In silico approaches to identify the functional and structural effects of non-synonymous SNPs in selective sweeps of the Berkshire pig genome

Donghyun Shin¹, Jae-Don Oh¹, Kyeong-Hye Won¹, and Ki-Duk Song^{1,2,*}

* Corresponding Author: Ki-Duk Song
Tel: +82-63-219-5523, Fax: +82-63-270-5937,
E-mail: kiduk.song@gmail.com

¹ Department of Animal Biotechnology, Chonbuk National University, Jeonju 54896, Korea

² The Animal Molecular Genetics and Breeding Center, Chonbuk National University, Jeonju 54896, Korea

ORCID

Donghyun Shin

<https://orcid.org/0000-0002-0819-0553>

Jae-Don Oh

<https://orcid.org/0000-0001-7756-1330>

Kyeong-Hye Won

<https://orcid.org/0000-0001-8112-2840>

Ki-Duk Song

<https://orcid.org/0000-0003-2827-0873>

Submitted Mar 19, 2017; Revised Jun 24, 2017;
Accepted Jan 30, 2018

Objective: Non-synonymous single nucleotide polymorphisms (nsSNPs) were identified in Berkshire selective sweep regions and then were investigated to discover genetic nsSNP mechanisms that were potentially associated with Berkshire domestication and meat quality. We further used bioinformatics tools to predict damaging amino-acid substitutions in Berkshire-related nsSNPs.

Methods: nsSNPs were examined in whole genome resequencing data of 110 pigs, including 14 Berkshire pigs, generated using the Illumina HiSeq2000 platform to identify variations that might affect meat quality in Berkshire pigs.

Results: Total 65,550 nsSNPs were identified in the mapped regions; among these, 319 were found in Berkshire selective-sweep regions reported in a previous study. Genes encompassing these nsSNPs were involved in lipid metabolism, intramuscular fatty-acid deposition, and muscle development. The effects of amino acid change by nsSNPs on protein functions were predicted using sorting intolerant from tolerant and polymorphism phenotyping V2 to reveal their potential roles in biological processes that may correlate with the unique Berkshire meat-quality traits.

Conclusion: Our nsSNP findings confirmed the history of Berkshire pigs and illustrated the effects of domestication on generic-variation patterns. Our novel findings, which are generally consistent with those of previous studies, facilitated a better understanding of Berkshire domestication. In summary, we extensively investigated the relationship between genomic composition and phenotypic traits by scanning for nsSNPs in large-scale whole-genome sequencing data.

Keywords: Berkshire; Functional Annotation; Next Generation Sequencing; Non-synonymous Single Nucleotide Polymorphisms; Selective Sweep

INTRODUCTION

The domestic pig, *Sus scrofa domestica*, has long been an important source of dietary protein for humans and has undergone natural selection in response to various environmental factors over time. Furthermore, pig breeds have been subject to intensive artificial selection to increase economically important traits such as reproduction, growth rate, stress resistance, and meat quality [1,2]. In particular, Berkshire pigs have been under intensive artificial selection, particularly to ensure rapid and efficient muscle accumulation and desirable pork qualities [3]. As a result of strong selection, the quality of Berkshire meat is considered superior, as it contains a high proportion of neutral lipid fatty acids and good marbling fat [4], both of which are important factors contributing to palatability traits such as tenderness and juiciness [5]. These unique meat qualities, as well as biochemical and physical features, have been investigated in a sensory panel analysis [6-8]. Previous studies have also focused

on potential associations of candidate genes with meat quality in Berkshire pigs [9-11].

Several studies have investigated genetic factors related to meat quality in the Berkshire population using next generation sequencing and whole genome sequencing data [1,6,7,10, 11]. This technology provides a powerful approach for identifying a massive number of single nucleotide polymorphisms (SNPs) in genome-wide sequence data [12,13,]. Selective sweeps, which were inferred from whole genome sequencing data using cross-population extended haplotype homozygosity (XP-EHH) and a cross-population composite likelihood ratio method (XP-CLR), have revealed distortions in genetic patterns between Berkshire and two other maternal breeds (Yorkshire and Landrace) [1]. These selective sweeps reveal strong selection signatures in genomic regions that harbor economic trait-related genes. Therefore, the identification of positively selected genetic regions, especially in Berkshire pigs might allow the exploration of genetic mechanisms related to Berkshire-specific phenotypic traits mentioned above, including rapid and efficient muscle accumulation and desirable pork qualities etc.

Although many SNPs are phenotypically neutral, nonsynonymous SNPs (nsSNPs) that are located in protein-coding regions and cause amino-acid substitutions in corresponding protein products, can alter protein structure, stability, or function and are typically associated with disease. In human, nsSNPs cause approximately 50% mutations involved in inherited genetic diseases [14-16]. In addition, multiple nsSNPs in innate immune genes can affect susceptibility to infection, as well as the development of inflammatory disorders and autoimmune diseases [17-19]. The effects of nsSNPs on porcine genetic mechanisms remain under investigation.

