Article
Discoveries

Genomic analysis revealed a convergent evolution of LINE-1 in coat color: A case study in water buffaloes (*Bubalus bubalis*)

Dong Liang[1*], Pengju Zhao[1*], Jingfang Si[1], Lingzhao Fang[2], Erola Pairo-Castineira[2], Xiaoxiang Hu[3], Qing Xu[4], Yali Hou[5], Yu Gong[6], Zhengwen Liang[7], Bing Tian[8], Huaming Mao[9], Marnoch Yindee[10], Md Omar Faruque[11], Siton Kongvongxay[12], Souksamlane Khamphoumee[12], George E. Liu[13], Dongdong Wu[14], J. Stuart F. Barker[15], Jianlin Han[16,17#], Yi Zhang[1#]

[1]National Engineering Laboratory for Animal Breeding, Key Laboratory of Animal Genetics and Breeding and Reproduction of MOAR, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China.

[2]Medical Research Council Human Genetics Unit at the Medical Research Council Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, United Kingdom.

[3]State Key Laboratory of AgroBiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China.

[4]College of Life Sciences and Bioengineering, Beijing Jiaotong University, Beijing 100044, China.

[5]Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China.

[6]Guizhou Domestic Animal Genetic Resources Management Station, Guiyang 550001, China.

[7]Agriculture and Rural Affairs Bureau of Fenggang County, Zunyi 564200, China.

[8]Animal Disease Prevention and Control Station of Zunyi City, Zunyi 564200, China.

[9]College of Animal Science and Technology, Yunnan Agricultural University, Kunming 650201, China.

[10]Akkhararatchakumari Veterinary College (AVC), Walailak University, Nakorn Si Thammarat 80161, Thailand.

[11]Department of Animal Breeding and Genetics, Bangladesh Agricultural University, Mymensingh-2202, Bangladesh.

[12]Livestock Research Center, National Agriculture and Forestry Research Institute, Ministry of Agriculture and Forestry, Vientiane, Lao PDR.

[13]Animal Genomics and Improvement Laboratory, Henry A. Wallace Beltsville Agricultural Research Center, Agricultural Research Service, USDA, Beltsville, Maryland 20705, USA.

[14]Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China.

[15]School of Environmental and Rural Science, University of New England, Armidale, NSW 2351, Australia.

[16]International Livestock Research Institute (ILRI). P.O. Box 30709, Nairobi 00100, Kenya.

[17]CAAS-ILRI Joint Laboratory on Livestock and Forage Genetic Resources, Institute of Animal Science, Chinese Academy of Agricultural Sciences (CAAS), Beijing 100193, China.

[*]contributed equally

[#]Corresponding authors
Yi Zhang: yizhang@cau.edu.cn;
Jianlin Han: h.jianlin@cgiar.org

Abstract

Visible pigmentation phenotypes can be used to explore the regulation of gene expression and the evolution of coat color patterns in animals. Here, we performed whole-genome and RNA sequencing and applied GWAS, comparative population genomics and biological experiments to show that the 2,809 bp long LINE-1 insertion in the *ASIP* (agouti signaling protein) gene is the causative mutation for the white coat phenotype in swamp buffalo (*Bubalus bubalis*). This LINE-1 insertion (3' truncated and containing only 5' UTR) functions as a strong proximal promoter that leads to a 10-fold increase in the transcription of *ASIP* in white buffalo skin. The 165 bp of 5' UTR transcribed from the LINE-1 is spliced into the first coding exon of *ASIP*, resulting in a chimeric transcript. The increased expression of *ASIP* prevents melanocyte maturation, leading to the absence of pigment in white buffalo skin and hairs. Phylogenetic analyses indicate that the white buffalo-specific *ASIP* allele originated from a recent genetic transposition event in swamp buffalo. Interestingly, as a similar LINE-1 insertion has been identified in the cattle *ASIP* gene, we discuss the convergent mechanism of coat color evolution in the Bovini tribe.

