by 3,738, and that in intron regions increased by 23,505.

For the three panels, the proportion of variants found in intergenic and intron regions was approximately 59.39% and 28.95%, respectively, when mapping was conducted using the *UMD3.1* reference. However, when the SNPs were annotated based on *ARS-UCD1.2*, approximately 35.39% of the effects were found in intergenic regions and 50.71% in intron regions. Altogether, the number of mutations in intergenic regions decreased, with a concurrent increase in intron regions. Furthermore, minor changes and genetic variations were observed in other regions (**Table 6**).

Table 6. A number of effects by variant type

	Hanwoo v1		Bovine v2		Bovine v3	
	UMD3.1	*ARS-UCD1.2*	*UMD3.1*	*ARS-UCD1.2*	*UMD3.1*	*ARS-UCD1.2*
3 prime UTR variant	339	631	319	586	309	574
5 prime UTR premature start codon gain variant	11	31	11	30	11	28
5 prime UTR variant	53	163	54	155	50	146
downstream gene variant	2,728	4,709	2,475	4,292	2,429	4,215
Initiator codon variant	1	-	1	-	1	-
intergenic region	34,777	30,506	31,345	27,462	30,791	27,053
intron variant	16,912	43,567	15,320	39,752	15,013	38,518
missense variant	407	789	332	645	326	638
non-coding transcript exon variant	36	74	37	70	34	71
splice acceptor variant	2	3	1	-	1	-
splice donor variant	3	5	-	2	1	2
splice region variant	80	172	70	151	70	150
stop gained	28	65	20	50	18	45
Stop lost	3	6	2	4	2	4
Stop retained variant	-	2	-	-	-	-
synonymous variant	573	1,009	525	923	520	905
upstream gene variant	2,538	4,276	2,314	3,925	2,291	3,851

In annotation results, effects were classified according to SNP type. The number of intergenic region effects decreased while intron variant increased in all panels when mapping was conducted using the ARS-UCD1.2 genome. Furthermore, there were minor changes and genetic variations in other regions.

Functional effects can also be categorized as causing transition or transversion. A transition is a mutation that changes from pyrimidine to pyrimidine base or purine to purine base, whereas transversion is a variant of a pyrimidine-to-purine base or purine-to-pyrimidine base. In the annotation result, the numbers of transitions and transversions are different between two references in all three panels. However, the conversion/transversion variant ratios were almost the same (**Table 7**).

Table 7. The number of transitions and transversion variants in each panel

	Hanwoo v1		Bovine v2		Bovine v3	
	UMD3.1	*ARS-UCD1.2*	*UMD3.1*	*ARS-UCD1.2*	*UMD3.1*	*ARS-UCD1.2*
Transitions[y]	64,970,889	64,727,580	5,674,827	5,653,530	1,689,642	1,683,372
Transversions[z]	18,453,817	18,363,577	1,587,304	1,579,640	469,087	466,886
Ts/Tv ratio	3.52	3.52	3.58	3.58	3.60	3.61

[y]Transition (Ts) is a mutation that changes from pyrimidine to pyrimidine base or purine to purine base;
[z]Transversion (Tv) is a variant of a pyrimidine-to-purine base or purine-to-pyrimidine base

5. Location changes in Hanwoo economic trait-related candidate genes between the references

To assess changes in physical position and information, the distributions of the *PLAG1, ACSS2,* and *CAST* gene locations, as well as information on genetic variations, were confirmed. These well-known candidate genes were associated with the economic traits of the Hanwoo breed in previous studies. The positions of the genes were verified using reference GTF annotation files downloaded from Ensembl.

The *PLAG1* (*PLAG1* zinc finger) gene is reportedly significantly associated with the carcass weight of Hanwoo cattle (S. H. Lee et al., 2012). It is located on chromosome 14 in the 25,007,291–25,009,296 bp region (length, 2,005 bp). For

Hanwoo v1, upon annotating using *UMD3.1*, one variant in the gene's downstream region (14:25,003,338) was found. In *ARS-UCD1.2*, both the locations of the gene and variant changed. In this reference genome, *PLAG1* is located in chromosome 14 in the 23,330,541–23,375,751 bp region (45,210 bp). The SNP, located in the downstream region of *UMD3.1*, moved to the 14:23,326,588 position of *ARS-UCD1.2*, still downstream of the *PLAG1* gene. No mutations in the *PLAG1* gene were detected with the other panels, regardless of the reference used (**Figure 2-A**).