In this study, we performed a genome-wide annotation of nsSNPs of the genes located in selective sweeps of Berkshire pigs in regard to protein function, using *in silico* bioinformatics tools. Notably, nsSNPs are relevant to biological functions reflective of breed characteristics [20,21]. Accordingly, we integrated and reanalyzed the whole-genome sequencing data of Berkshire, Duroc, Korean native pig, Landrace, miniature pig, and Yorkshire pigs, and wild boar, to identify genomic variants. The nsSNPs were identified in Berkshire selective sweep regions and investigated to discover genetic mechanisms that were potentially associated with Berkshire domestication and meat quality. We further used bioinformatics tools to predict damaging amino-acid substitutions caused by novel nsSNPs of genes in Berkshire pig genome.

MATERIALS AND METHODS

Sample preparation and whole genome sequencing

The whole genome sequence data set used in this study included data of 14 Berkshire, 26 Duroc, 14 Korean native pigs,

14 Landrace (Danish), 16 miniature pigs, and 16 Yorkshire (Large White) pigs, and 10 wild boars. All data were obtained from the NCBI Sequence Read Archive database (SRP047260, SRP052927, PRJNA254936, and PRJNA281548). Especially, Berkshire, Landrace, and Yorkshire (Large White) among the 110 pigs were used in a previous study [22] to explore the evidence suggesting a genetic basis for the positive selection of meat-quality traits in Berkshire pigs.

FastQC software [23] was used to perform a quality check of raw sequence data, and Trimmomatic-0.32 [22] to remove potential adapter sequences before sequence alignment. Paired-end sequence reads were mapped to a pig reference genome (Sscrofa 10.2.75) from the Esembl database, using Bowtie2 [24] with default settings. The open-source software packages, i.e., Picard tools (<https://broadinstitute.github.io/picard/>), SAMtools [25], and genome analysis toolkit (GATK) [26] were used for downstream processing and variant calling. The Picard “CreateSequenceDictionary” and “MarkDuplicates” command-line tools were used to read reference FASTA sequences for writing bam files with only a sequence dictionary and to filter potential polymerase chain reaction duplicates, respectively. We used SAMtools to create index files for the reference and bam files, and then used the GATK “Realigner-Target-Creator” and “IndelRealigner” arguments to perform a local realignment of sequence reads to correct misalignments caused by small insertions and deletions. Base quality scores also were recalibrated to ensure accuracy and correct machine cycle-related and sequence context-related quality variations. GATK “Unified Genotyper” and “Select Variants” arguments were used to call variants according to the following filtering criteria: i) a Phred-scaled quality score of <30; ii) read depth of <5; and iii) MQ0 (total count of mapping quality zero reads across all samples ≥ 4). Phred-scaled p-values (Fisher’s exact test) >200 were filtered out to decrease false-positive calls because of strand bias. The “vcf-merge” tool in the VCF tools package [27] were used to merge all variant-calling format files for the 110 samples.

Population structure analysis

A population structure analysis was performed to infer the population structures from whole-genome sequence data of 110 pigs. The STRUCTURE program (version 2.3.4) was used to evaluate substructure among the 110 individuals from seven breeds. A Bayesian clustering analysis in this program [28,29] was used to estimate the population structure, using 65,550 nsSNPs identified from the whole-genome sequencing data of all 110 pigs. We determined that an initial burn-in of 100,000 iterations, followed by 100,000 iterations for parameter estimation, would be sufficient to ensure the convergence of parameter estimates. To identify admixtures (K parameter, STRUCTURE), we analyzed the dataset after allowing for K values of 4 and 6.

Identification of nsSNPs in Berkshire selective-sweep regions

ClueGO [30] was used to perform a gene ontology (GO) network analysis for the identified genes to infer the biological meanings of the related nsSNPs. In addition, empirical p-values were defined according to rankings to further investigate multiple nsSNP-containing genes. In the Berkshire population, genes were ranked in descending order by nsSNP number with respect to the empirical p-value ranking. In this study, “p-values” denotes empirical p-values; here, a low p-value indicates that a locus is an outlier relative to the rest of the genome.