Key words: white coat color, water buffalo, *ASIP* gene, LINE-1, transposon, convergent evolution

Introduction

Animal pigmentation is one of the most visible and variable traits shaped by natural and/or artificial selection. As a visible phenotypic marker, pigmentation has played an important role in our understanding of inheritance, development and evolutionary theory (Hoekstra 2006; Mort et al. 2015; Cuthill et al. 2017). In mammals, basic coat coloration is determined by the ratio of two pigments – eumelanin and pheomelanin. There are almost 200 color genes identified in mice (Mort et al. 2015). These genes act during developmental and cellular processes, including melanocyte development, melanogenesis, pigment transport and transfer (Cieslak et al. 2011). Coat color, an important form of camouflage in the wild ancestors of domestic animals, is likely under strong purifying selection. The relaxation of natural selective constraints and human mediated positive selection for different coat color phenotypes in domestic animals are believed to be the primary driving mechanisms leading to their significantly enriched allelic variation in coat-color-associated genes (Norris and Whan 2008; Fang et al. 2009; Cieslak et al. 2011; Henkel et al. 2019; Bruders et al. 2020). Domestic animal genetic resources provide an excellent opportunity to study the causative mutations and regulatory mechanisms responsible for coat color diversity (Henkel et al. 2019; Bruders et al. 2020).

The domestic Asian water buffalo (*Bubalus bubalis*) is an important animal resource, with a current global population of ca. 202 million supplying draught power, milk and/or meat in at least 67 countries on five continents (http://www.fao.org/faostat/). Two types have been recognized – river and swamp (Zhang et al. 2020). River buffaloes are native to the Indian sub-continent and have spread west as far as to the Balkans, Greece, Egypt and Italy within recorded historical times, while swamp buffaloes are found throughout south-east Asia, from Assam in India and Bangladesh in the west to the Yangtze valley of China in the east. Swamp buffalo is usually dark gray to black with white chevrons (one or two white stripes on the throat) and socks, and relatively straight, long and pale-colored horns. White swamp buffaloes (a common variant) have white hairs over the entire body, overlying pink skin, but their eyes are dark, the same as those of black buffaloes (supplementary fig. S1, Supplementary Material online). In some Asian countries, such as China and Indonesia, white buffaloes are particularly valued because the white coat phenotype is preferred for cultural (ceremonial slaughter at funerals) or religious reasons. Although a dominant gene was shown to determine the white coat phenotype nearly 60 years ago (Rife & Buranamanas 1959; Rife 1962), its molecular basis remains unknown.

In this study, we performed whole-genome and RNA sequencing using both next generation sequencing (NGS) and long-read sequencing strategies, and applied genome-wide association study (GWAS), population genomics and biological experiments to explore the genetic mechanism underlying the white coat phenotype of swamp buffaloes. We demonstrated that a transposable element (TE) insertion, functioning as the active promoter of the *ASIP* (agouti signaling protein) gene, is the causal mutation for the white coat phenotype.

Results

Mapping of the white coat phenotype in swamp buffaloes

Whole-genome sequencing (WGS) was conducted for 22 white and 41 black swamp buffaloes that were randomly sampled from five populations (supplementary table S1_A, Supplementary Material online). In total, 2,003 Gb of sequence data were generated, resulting in averages over the 63 animals for depth (11.95X) and coverage (98.27%) of the river buffalo reference genome (UOA_WB_1, GCF_003121395.1, Low et al. 2019) (supplementary table S2, Supplementary Material online). After quality control, 10,999,832 SNPs remained for GWAS analysis of the white coat phenotype. We applied Fisher's exact test with the dominant gene effect model in the PLINK software v1.07 (Purcell et al. 2007) and identified the most significant peak with the highest level of significance located on buffalo chromosome (BBU) 14 associated with the white coat phenotype (fig. 1A). This region contained 407 genome-wide statistically significant SNPs (Fisher's exact test, Bonferroni-corrected P-value < 0.05, $-\log_{10} P = 8.34$; fig. 1B; supplementary table S3, Supplementary Material online), spanning 1.07 Mb

(BBU14:19,332,562-20,392,733). It harbors five pseudogenes and 23 genes (16 protein-coding genes and seven RNA genes), including the well-known color gene *ASIP* (fig. 1E; supplementary table S4, Supplementary Material online).

Meanwhile, population genomic analyses were done to detect any signatures of selection underlying the white coat phenotype. A genomic scan for population genetic differentiation measured by F_{ST} was conducted using a sliding window of 50 Kb length with an increment of 25 Kb. The results showed that the top 20 windows of F_{ST} values were mainly located in a region on BBU14 (BBU14:19,575,001-20,000,000), overlapping with the GWAS signals (figs. 1A, 1B; supplementary table S5, Supplementary Material online). This finding suggested that this region carried a signature of positive selection in white buffaloes. However, the measure of polymorphism (nucleotide diversity (Pi)), showed similar patterns in the GWAS signal region in white and black buffaloes (data not shown), which could largely be explained by the dominant inheritance of white coat color, so that the majority of the white buffaloes were heterozygotes.