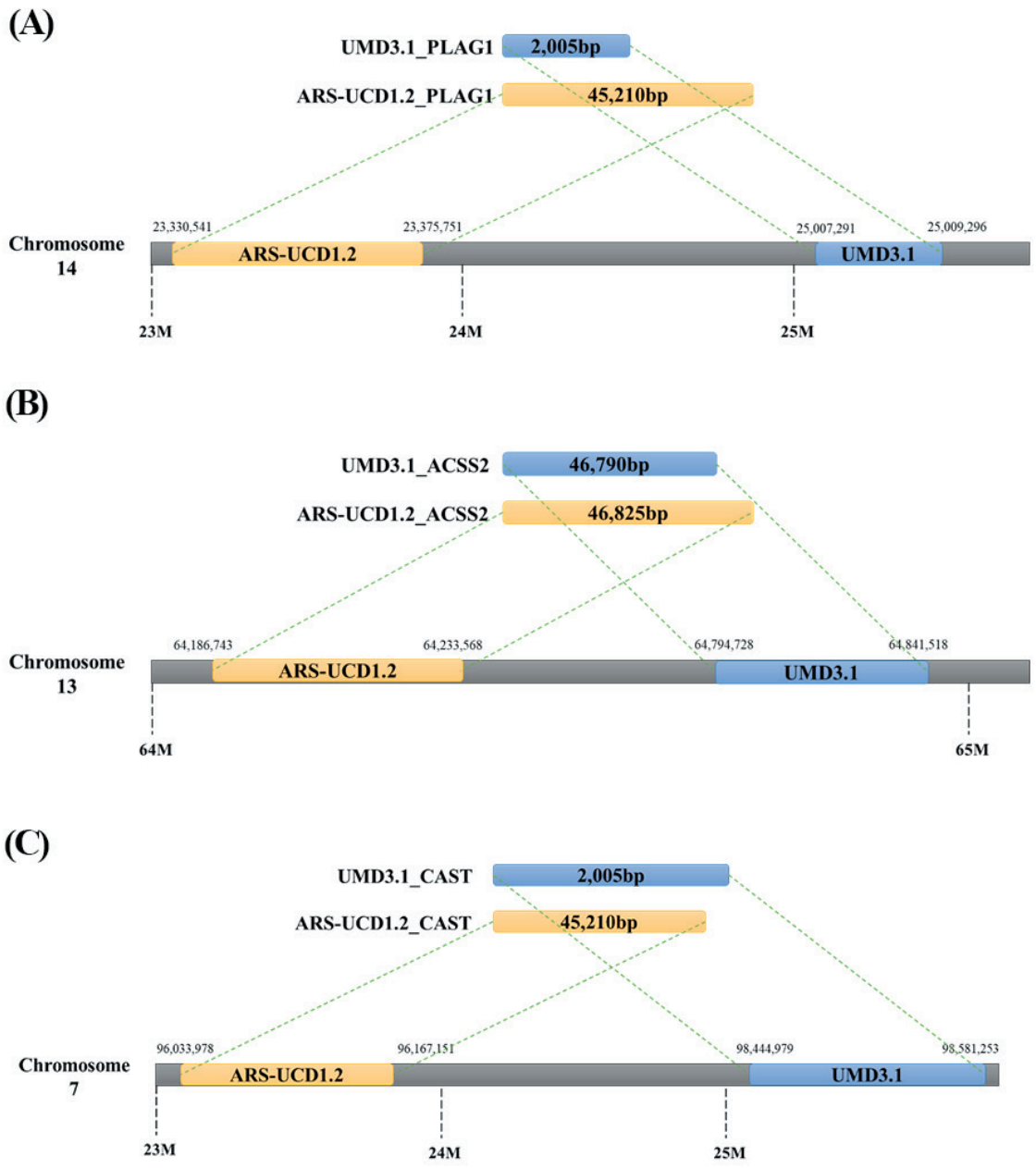

Figure 2. Gene position change according to reference genomes.
In this figure, the blue box represents the gene for the genome and the yellow box for *ARS-UCD1.2* genome. And the gray box is the genome of the chromosomes. As shown in this figure, the size and the location of the gene are changed according to the reference genomes. Due to these changes, the number and the location of SNPs are changed.

The *ACSS2* (acetyl-CoA synthetase short chain family member 2) gene is reportedly important in the Hanwoo marbling score (Kim et al., 2008). The location of this gene shifted from the 64,794,728 - 64,841,518 bp region (46,790 bp) in chromosome 13 to the 64,186,743 - 64,233,568 bp region (46,825 bp). The annotation results for the three panels indicated one intron (13:64,804,947) and one synonymous variant (13:64,840,460) in the gene with *UMD3.1* as the reference. The locations of these SNPs shifted to positions 64,196,996 and 64,232,510, respectively, and were still found within the *ACSS2* gene (**Figure 2-B**).

Finally, the *CAST* (calpastatin) gene is associated with meat tenderness (S.-H. Lee et al., 2014) and is located in the 98,444,979 - 98,581,253 bp region (136,274 bp) in chromosome 7 with *UMD3.1* as the reference. However, its location shifted to the 96,033,978 - 96,167,151 bp region (133,173 bp) in the same chromosome of the current reference. The annotation results for the three panels indicated the presence of two intron variants and one missense variant, which are located at 7:98498047, 98566391, and 7:98534197, with *UMD3.1* as the reference. In *ARS-UCD1.2*, these variants are moved to 7:96087310, 7:96152289, and 7:96119746 (**Figure 2-C**).

According to the above results, the locations and sizes of the genes changed according to the reference used. In addition, the physical positions and genetic information of all SNPs annotated in the three panels changed when the updated reference genome was used.