Prediction of damaging amino-acid substitutions in nsSNPs in Berkshire pigs

In this study, the functional effects of amino acid sequence change by nsSNPs were predicted using the following *in silico* algorithms: sorting intolerant from tolerant (SIFT) [31,32] and polymorphism phenotyping V2 (Polyphen-2) [33]. SIFT uses sequence homology to predict the effects of the amino-acid substitution of interest on protein functions. nsSNP-based sequence homology tests were performed to identify important amino-acid substitutions that might affect biological functions via structural modifications of proteins. In our analysis of 319 nsSNPs from Berkshire selective-sweep regions (Table 1), a SIFT score of <0.05 indicated that the SNP was deleterious and could strongly affect protein function. Moreover, a PolyPhen-2 (version 2.2.2) analysis were performed with specific empirical rules to predict the effects of amino-acid substitutions on the structures and functions of proteins [3], using amino-acid encoded by genes containing target SNPs from the Esembl database. After SIFT analysis, proteins resulting from 319 nsSNPs in Berkshire selective-sweep regions were tested using Polyphen-2. Tested nsSNPs were classified as probably damaging, possibly damaging, or benign if they received Polyphen-2 scores (range: 0 to 1) of >0.95, 0.5 to 0.95, or <0.5, respectively [33]. In this study, we expected that probably damaging and possibly damaging SNPs would affect protein functions. The results from our analyses of the effects of nsSNPs on protein function are shown in Tables 1, 2.

RESULTS

DNA sequencing data preprocess and genetic variant calling

After excluding tri-allelic SNPs (283,097 SNPs; 0.91% of 31,113,755 total SNPs), a total of 30,830,658 SNPs were extracted from the whole-genome sequences of 110 pigs, including 14 Berkshire individuals (Supplementary Table S1), and annotated all extracted SNPs using SnpEff version 4.1a [34]. Subsequently, all SNPs were divided into 19 functional classes, including nsSNPs (Supplementary Figure S1). Most SNPs were located in intergenic or intron regions, and 65,550 nsSNPs (0.223% of total SNPs) were identified in 12,469 protein-coding genes from all SNPs in genic regions (except splice regions). Of these nsSNPs, SNP frequency of 26,644 were not zero, and 8,177 genes in Berkshire population contained two or more SNPs. A population structure analysis was performed based on genotypic information to evaluate the genetic relationships among breeds, and found that Landrace and Yorkshire pig breeds and miniature pigs were clustered into one group and were distinct from Duroc pigs, Korean native pigs, and wild boars ($K = 4$ in Figure 1). However, Berkshire populations were distinct from Yorkshire, Landrace, and miniature pigs at $K = 6$ (Figure 1).

Identification of nsSNPs in Berkshire selective-sweep regions

In this study, 155 genes (56 genes from XP-EHH plus 114 genes from XP-CLR) were identified in selective-sweep regions of the Berkshire genome. Furthermore, 319 of the total 65,550 nsSNPs belonged to 73 of 155 genes in Berkshire selective sweep regions (Supplementary Table S2). Genotype information for these 319 SNPs in Berkshire, Landrace, and Yorkshire populations is presented in Figure 2. In addition, by defining empirical p-values according to rankings to further investigate multiple nsSNP-containing genes, and observed 12,469 genes in our 110-pig population contained multiple SNPs. In the Berkshire population, we ranked genes in descending order by nsSNP number with respect to the empirical p-value ranking. The p-values were assigned to 12,469 genes;

Table 1. Summarization of nonsynonymous single amino-acid variations in genes from a Berkshire selective-sweep according to SIFT and Polyphen-2 scores

		Polyphen-2				
		Possibly damaging	Probably damaging	Benign	Unknown	Total
SIFT	Deleterious	5	21	17	1	44
	Deleterious low confidence	0	1	3	1	5
	Tolerated	13	17	204	7	241
	Tolerated low confidence	0	1	17	7	25
	Unknown	0	0	4	0	4
	Total	18	40	245	16	319

SIFT, sorting intolerant from tolerant; Polyphen-2, polymorphism phenotyping V2.

Table 2. Top 27 nonsynonymous single-nucleotide polymorphisms (SNPs) with strong effects on residues that stabilize protein functions