Validation of the GWAS signals using independent samples

The above analyses provided the first line of evidence that the peak on BBU14 likely represented a candidate locus responsible for white coat color in swamp buffaloes. To validate the GWAS signals, we performed an association analysis using a panel of 20 significant variants, but based on a larger collection of samples, including 80 white and 122 black buffaloes that were sampled from Thailand, Bangladesh and China (supplementary tables S1_A and S6, Supplementary Material online). These variants were generally distributed evenly in the target genomic region with relatively more variants falling in the location of *ASIP*, the strong candidate gene. As expected, the larger sample size resulted in more significant associations with $-\log_{10} P$-values of 18.7–50.2 (Fisher's exact test) (supplementary tables S7 and S8, Supplementary Material online). In particular, four adjacent SNPs located in the first intron and upstream of *ASIP* (BBU14:19,970,628, BBU14:20,048,647, BBU14:20,098,786 and BBU14:20,111,204) showed the strongest signals (supplementary table S8, Supplementary Material online). In addition, linkage disequilibrium (LD) analysis indicated that 13 SNPs in this region were tightly linked, forming an LD block that harbored *ASIP* (fig. 1C). As shown in a haplotype bifurcation diagram (see Methods), the white buffalo-specific core haplotype in this region showed long-range homozygosity, which was distinct from that of black buffaloes (fig. 1D). Taken together, these evidences suggested that the genomic region harboring *ASIP* was strongly associated with the white coat phenotype.

Given its critical role in mammalian pigmentation (Cieslak et al. 2011), we considered *ASIP* as a candidate gene and explored for any functional mutations within this gene that may be responsible for white coat phenotype. A total of 51 SNPs in *ASIP* were significantly associated with white coat phenotype (supplementary tables S3, S9, Supplementary Material online) in the

GWAS analysis. However, none was predicted to have any functional effects on the agouti protein (supplementary table S9, Supplementary Material online). Furthermore, as a validation, we used Sanger sequencing to detect the variants of *ASIP* in three DNA pools (one for white buffaloes and two for black buffaloes, see Methods). We identified 11 SNPs, of which nine were shared with the WGS results while the other two showed no association with the white coat phenotype because they occurred in only one of the two black buffalo DNA pools (supplementary table S10, Supplementary Material online). The translation of *ASIP* was validated to ensure that there was no missense mutation that altered the agouti protein structure and led to the white coat.

We then sought to explore if large structural variants were associated with coat color in the region of the GWAS signals. Three software tools, mrFAST v2.6.1.0 (Alkan et al. 2009), CNVnator v0.3.3 (Abyzov et al. 2011) and BreakDancer v1.1.2 (Chen et al. 2009), were used (see Methods), but based on our short-read WGS data, no structural variant private to white buffaloes was detected (supplementary tables S11-14, Supplementary Material online).

Upregulated expression of *ASIP* in white buffalo skin

To track down the potential causative gene(s), we compared transcription profiles of the 23 genes annotated in the significant GWAS region. Skin biopsies of six animals (three each of white and black swamp buffaloes) were used for whole transcriptome sequencing (RNA-seq) (supplementary table S15, Supplementary Material online). Based on the RNA-seq data, *ASIP* showed a 10.3-fold increase in transcription in the skin of white buffaloes (transcripts per million, TPM: 28.04 ± 6.34) as compared with black buffaloes (TPM: 2.96 ± 0.73) (fig. 2A; supplementary table S16, Supplementary Material online). This striking difference was further verified by real-time quantitative PCR (relative expression: 1.25 ± 0.75 for black buffaloes vs 12.86 ± 4.88 for white buffaloes; student t-test $P < 0.001$; fig. 2A; supplementary tables S17-18, Supplementary Material online). Further, we characterized the tissue-specific expression profile using our newly generated RNA-seq data and published data from 55 tissue and cell types of river buffaloes (supplementary table S19, Supplementary Material online). Interestingly, among the genes in the significant GWAS region, only *ASIP* showed tissue-specific expression in skin tissue (supplementary fig. S2, Supplementary Material online). These findings showed that the white coat phenotype in swamp buffaloes might be the result of a *cis*-regulatory variant that elevated *ASIP* expression in the skin.

Identification of the white buffalo-specific *ASIP* transcript

To address whether *ASIP* transcripts were different between white and black buffaloes, we initially visualized RNA-seq reads in IGV (Integrative Genomics Viewer, Robinson et al. 2017). This revealed distinct patterns of overlapping reads that were mapped to *ASIP* exons

(supplementary fig. S3, Supplementary Material online). In black buffaloes, as expected, reads were aligned to both non-coding exons and coding exons with similar read counts (depth) across all three coding exons. In white buffaloes, however, reads were mainly aligned to the coding exons. In particular, read counts on exon 2 (the first coding exon) decreased gradually from the 5' end to the 3' end, implicating a distinct transcript. Transcript assembly and quantification based on the RNA-seq data using the Stringtie software v2.0 (Pertea et al. 2015) showed two abundantly expressed transcripts, one in black buffaloes and another in white buffaloes (fig. 2B).