DISCUSSION

This study was conducted to confirm differences in genetic information when two different cattle genome sequences, *UMD3.1* and *ARS-UCD1.2*, are used. In addition, we tracked how information on SNPs changed with the use of an updated reference genome and different types of array panels.

1. Differences in the reference genome

The overall length and ungapped length of *ARS-UCD1.2* increased compared to *UMD3.1*. In particular, the scaffold N50 and contig N50 for *ARS-UCD1.2* are approximately 1.63-fold and 267.11-fold higher, respectively, indicating an improvement in the quality of the sequence (**Table 1**). The differences are likely due to advancements in sequencing technology. For example, *UMD3.1* was assembled using Sanger sequencing technology, whereas *ARS-UCD1.2* was developed using NGS technologies such as PacBio and the Illumina platforms NextSeq 500, Illumina HiSeq, and Illumina GAII. The NGS method represents an improvement over Sanger sequencing, with higher accuracy and the ability to process massive volumes of DNA in a single session (Grada & Weinbrecht, 2013). As a result, the *ARS-UCD1.2* genome coverage is approximately nine times larger than the *UMD3.1* sequence. Highly reliable SNP and indel calls can be made due to the 20-fold greater depth of NGS data. In addition, the accuracy and quality of mapping have been improved not only for annotation but also for imputation (Muzzey, Evans, & Lieber, 2015; Nicolazzi et al., 2019). Mapping the three panels to the two reference genomes showed that the genome total length increased by 45,431,493 bp (1.70%). However, the effective genome length decreased by 22,696,727 bp (0.90%) for three panels. The current reference is associated with fewer instances of haplotype non-inheritance, suggesting that it better matches true DNA sequences (Null, VanRaden, Rosen, O'Connell, & Bickhart, 2019). Therefore, *ARS-UCD1.2* is may a much more accurate sequence compared to *UMD3.1*.