SNP	SNP rs #	CHR	POS	REF	ALT	Gene (Ensembl gene ID)	FREQ BKS	FREQ other breed	SIFT result	SIFT score	Polyphen-2 result	Polyphen-2 score
1_136765966	rs325343029	1	136,765,966	G	T	CEP152 (ENSSSCG00000004657)	0.893	0.401	Deleterious	0.020	possibly damaging	0.717
4_8206635	rs331950721	4	8,206,635	C	T	TG (ENSSSCG00000005948)	0.071	0.010	Deleterious	0.000	probably damaging	1.000
4_8315405	rs327392897	4	8,315,405	G	A		0.071	0.057	Deleterious	0.030	probably damaging	0.986
4_8406938	rs328014000	4	8,406,938	G	A		0.036	0.047	Deleterious	0.010	probably damaging	0.976
4_8406980	rs337167711	4	8,406,980	T	C		0.321	0.047	Deleterious	0.000	probably damaging	1.000
5_105863044	rs336428086	5	105,863,044	C	T	OTOGL (ENSSSCG00000000943)	0.500	0.000	Deleterious	0.000	probably damaging	0.999
6_155183186	.	6	155,183,186	C	T	SZT2 (ENSSSCG000000003941)	0.036	0.005	Deleterious	0.000	probably damaging	0.998
6_156247324	rs331099959	6	156,247,324	A	G	CCDC30 (ENSSSCG000000003965)	0.036	0.016	Deleterious	0.010	probably damaging	0.995
6_156274435	rs321401589	6	156,274,435	T	G		0.214	0.141	Deleterious	0.000	probably damaging	0.998
6_156274449	rs330010762	6	156,274,449	C	T		0.321	0.266	Deleterious	0.030	probably damaging	0.991
6_156282099	.	6	156,282,099	G	T		0.500	0.453	Deleterious	0.040	probably damaging	1.000
6_156282110	.	6	156,282,110	T	G		0.464	0.432	Deleterious	0.000	probably damaging	0.999
6_156287368	rs342319002	6	156,287,368	C	G		0.036	0.031	Deleterious	0.000	probably damaging	0.999
6_156287384	rs335124185	6	156,287,384	C	T		0.071	0.125	Deleterious	0.000	probably damaging	1.000
6_156314992	rs323439800	6	156,314,992	C	A		0.286	0.260	Deleterious	0.040	probably damaging	0.993
8_10894255	rs335723841	8	10,894,255	A	C	PROM1 (ENSSSCG000000008745)	0.321	0.391	Deleterious	0.040	probably damaging	1.000
9_2311094	rs343636299	9	2,311,094	C	T	OVCH2 (ENSSSCG000000021619)	0.107	0.073	Deleterious	0.020	probably damaging	1.000
9_9183647	rs344056896	9	9,183,647	G	A	C2CD3 (ENSSSCG000000014835)	0.143	0.109	Deleterious	0.000	possibly damaging	0.814
10_71332113	rs319726102	10	71,332,113	G	A	FAM208B (ENSSSCG000000011136)	0.036	0.135	Deleterious	0.010	possibly damaging	0.781
12_3353997	.	12	3,353,997	G	A	USP36 (ENSSSCG000000017165)	0.071	0.005	Deleterious	0.030	probably damaging	0.987
12_46832762	rs334891445	12	46,832,762	A	G	SPAG5 (ENSSSCG000000017758)	0.607	0.260	Deleterious low confidence	0.000	probably damaging	0.999
13_190959574	.	13	190,959,574	T	C	USP25 (ENSSSCG000000024623)	0.143	0.000	Deleterious	0.010	possibly damaging	0.916
13_216635063	rs325814243	13	216,635,063	G	C	RRP1B (ENSSSCG000000021698)	0.107	0.026	Deleterious	0.000	probably damaging	0.999
13_216641266	rs323442412	13	216,641,266	C	T		0.071	0.037	Deleterious	0.020	probably damaging	0.998
14_58666614	.	14	58,666,614	G	A	ACTN2 (ENSSSCG000000010144)	0.036	0.005	Deleterious	0.010	probably damaging	1.000
16_6458337	rs343777401	16	6,458,337	C	T	MYO10 (ENSSSCG000000016794)	0.143	0.365	Deleterious	0.000	probably damaging	1.000
16_34097125	rs331591905	16	34,097,125	C	T	ITGA1 (ENSSSCG000000016885)	0.036	0.031	Deleterious	0.030	possibly damaging	0.921

SIFT, sorting intolerant from tolerant; Polyphen-2, polymorphism phenotyping V2.

those with p-values of <0.05 (5%) as were thought to contain significantly large numbers of nsSNP and were assumed to be strongly associated with Berkshire characteristics. Thirteen of 73 genes with nsSNPs in the Berkshire selective-sweep region had p-values of <0.05 (Supplementary Table S2).

nsSNPs in the Berkshire selective-sweep region

Detailed information of genes that were analyzed in this study is presented in Supplementary Table S2, and the findings of a GO network analysis for 73 genes are presented in Figure 3 and Supplementary Table S3. Through our GO network analysis, 17 of 73 genes, i.e., ATP binding cassette subfamily A

Figure 1. STRUCTURE-based population structure analysis. Individuals are represented by vertical bars, and the length of each colored segment in each vertical bar represents the proportion contributed by an ancestral population. (a) Four colors (K = 4) mostly represent the population structures of 110 individuals. (b) Six colors (K = 6) represent the population structures of 110 individuals.