To isolate full-length transcripts and characterize their transcription initiation sites, 5' and 3' Rapid Amplification of cDNA Ends (RACE) PCR experiments for skin samples of one white buffalo and one black buffalo were done (supplementary table S17, Supplementary Material online). The RACE-PCR products were subject to conventional cloning, followed by Sanger sequencing. Sequences were determined for multiple clones of white (16 clones in 3' end and 17 in 5' end) and black (14 in 3' end and 14 in 5' end) buffaloes. Comparison of clones from black buffalo showed six alternative transcripts that shared the same coding exons and 3' UTR but differed in 5' UTR (fig. 2C; supplementary fig. S4, Supplementary Material online). White buffalo, however, had only one transcript. Interestingly, while sharing the same coding exons and 3' UTR with black buffalo, the white buffalo-specific transcript contained an unknown 165 bp sequence at 5' UTR that could not be aligned to the buffalo reference genome (fig. 2C; supplementary fig. S4, Supplementary Material online).

To further characterize this white buffalo-specific transcript, a BLASTN search assigned this unknown 165 bp fragment to a bovine LINE-1 transposon element (L1-BT, GenBank accession no. DQ000238) with a sequence identity of 98%. This suggested that the presence of a LINE-1 insertion upstream of *ASIP* led to a chimeric transcript in white buffalo.

Genomic position of the white buffalo-specific LINE-1 insertion in *ASIP*

To position the LINE-1 insertion, we analyzed the soft-clipped reads that mapped upstream of *ASIP* (BBU14:19,952,567-20,083,962) in our WGS data. In contrast to aligned reads, soft-clipped reads were partially mapped to the buffalo reference genome and contained unmapped sequences, suggestive of structural variants. To improve the efficiency of comparative analysis, we pooled data of 10 randomly selected samples from each of the two coat color phenotypes, and compared the counts of soft-clipped reads on each genomic position between white and black buffaloes (supplementary file 1, Supplementary Material online). The position at BBU14:19,996,806 showed the top signal (the difference of 38 in the counts of soft-clipped reads) that was further verified in the IGV (supplementary fig. S5, Supplementary Material online). The soft-clipped reads mapped to this position were divided into two categories: left soft-clipped reads (truncated at BBU14:19,996,806) and right soft-clipped reads (truncated at BBU14:19,996,791). The 16 bp fragment (TGCTACTTTCTTTTTG) between these two reads showed much higher

read depth than its flanking regions in white buffaloes (supplementary fig. S5, Supplementary Material online), indicating the presence of an insertion variant. Then, we *de novo* assembled all the soft-clipped reads containing the 16 bp fragment, yielding two contigs: one of 269 bp on the left connecting to the upstream flanking sequence at position BBU14:19,996,791 and another of 257 bp on the right joining the downstream flanking sequence at position BBU14:19,996,806 (fig. 2D; supplementary fig. S5, Supplementary Material online). The 165 bp 5' UTR of the white buffalo-specific transcript was perfectly aligned to the contig on the right. Thus, we positioned the white buffalo *ASIP* LINE-1 insertion at BBU14:19,996,791-19,996,806 and obtained its head- and tail-end DNA sequences. This LINE-1 insertion was 44.2 Kb away from the first coding exon (exon 2, located at BBU14:19,952,408-19,952,577) of *ASIP*.

Sequence of the complete LINE-1 insertion in white buffalo *ASIP*

To determine the complete sequence of the LINE-1 insertion, we sequenced one white buffalo and one black buffalo using Nanopore long-read sequencing technology. We generated 80.64 Gb of filtered data, including 7,833,594 and 9,759,939 reads with mean lengths of 5,103 bp and 4,166 bp in white buffalo and black buffalo, respectively. The reads aligned to the genomic region of *ASIP* (BBU14:19946809-20113374) were extracted for *de novo* assembling. Using the Canu assembler v1.8 (Koren et al. 2017), 33 contigs were assembled for the white buffalo with a mean length of 12,024 bp, of which the longest contig was 67,524 bp (supplementary file 2, Supplementary Material online). By aligning the two partial LINE-1 fragments (269 bp and 257 bp) assembled from short-reads to this longest contig, we resolved the complete structure of the LINE-1 insertion. It was 2,809 bp in length flanked by the 16 bp direct repeat (TGCTACTTTCTTTTTG) that was characterized as the target site duplication (TSD) of the LINE-1 element (fig. 2C). In black buffalo, however, there was only one copy of this 16 bp sequence and the LINE-1 fragment was not detected.