2. Differences in SNP panel information according to reference genome annotations

Illumina developed the Bovine LD v2.0 Genotyping BeadChip (Bovine v2) with 54,609 SNPs for cost-effective bovine genotyping (https://www.illumina.com/documents/products/datasheets/datasheet_bovineLD.pdf)(Gunderson, Steemers, Lee, Mendoza, & Chee, 2005). The BovineSNP50 v3 BeadChip (Bovine v3) is an upgrade version of Bovine v2, which has additional features such as parentage markers identified by researchers from the US Meat Animal Research Center and US Department of Agriculture Agricultural Research Service (ARS). They compared Holstein bacterial artificial chromosome (BAC) sequence data to the bovine genome assembly (https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_bovine_snp50.pdf) (Heaton et al., 2005). In South Korea, the National Institute of Animal Science developed the Hanwoo BeadChip v1 (Hanwoo v1) for customized studies of the Hanwoo breed. To this chip, 44 markers for 24 hereditary diseases, such as abortion, achondrogenesis, and arthrogryposis, and 4,383 markers related to carcass traits of Hanwoo bulls were added (https://www.korea.kr/news/pressReleaseView.do?newsId=156250636). After lift-over process, approximately 96.5%, 96.5%, and 96.0% of the SNPs in Hanwoo v1, Bovine v2, and Bovine v3 could be located in *ARS-UCD1.2*. The change tendencies in genetic information were similar among the three panels.

3. Changes in SNP information related to amino acids and proteins

The increase in the missense/silent ratio indicates a corresponding increase in the number of genetic mutations affecting amino acid synthesis. Furthermore, when genetic variations were classified according to SNP type, the number of SNPs in intergenic areas decreased, whereas in intron regions increased. Especially, genetic variations in intron regions can have significant effects on alternative splicing (Talerico & Berget, 1990). Therefore, it is necessary to map genetic variations related to the production of amino acids, including intron and intergenic variants, again to check if they are still relevant.

4. Changes in the genetic information of Hanwoo phenotype-related candidate genes

This study identified changes in genetic information by targeting three candidate genes (*PLAG1*, *ACSS2*, and *CAST*) related to the economic traits of Hanwoo cattle. These genes mapped to both *UMD3.1* and *ARS-UCD1.2* in Hanwoo v1. However, the locations and sizes of the genes differed between the two reference genomes. In addition, the number and types of genetic variants located in the current reference genome also changed with these changes. In a wide-ranging study, the *CPAMD8* gene, associated with congenital Morgagnian cataracts in Holstein calves, was mapped to the whole genome using SAS version 9.4. In the *UMD3.1* reference genome used in this study, the *CPAMD8* gene is located in the 5,995,888–6,095,676 bp region in chromosome 7. But the *ARS-UCD1.2* reference genome is located in the 6,073,378–6,174,087 bp region in the same chromosome (Braun et al., 2019). As the reference sequence has been updated, the genetic information associated with the sequence has changed completely. Therefore, mapping in previous studies using *UMD3.1* may need to be re-analyzed using *ARS-UCD1.2*, and the results should be compared. As the genetic information differs between the two references, it is recommended that only a single reference be used for analysis.

CONCLUSION

Microarrays mapped to the *UMD3.1* reference genome were remapped to the *ARS-UCD1.2* version, achieving 96.0-96.5%. As a result, the number of SNPs affecting protein expression, the number of effects by SNP type, and the location of candidate genes related to Hanwoo economic traits were different between the reference genomes. Therefore, the map file of *UMD3.1*-based SNP chip data needs to update and use the location information of *ARS-UCD1.2*, a new reference genome, for accurate comparison with other currently published studies.

ACKNOWLEDGEMENTS

This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (321082-3)

REFERENCES

Boddhireddy, P., Prayaga, K., Barros, P., Lôbo, R., & DeNise, S. (2014). Genomic predictions of economically important traits in Nelore cattle of Brazil. Paper presented at the Proceedings of the 10th World Congress of Genetics Applied to Livestock Production (Vancouver, BC).