Figure 2. Genotypes of 319 nonsynonymous single-nucleotide polymorphisms (SNPs) in Berkshire selective-sweep regions. The genotype patterns of 319 nonsynonymous SNPs in Berkshire selective-sweep regions extracted from 110 individuals are represented by a heat map. Box colors represent the genotypes of individuals determined using whole-genome sequencing data. Dark blue boxes indicate that both alleles were the minor allele. Blue boxes indicate that one each of two alleles was the minor and major allele. Sky blue boxes indicate meant that both alleles were the major allele. At left, each SNP name is listed as chromosome, position, and minor allele type. The gray box at the bottom of the figure indicates the three evaluated breeds. Because previous Berkshire selective sweeps were extracted from Berkshire, Yorkshire, and Landrace whole-genome data, we displayed SNP information for these breeds.

member 13 (*ABCA13*), actin alpha2 (*ACTN2*), ATPase phospholipid transporting 9A (*ATP9A*), C2 calcium dependent domain containing 3 (*C2CD3*), centrosomal protein 152 (*CEP152*), growth hormone (*GH*), glucagon like peptide 2 receptor (*GLP2R*), GNAS complex locus (*GNAS*), interleukin 22 receptor subunit alpha 2 (*IL22RA2*), microtubule associated protein 1 light chain 3 beta (*MAP1LC3B*), I-BAR domain containing (*MTSS1*), olfactomedin 1 (*OLFM1*), slow-mo homolog 2 (*SLMO2*), prominin 1 (*PROM1*), sperm associated antigen 5 (*SPAG5*), transforming growth factor beta receptor 3 (*TGFBR3*), ubiquitin specific peptidase 36 (*USP36*), were found associated with six major GO terms. *TGFBR3* and *GLP2R*, which have been previously reported

as major candidate meat-quality genes according to the signatures of selection scan of the whole genome sequences in Berkshire pigs [1], were closely related to the cellular response to glucagon stimulus, one of six significant GO terms (Supplementary Table S3). In the previous study [35], a SNP in the intron region of *GLP2R* associated significantly with the intramuscular fat in a Large White×Minzhu intercross population. Moreover, we detected a novel nsSNP (Chr12:574,612,64) in the exon region of *GLP2R* [35]. In general, glucagon induces lipolysis in mammals under conditions of insulin suppression, and we considered that amino-acid substitutions caused by nsSNPs in the Berkshire genome might affect lipolysis, thus affecting the high proportions of neutral lipid fatty acids and

Figure 3. Gene ontology (GO) network of genes related to non-synonymous single nucleotide polymorphisms (nsSNPs) in the Berkshire selective-sweep region. We obtained significant GO analysis results using genes related to nonsynonymous SNPs in the Berkshire selective-sweep region and our criteria in ClueGO Cytoscape packages (number of genes = 2, sharing group percentage = 50.0). Subsequently, these results were largely divided into six clusters.

marbling fat associated with this breed. *TGFBR3* plays important roles in muscle and adipose tissue development [36], and increased transcription of *TGFBR3* mRNAs was observed in pigs with high lipid deposition and composition. We therefore assumed that four genes (*GNAS*, *OLFM1*, *GLP2R*, *TGFBR3*) related to cellular responses to glucagon stimulus would be affected by selective pressure over time, to establish characteristics of Berkshire that have been selected artificially.

Three genes (*C2CD3*, *CEP152*, *SPAG5*) were associated with protein localization to the centrosome, cytoskeleton, and microtubule cytoskeleton (Figure 3). Rutkowski et al [37] reported that the coalescing mechanism associated with large lipid droplet formation was microtubule-dependent, and therefore disruption of cytoskeletal remodeling would affect lipid droplet maintenance and dynamics during lipolysis. Furthermore, dynamic changes in adipocyte size (expansion and reduction) are thus critically dependent on the environments [37]. Our comparison of SNP frequencies in *C2CD3*, *CEP152*, and *SPAG5* led us to consider that the high levels of genetic variation might be linked to the high lipid deposition in Berkshire pigs (Figure 4). *ABCA13*, *ATP9A*, and *PRELID3B* are involved in phospholipid transport. Notably, the fatty acids of phospholipids are particularly important in developing pork flavor, and obvious genetic effects have been observed on the total fatty-acid concentrations in muscle tissues. Another study observed higher phospholipid concentrations in the *longissimus* and *psoas* muscles of Berkshire purebred pigs compared to Large White pigs [38]. Once engaged with a specific receptor or ligand, GH and IL22RA2 can transmit signals via tyrosine-phosphorylated STAT3 to regulate the transcriptional

activities of lipid metabolism-related genes [39].