This LINE-1 was 3' truncated and contained only 5' UTR of a full-length LINE element. It was located upstream of the first coding exon and in the same orientation as that of the *ASIP* transcription. Therefore, the promoter of LINE-1 could act as a strong alternative promoter to drive *ASIP* expression in white buffaloes. The 165 bp of 5' UTR transcribed from the LINE-1 was spliced into the first coding exon, creating the chimeric *ASIP* transcript in white buffaloes (fig. 2C).

To validate the association of the presence of this LINE-1 element with the white coat phenotype, we developed a genotyping assay and examined 91 white and 194 black buffaloes (fig. 2E; supplementary table S1_B, Supplementary Material online). The result showed that the LINE-1 was perfectly associated with the white coat phenotype. All black buffaloes were wild-type homozygotes. White buffaloes were either heterozygous or homozygous for the LINE-1 insertion, confirming that the white coat phenotype is inherited as a dominant Mendelian trait.

Effect of increased *ASIP* expression on melanocyte development

To explore the regulatory mechanism of the white coat phenotype, we investigated the gene expression profiles based on skin RNA-seq data generated in the current study. Transcriptome analysis revealed a total of 344 differentially expressed genes (DEGs), of which 148 DEGs were down-regulated while 196 DEGs were up-regulated in white buffalo skin (FDR < 0.01 and fold change ≥ 2 as the thresholds). A functional annotation showed that the down-regulated DEGS were enriched in melanocyte biology-related Gene Ontology (GO) terms (e.g., melanin metabolic process, melanin biosynthetic process) and KEGG pathways (e.g., tyrosine metabolism, melanogenesis), indicating that the melanocyte function might be diminished in white buffalo skin (supplementary table S20, Supplementary Material online). We found that five out of the 11 skin-color-associated genes (*TYR*, *DCT/TYRP2*, *TYRP1*, *PMEL* and *OCA2*) showed significantly lower or no expression ($P < 0.01$) in white buffaloes, five (*KITLG*, *MITF*, *MC1R*, *EDNRB* and *SOX10*) displayed no difference ($P > 0.05$) between white and black buffaloes while only *KIT* had a slightly higher expression ($P < 0.05$) in white buffalo skin (fig. 3A; supplementary table S16, Supplementary Material online). We then focused on the tyrosinase-related family genes *TYRP2* and *TYRP1* that are consecutively expressed in melanocytes during their migration in the dermis and maturation, as markers of early and late differentiation, respectively (Steel et al. 1992; Botchkareva et al. 2003; Manceau et al. 2011). While *TYRP2* was expressed in both white and black buffaloes, *TYRP1* was expressed only in black buffaloes (fig. 3A; supplementary table S16, Supplementary Material online), indicating that the melanocyte was fully differentiated in black buffaloes but not in white buffaloes. This was further supported by the immunohistochemical staining of skin samples. The fully differentiated (Trp1+) melanocytes were observed at the dermal–epidermal junction in black buffaloes, while no Trp1+ signal was present in white buffaloes (fig. 3B). Melanin pigment was present near the melanocytes in black buffaloes but not in white buffaloes (fig. 3C). Collectively, these results indicated that the over-expression of *ASIP* prevented melanocyte maturation, leading to the absence of pigment in white buffalo skin and hairs (fig. 4).

Interestingly, the up-regulated DEGs in white buffalo skin were also overrepresented in the growth-related GO terms, such as the development of skeletal system, tissue, organ and connective tissue (supplementary table S20, Supplementary Material online), implicating possible pleiotropic effect of *ASIP* over-expression on physiology and metabolism in white buffaloes.