Braun, M., Struck, A.-K., Reinartz, S., Heppelmann, M., Rehage, J., Eule, J. C., . . . Distl, O. (2019). Study of congenital Morgagnian cataracts in Holstein calves. PLoS One, 14(12), e0226823. doi:10.1371/journal.pone.0226823

Chen, Y., Cunningham, F., Rios, D., McLaren, W. M., Smith, J., Pritchard, B., . . . Flicek, P. (2010). Ensembl variation resources. BMC Genomics, 11(1), 293. doi:10.1186/1471-2164-11-293

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., . . . Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly, 6(2), 80-92. doi:10.4161/fly.19695

Grada, A., & Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. The Journal of investigative dermatology, 133(8), e11.

Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G., & Chee, M. S. (2005). A genome-wide scalable SNP genotyping assay using microarray technology. Nature Genetics, 37(5), 549-554. doi:10.1038/ng1547

Heaton, M. P., Keen, J. E., Clawson, M. L., Harhay, G. P., Bauer, N., Shultz, C., . . . Laegreid, W. W. (2005). Use of bovine single nucleotide polymorphism markers to verify sample tracking in beef processing. Journal of the American Veterinary Medical Association, 226(8), 1311-1314.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. Genome Res, 12(6), 996-1006. doi:10.1101/gr.229102

Kim, N.-K., Kim, S.-K., Heo, K.-N., Yoon, D., Lee, C.-S., Im, S.-K., & Park, E.-W. (2008). Expression profiles of triacylglycerol biosynthesis genes on fattening stages in Hanwoo. Journal of Animal Science and Technology, 50(3), 293-300. Kong, J., Huh, S., Won, J.-I., Yoon, J., Kim, B., & Kim, K. (2019). GAAP: A Genome Assembly + Annotation Pipeline. BioMed Research International, 2019, 4767354. doi:10.1155/2019/4767354

Lee, S.-H., Kim, S.-C., Chai, H.-H., Cho, S.-H., Kim, H.-C., Lim, D., . . . Hong, S.-K. (2014). Mutations in calpastatin and μ-calpain are associated with meat tenderness, flavor and juiciness in Hanwoo (Korean cattle): Molecular modeling of the effects of substitutions in the calpastatin/μ-calpain complex. Meat Science, 96(4), 1501-1508. https://doi.org/10.1016/j.meatsci.2013.11.026

Lee, S. H., Lim, D., Jang, G. W., Cho, Y. M., Choi, B. H., Kim, S. D., . . . Park, E. W. (2012). Genome Wide Association Study to Identity QTL for Growth Taits in Hanwoo. Journal of Animal Science and Technology, 54(5), 323-329.

Muzzey, D., Evans, E. A., & Lieber, C. (2015). Understanding the Basics of NGS: From Mechanism to Variant Calling. Current Genetic Medicine Reports, 3(4), 158-165. doi:10.1007/s40142-015-0076-8

Nicolazzi, E. L., Bacheller, L. R., Fok, G. C., Gaddis, K. L. P., Jensen, L., Megonigal Jr, J. H., . . . Wiggans, G. R. (2019). Enhancements to US genetic and genomic evaluations in 2018 and 2019. Interbull Bulletin(55), 26-29.

Null, D., VanRaden, P. M., Rosen, B., O'Connell, J., & Bickhart, D. (2019). Using the ARS-UCD1.2 reference genome in US evaluations. Interbull Bulletin(55), 30-34.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., . . . Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics, 81(3), 559-575. https://doi.org/10.1086/519795

Talerico, M., & Berget, S. M. (1990). Effect of 5'splice site mutations on splicing of the preceding intron. Molecular and cellular biology, 10(12), 6299-6305.

Uemoto, Y., Abe, T., Tameoka, N., Hasebe, H., Inoue, K., Nakajima, H., . . . Kobayashi, E. (2011). Whole-genome association study for fatty acid composition of oleic acid in Japanese Black cattle. Animal Genetics, 42(2), 141-148. https://doi.org/10.1111/j.1365-2052.2010.02088.x