We ranked genes by empirical p-values based on number of nsSNP and number of nsSNPs among the population of 110 pigs. Overall, 715 of 12,469 genes had p-values of <0.05 and contained at least eight SNPs. We assumed that these 715 genes would be significantly affected in Berkshire pigs; in other words, those genes would be strongly associated with Berkshire domestication and characteristic establishment. Among these 715 genes, we identified 13 nsSNP-containing genes with significant p-value-based ranks in the Berkshire selective-sweep region. The highest rank gene was *C2CD3* among 13 genes, which encodes a protein that regulates centriole elongation. In this study, *C2CD3*, along with *CEP152* and *SPAG5*, was associated with protein localization to the centrosome, cytoskeleton, and microtubule cytoskeleton in the GO network analysis. Interestingly, *CEP152* (p-value = 0.045) and *SPAG5* (p-value = 0.036) were also included among the 13 significant genes. Details of the nsSNP analysis results are summarized in Supplementary Figure S1.

SIFT analysis of nsSNPs in the Berkshire selective-sweep region

According to the SIFT analysis, 49 of 319 nsSNPs were classified as deleterious (some were deleterious with low confidence). PolyPhen-2 was used to calculate the true-positive rate as a fraction of predicted mutations (Figure 1), and 58 nsSNP-related amino-acid variants (sum of possibly damaging and probably damaging in Table 1) in the Berkshire selective-sweep region may cause deleterious functional effects on the associated genes; these findings overlapped with SIFT results for

Figure 4. Single-nucleotide polymorphism (SNP) frequency in *C2CD3*, *SPAG5*, and *CEP152*. *C2CD3*, *CEP152*, and *SPAG5* were associated with protein localization to the centrosome in this gene ontology network analysis. As shown, we compared the allele frequencies of SNPs in these three genes in the Berkshire and other porcine populations (a: SNPs in *C2CD3*, b: *SPAG5*, and c: *CEP152*). Blue bars indicate allelic frequency in the Berkshire population, and orange bars indicate allelic frequencies in other populations. *C2CD3*, C2 calcium dependent domain containing 3; *SPAG5*, sperm associated antigen 5; *CEP152*, centrosomal protein 152.

27 nsSNPs. Therefore, it is reasonable to assume that these 27 SNPs may have a considerable impact on biological mechanisms during Berkshire domestication (Table 1). Thyroglobulin (*TG*), coiled-coil domain-containing protein 30 (*CCDC30*), and ribosomal RNA processing 1B (*RRP1B*) each contained multiple SNPs and have strong effects on protein stabilization (Table 2). In particular, the ranked p-value of *TG* and *CCDC30* were 0.05 or less. *CCDC30* contained 25 nsSNPs in the Berkshire population; of these, eight SNPs, i.e., rs331099959, rs321401589, rs330010762, rs342319002, rs335124185, rs323439800 and 2 novel SNPs, were identified to have considerable effects in both the SIFT and PolyPhen-2 analyses. However, the detailed physiological mechanism of *CCDC30* in the Berkshire domestication process remains to be studied. *TG* plays an important role in adipocyte growth and differentiation. In this study, among the 17 nsSNPs of *TG* in Berkshire pigs, 4 nsSNPs were identified to have considerable impacts on protein stabilization. Notably, a previous study identified *TG* as a major candidate genes related to meat quality via Berkshire positive selection scans (XP-EHH and XP-CLR) [1]. In the Berkshire population, *RRP1B*, which encodes a ribosomal

RNA processing protein homolog, contained two nsSNPs with potentially strong effects on protein residue stabilization. Previously, Grant et al [40] demonstrated the differential transcription of *RRP1B* in lean and obese individuals [40], which led us to presume that *RRP1B* may be strongly correlated with meat quality in domesticated Berkshire pigs. Although *C2CD3* contained the highest number of Berkshire nsSNPs among the 12,469 evaluated genes, only one of the 36 SNPs was thought to have a strong effect on protein function. Like *C2CD3*, *CEP152*, and *SPAG5* each had ranking p-values of <0.05 but there is only one SNP thought to cause a significant amino-acid substitution (Table 2). Similarly, otogelin like (*OTOGL*), seizure threshold 2 (*SZT2*), family with sequence similarity 208 member B (*FAM208B*), and integrin subunit alpha 1 (*ITGA1*) each had ranking p-values of <0.05 but only one SNP expected to affect protein function.

DISCUSSION

Regarding the interest of the meat-production industry in quality improvement, a previous study performed a selective

sweep-based genetic investigation of Berkshire pigs to determine information essential to domestic breeding [1]. Various SNPs were used in these selective sweeps. A nonsynonymous substitution of amino acid by a nsSNP may result in biological changes and are therefore subject to natural selection. An evaluation of positive selection throughout the Berkshire whole genome sequence data indicated the need for an advanced, nsSNP-based interpretation. In this study, we aimed to predict and interpret the impacts of amino acid substitutions by nsSNPs of Berkshire pigs by *in silico* tools, under the assumption that these elements might associate with important and previously unrevealed biological mechanisms related to Berkshire characteristics. Our deep analyses identified 319 nsSNPs in the Berkshire selective sweep regions that were subjected to a GO network analysis. Here, the SNPs were classified according to six major GO terms, most of which were associated with process involved in lipid metabolism.