Origin of the LINE-1 insertion in white buffalo *ASIP*

Mammalian genomes host hundreds of thousands of LINE-1 elements that have accumulated since the origin of mammals (Boissinot & Sookdeo 2016). Although the great majority of LINE-

1s are inactive, some retain the ability to retrotranspose (Sassaman et al. 1997; Richardson et al. 2015). To investigate if the LINE-1 insertion in white buffalo *ASIP* occurred recently in the buffalo species or was shared with the other related bovine species, we explored its evolutionary origin based on the sequence similarity at two levels – among and within species. First, full-length LINE-1 elements were identified from the reference genomes of water buffalo and two related species (taurine cattle and yak), using the RepeatMasker software v4.07 (http://www.repeatmasker.org) and compared with the bovine L1-BT transposon element (GenBank accession no. DQ000238) as the reference library. A total of 6,986 full-length LINE-1 copies, including 2,516 from river buffalo, 1,571 from swamp buffalo, 1,617 from taurine cattle and 1,282 from yak genomes (supplementary tables S21-24, Supplementary Material online), were obtained and used to construct an approximately maximum-likelihood tree. A primary analysis showed that the LINE-1 copies were closely related between swamp and river buffaloes (supplementary fig. S6, Supplementary Material online). To improve the visualization of the phylogeny of LINE-1s from different species, we used the same color for swamp and river buffaloes in the tree (fig. 5A). This tree held two major clades: (1) One on the left consisted of water buffalo LINE-1 copies mixed with those of taurine cattle and yak. They all had comparably long branch lengths, representing ancient and inactive LINE-1 copies; and (2) On the right another clade was divided into several sub-clades of those with mixed LINE-1 copies of all three species, but having relatively short branch lengths as well as three species-specific subclades. These species-specific LINE-1 copies displayed the shortest branch lengths and tended to cluster tightly to each other within species, suggesting their recent evolutionary origins. The LINE-1 copy in white buffalo *ASIP* clustered with the water buffalo-specific subclade, indicating it to be a young copy derived from the water buffalo-specific LINE-1 copies.

Next, we characterized water buffalo-specific LINE-1 copies and the evolutionary origin of the *ASIP* LINE-1. Using the 2,809 bp long *ASIP* LINE-1 as a probe, we identified 1,500 and 1,267 LINE-1 elements in the river and swamp buffalo reference genomes, respectively, of which 1,009 and 766 were retained after filtering based on their sequence identities (> 80%; supplementary tables S25-26, Supplementary Material online). A minimum spanning (MS) tree analysis categorized these LINE-1 elements into 21 distinct subfamilies, each containing from 51 to 139 copies (*P*-values for subfamily partition ranging from 8E-215 to 7E-124) (fig. 5B; supplementary table S27, Supplementary Material online). The LINE-1 copy in white buffalo *ASIP* belonged to a young subfamily (sub20 in fig. 5B).

Finally, we did a population genetics analysis to characterize the relationship of haplotypes in the genomic region of 10 Kb flanking the insertion point of LINE-1 upstream of *ASIP* (BBU14:19,991,854-20,001,504) in both white and black buffaloes. We identified 42 haplotypes in 73 buffaloes (63 swamp buffaloes from our WGS data and 10 river buffaloes from published data (Whitacre et al. 2017). A median-joining network defined three haplogroups – one for river buffaloes and two for swamp buffaloes, namely SW1 and SW2. All white buffaloes were in the

SW2 group (fig. 5C). This finding was also supported by the maximum likelihood evolutionary tree and sequence alignment (supplementary fig. S7, Supplementary Material online). These results indicated that the haplotype carrying the LINE-1 insertion was closely related to a haplogroup of the non-white swamp buffaloes, in line with the scenario that white buffaloes originated from a recent genetic transposition event within the swamp buffalo rather than due to introgression from the river buffalo or another species.

Discussion

In this study, we combined evidence from GWAS, RNA-seq, long-read sequencing and histological data to demonstrate that a LINE-1 insertion functioning as the active promoter of *ASIP* is the causal mutation for the white coat phenotype in swamp buffaloes. To our knowledge, this is the first morphological trait in water buffalo (Online Mendelian Inheritance in Animals (OMIA) 000213-89462 at https://omia.org/OMIA000213/89462/) to have its molecular mechanism uncovered.

White coat color, a common phenotypic variant in mammals, may be due to albinism or leucism. The former results from a disruption of pigment synthesis, usually caused by mutations of the tyrosinase gene (*TYR*), whereas the latter is caused by the absence of mature melanocytes in skin (Cieslak et al. 2011). Mutations in genes involved in melanocyte development, such as *KIT* (e.g., Haase et al. 2007) and *MITF* (e.g., Karlsson et al. 2007), can lead to leucism. *ASIP* encodes the agouti signaling protein, which has been well characterized as having an important role in melanin synthesis. It acts as an antagonist to the alpha-MSH (melanocyte-stimulating hormone) for the melanocortin-1 receptor (*MC1R*), leading to an increased pheomelanin synthesis in melanocytes (Furumura et al. 1998; Schiaffino 2010). The elevated expression of *ASIP* increases pheomelanin production whereas its decreased expression or loss of function mutations tends to result in the exclusive production of eumelanin and thus dark pigmentation (non-agouti phenotype) in rodents (Kingsley et al. 2009; Hubbard et al. 2010; Tanave et al. 2019). Recent studies also demonstrate that *ASIP* is involved in melanocyte development. It not only inhibits forward differentiation of melanoblasts (unpigmented melanocyte precursors) (Sviderskaya et al. 2001), but also induces rapid dedifferentiation of cultured melanocytes to the morphology of melanoblasts (Hida et al. 2009; Le Pape et al. 2009). *ASIP* is involved in the formation of a stripe pattern and dorso-ventral patterning in mammals (Girardot et al. 2006; Manceau et al. 2011; Mallarino et al. 2016), birds (Haupaix et al. 2018; Inaba et al. 2019; Robic et al. 2019) and teleost fishes (Ceinos et al. 2015; Kratochwil et al. 2018; Cal et al. 2019). The *cis*-regulatory variation in *ASIP* has also been shown to facilitate the adaptive winter camouflage polymorphism in snowshoe hares (Jones et al. 2018). In this study, we illustrate that a regulatory mutation leading to the 10-

fold increase of *ASIP* expression prevents melanocyte differentiation and thus results in the white coat phenotype in swamp buffaloes.

TEs are a key source of genomic structural variations in both eukaryotic and prokaryotic genomes. Recent evidence indicates that, in humans and model organisms, TEs play important roles in gene regulation, by contributing promoters and transcription factor binding sites and by affecting chromatin structures to change the expression of nearby genes (Merenciano et al. 2016; Burns 2017; Jang et al. 2019; De Cecco et al. 2019; Diehl et al. 2020). For example, viable yellow agouti (A^{vy}) mice carry an intracisternal A-particle (IAP) retrotransposon inserted into the *ASIP* locus and the cryptic promoter within the IAP 5' long terminal repeat (LTR) acts to drive the ectopic expression of *ASIP*, resulting in altered coat color, obesity and an increased incidence of tumors (Duhl et al. 1994; Michaud et al. 1994; Klebig et al. 1995). However, the regulatory function and evolution of TEs have not been well characterized in agricultural animals (Girardot et al. 2006; Dreger & Schmutz 2011).

LINE-1 is the most abundant type of retrotransposon in mammalian genomes (Richardson et al. 2015), and mounting evidence indicates that, in humans, the insertion of a LINE-1 element can affect the expression of neighboring genes, causing phenotypic variation and diseases (Burns 2017; De Cecco et al. 2019; Jang et al. 2019). One potential mechanism by which LINE-1 affects gene expression is by introducing regulatory elements or promoters (Faulkner et al. 2009; Elbarbary et al. 2016). A full-length LINE-1 is typically 6-8 Kb and contains a promoter within its 5' UTR, two open reading frames (ORF1 and ORF2), a short 3' UTR and a poly(A) tail (Moran et al. 1996). The 5' UTR of LINE-1 has bidirectional promoter activity – a sense promoter that drives the transcription of the ORF-1 and ORF-2 proteins required for retrotransposition and an antisense promoter that affects the transcription of its upstream genomic region (Speek 2001; Nigumann et al. 2002; Beck et al. 2011). The transcribed 5' LINE-1 antisense sequences are usually spliced to the exons of neighboring genes to form chimeric transcripts (Speek 2001). Recent studies show that LINE-1 antisense promoter-driven transcriptions are common in humans (Faulkner et al. 2009; Criscione et al. 2016). In this study, we identify a 2,809 bp long LINE-1 insertion upstream of the first coding exon of *ASIP*, which acts as an active proximal promoter (~44 Kb away from the first coding exon) to initiate the transcription of *ASIP* in white buffaloes. In contrast, the wild type allele of *ASIP* initiates the transcription from a distal promoter (~72 Kb away from the first coding exon) in black buffaloes. The promoter activity of LINE-1 could be enhanced by the upstream flanking sequence (Lavie et al. 2004), inducing an increased expression of *ASIP*. However, a different mechanism is found in white sheep, where Norris and Whan (2008) identified a tandem duplication encompassing *ASIP* and two neighboring genes, *AHCY* and *ITCH*, which enhanced the expression of *ASIP* activated by a duplicated copy of the nearby *ITCH* promoter.