We investigated the nsSNPs of seven major meat-quality candidate genes, i.e., *TG*, fatty acid binding protein 1, Akirin 2, *TGFBR3*, junctophilin 3, intercellular adhesion molecule 2, and endoplasmic reticulum to nucleus signaling 1, detected via positive selection scans in a previous study [1]. In the Berkshire population, three of the genes did not contain nsSNPs; of the remaining four, all genes contained at least one nsSNP, and *JPG3* contained three nsSNPs. Only *TG* contained 17 nsSNPs. Accordingly, we searched additional genes that have a potential impact on Berkshire domestication. First, we identified several genes associated with various biological processes associated with lipid metabolism (e.g., cellular responses to glucagon stimulus, protein localization to centrosome, and phospholipid transport and tyrosine phosphorylation of STAT3 protein) through a GO network analysis (Figure 3; Supplementary Table S3). Second, we conducted a differential search for genes that were enriched for nsSNPs specifically in the Berkshire population. Each nsSNP has been assigned a ranking based on p-values and identified 13 genes that met our criterion (p-value <0.05) (Supplementary Table S2). Of these, interestingly, only *TG* overlapped with previously detected major meat-quality candidate genes [1]. Furthermore, all three genes (*C2CD3*, *CEP152*, *SPAG5*) associated with protein localization to the centrosome in our GO ontology were included among this 13-gene set, and contained more nsSNPs than others in the Berkshire selective-sweep region. Surprisingly, most SNPs in these three genes were present at very high frequencies in the Berkshire population compared to other pig breeds. *C2CD3* contained 36 nsSNPs, the highest number in the Berkshire population, and most were present at high allele frequencies. We expect that the variants in these three genes likely played a crucial role in Berkshire domestication and therefore warrant additional studies in the future. Third, we used SIFT [32] and PolyPhen-2 [33] to predict the possible effects of the 319 nsSNPs in Berkshire selective-sweep regions

on protein function. Twenty-seven SNPs were predicted associated with biological functions at the protein levels. These results suggest that genes related to these 27 SNPs were associated with Berkshire meat quality.

Our results strongly suggest that genetic variants in Berkshire pigs are closely associated with important meat-quality factors such as increased lipid deposition and composition. Our study is distinct from others because of its use of nsSNPs to evaluate functional impacts of amino acid change caused by nsSNPs in the Berkshire genome. Nonetheless, this study has limitations; first, even various bioinformatics tools were adapted for the analyses but information that we studied in this study come from human variants database as there is no livestock specific database to evaluate functional impact of nsSNPs so far. Also, this study lacks experimental validation on the functionality of nonsynonymous amino acid substitution that is predicted by *in silico* analyses. Further study warrants experimental validation to prove the functionality of amino acid substitutions by nsSNPs of genes found in Berkshire pigs, which will allow us to identify causal variations to the traits of interest and ultimately to use variants as genetic markers.

Taken together, our study allowed an investigation of novel biological facts that could promote a better understanding of Berkshire domestication, and some of our results overlapped with and corroborated those of previous studies. We expected that many of our results will provide deep insights into the genetic factors underlying Berkshire domestication.

CONFLICT OF INTEREST

We certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript.

ACKNOWLEDGMENTS

This work was supported by a grant from the Next-Generation BioGreen 21 Program (No. PJ01110901, PJ01315101), Rural Development Administration, Republic of Korea.

REFERENCES

1. Jeong H, Song KD, Seo M, et al. Exploring evidence of positive selection reveals genetic basis of meat quality traits in Berkshire pigs through whole genome sequencing. *BMC Genet* 2015; 16:104.
2. Hazel LN. The Genetic basis for constructing selection indexes. *Genetics* 1943;28:476-90.
3. Li M, Tian S, Yeung CK, et al. Whole-genome sequencing of Berkshire (European native pig) provides insights into its origin and domestication. *Sci Rep* 2014;4:4678.

4. Wood JD, Nute GR, Richardson RI, et al. Effects of breed, diet and muscle on fat deposition and eating quality in pigs. *Meat Sci* 2004;67:651-67.
5. Ryu YC, Choi YM, Lee SH, et al. Comparing the histochemical characteristics and meat quality traits of different pig breeds. *Meat Sci* 2008;80:363-9.
6. Lee SH, Choi YM, Choe JH, et al. Association between polymorphisms of the heart fatty acid binding protein gene and intramuscular fat content, fatty acid composition, and meat quality in Berkshire breed. *Meat Sci* 2010;86:794-800.
7. Jeong DW, Choi YM, Lee SH, et al. Correlations of trained panel sensory values of cooked pork with fatty acid composition, muscle fiber type, and pork quality characteristics in Berkshire pigs. *Meat Science*. 2010;86:607-15.
8. Oh JD, Kim ES, Lee HK, Song KD. Effect of a *c-MYC* gene polymorphism (g.3350G>C) on meat quality traits in Berkshire. *Asian-Australas J Anim Sci* 2015;28:1545-50.
9. Lee SH, Choi YM, Choe JH, et al. Association between polymorphisms of the heart fatty acid binding protein gene and intramuscular fat content, fatty acid composition, and meat quality in Berkshire breed. *Meat Sci* 2010;86:794-800.
10. Suzuki K, Shibata T, Kadowaki H, Abe H, Toyoshima T. Meat quality comparison of Berkshire, Duroc and crossbred pigs sired by Berkshire and Duroc. *Meat Sci* 2003;64:35-42.
11. Kang YK, Choi YM, Lee SH, et al. Effects of myosin heavy chain isoforms on meat quality, fatty acid composition, and sensory evaluation in Berkshire pigs. *Meat Sci* 2011;89:384-9.
12. Cho IC, Park HB, Yoo CK, et al. QTL analysis of white blood cell, platelet and red blood cell-related traits in an F2 intercross between Landrace and Korean native pigs. *Anim Genet* 2011; 42:621-6.
13. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 2011;12:443-51.
14. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 2002;30:3894-900.
15. Radivojac P, Vacic V, Haynes C, et al. Identification, analysis, and prediction of protein ubiquitination sites. *Proteins* 2010; 78:365-80.
16. Doniger SW, Kim HS, Swain D, Corcuera D, Williams M, Yang S-P, Fay JC. A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genet* 2008;4:e1000183.
17. Daley D, Park JE, He J-Q, et al. Associations and interactions of genetic polymorphisms in innate immunity genes with early viral infections and susceptibility to asthma and asthma-related phenotypes. *J Allergy Clin Immunol* 2012;130:1284-93.
18. Azad AK, Sadee W, Schlesinger LS. Innate immune gene polymorphisms in tuberculosis. *Infect Immun* 2012;80:3343-59.
19. Netea MG, Wijmenga C, O'Neill LA. Genetic variation in Toll-like receptors and disease susceptibility. *Nat Immunol* 2012; 13:535-42.
20. Noreen M, Arshad M. Association of TLR1, TLR2, TLR4, TLR6, and TIRAP polymorphisms with disease susceptibility. *Immunol Res* 2015;62:234-52.
21. Ghosh M, Sodhi SS, Sharma N, et al. An integrated *in silico* approach for functional and structural impact of non-synonymous SNPs in the MYH1 gene in Jeju native pigs. *BMC Genet* 2016;17:35.
22. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114-20.
23. Andrews S. Fastqc: a quality control tool for high throughput sequence data [Internet]. Cambridge, UK: Babraham Institute; 2010 [cited 2017 Mar]. Available from: <http://www.bioinformaticsbabraham.ac.uk/projects/fastqc>
24. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357-9.
25. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078-9.
26. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297-303.
27. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics* 2011;27:2156-8.
28. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;155: 945-59.
29. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 2003;164:1567-87.
30. Bindea G, Mlecnik B, Hackl H, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 2009;25:1091-3.
31. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;4:1073-81.
32. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;31:3812-4.
33. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* 2013;Chapter 7:Unit7.20.
34. Cingolani P, Platts A, Wang le 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* 2012;6:80-92.
35. Luo W, Cheng D, Chen S, et al. Genome-wide association analysis of meat quality traits in a porcine Large White× Minzhu intercross population. *Int J Biol Sci* 2012;8:580-95.
36. Cánovas A, Quintanilla R, Amills M, Pena RN. Muscle transcriptomic profiles in pigs with divergent phenotypes for fatness traits. *BMC Genomics* 2010;11:372.
37. Rutkowski JM, Stern JH, Scherer PE. The cell biology of fat

- expansion. *J Cell Biol* 2015;208:501-12.
38. Wood J, Richardson R, Nute G, et al. Effects of fatty acids on meat quality: a review. *Meat Sci* 2004;66:21-32.
39. Kinoshita S, Ogawa W, Okamoto Y, et al. Role of hepatic STAT3 in the regulation of lipid metabolism. *Kobe J Med Sci* 2008;54: 200-8.
40. Grant R, Boler V, Ridge T, Graves T, Swanson K. Skeletal muscle tissue transcriptome differences in lean and obese female beagle dogs. *Anim Genet* 2013;44:560-8.