In nature, although convergence in phenotype is common, convergence at the molecular level is rather rare (Zou and Zhang 2015). However, in cattle as in white buffaloes, a LINE-1 element

(L1-BT) located between the non-coding and coding exons of *ASIP* is associated with the brindle coat color in Normande cattle (Girardot et al. 2006). Sequence comparisons indicate independent origins of the LINE-1 elements in white buffaloes and cattle. First, the two LINE-1 insertions are located in different genomic positions relative to *ASIP*, i.e., 44 Kb and 15 Kb from the first coding exon in white buffaloes and in cattle, respectively (fig. 6). Second, they belong to species-specific LINE-1 subclades (fig. 5A). Third, the two LINE-1 insertions have distinct TSDs to facilitate independent transposition events. Although the DNA structures of the LINE-1 elements are different in the two species, i.e., full-length LINE-1 (8.4 Kb) in cattle and 3' truncated LINE-1 (2.8 Kb) in white buffaloes, they share significant functional similarities (fig. 6). First, both LINE-1 elements have the same orientation as *ASIP* and transcribe a conserved sequence (~160 bp, 98% identity) from the LINE-1 that is spliced to the coding exons forming a chimeric transcript. Second, consistent with that in white buffaloes, the LINE-1 insertion also led to the over-expression of *ASIP* in cattle (Girardot et al. 2006; Albrecht et al. 2012). Therefore, the two independent LINE-1 insertions in *ASIP* lead to similar functional impacts, and our study presents a compelling case for a convergent mechanism affecting coat color evolution in the Bovini tribe (Martin & Orgogozo 2013; Cuthill et al. 2017).

In the human genome, more than 99% of LINE-1 copies are unable to move due to 5' truncation, rearrangement or mutation (Goodier and Kazazian 2008; Beck et al. 2011 Hancks and Kazazian 2016), with only a few remaining capable of retrotransposition (Brouha et al. 2003; Beck et al. 2011). Frequent 5' truncation is explained by an integration mechanism of LINE-1 retrotransposon – target primed reverse transcription (TPRT) (Luan et al. 1993). During TPRT, the LINE-1 endonuclease nicks genomic DNA, freeing a 3′ hydroxyl that serves as a primer for polymerizing the cDNA copy onto the host DNA. This process is frequently aborted, resulting in 5′ truncated LINE-1 copies. However, the LINE-1 copy of white buffalo *ASIP* is 3' truncated and contains only 5' UTR, which might be generated by an unconventional integration mechanism. This 3' truncation could be a special consequence of TPRT coupled with a reverse transcription/integration reaction to create an inversion in LINE-1 retrotransposition, a mechanism called 'twin priming' by Ostertag and Kazazian (2001). This result also highlights the important role that LINEs play in the evolution of many species.

Materials and Methods

Study samples

Three sets of swamp buffaloes were sampled from China, Thailand and Bangladesh (supplementary table S1_B, Supplementary Material online): 63 from China that were used for WGS and GWAS analysis, 202 from China, Thailand and Bangladesh that were used for an association study to validate the GWAS signals and 285 that were used to verify the candidate

causative mutation (LINE-1 insertion), which combined those used for WGS and the validation experiment, plus samples used only for genotyping the LINE-1 insertion.

Genomic DNA was extracted from blood or ear tissue using the phenol/chloroform method. The integrity and yield of genomic DNA were assessed and verified using agarose gel electrophoresis and a Nanodrop^{TM} spectrophotometer (Thermo Fisher Scientific, Waltham, USA), respectively.

WGS data generation and variant detection

Paired-end short insert (350 bp) libraries were constructed from genomic DNA and sequenced using the Illumina HiSeq X Ten system (Illumina, San Diego, CA, USA). Read pairs were aligned to the river buffalo reference genome (UOA_WB_1) using the BWA-MEM algorithm (http://bio-bwa.sourceforge.net/bwa.shtml) with the default parameters. PCR duplicates were removed with the MarkDuplicates module in the Picard Tools package v2.9.0 (http://broadinstitute.github.io/picard/). Realignment around indels was done using the GATK module IndelRealigner v3.8 (https://software.broadinstitute.org/gatk/). After variant calling by the GATK module UnifiedGenotyper, variant filtering was done using the parameters "QUAL < 30, QualByDepth (QD) < 2.0, RMS Mapping Quality (MQ) < 40.0, Mapping Quality Rank Sum Test (MQRankSum) < −12.5, Read Pos Rank Sum Test (ReadPosRankSum) < −8.0, Haplotype Score > 13.0".

GWAS and population genomics analyses

The VCFtools software v0.1.16 (Danecek et al. 2011) was used to convert the variant data file from VCF format to Plink format. For quality control filtering, we removed SNPs with call rates < 90% or with minor allele frequencies < 0.05 or departure from Hardy-Weinberg equilibrium < 10^{-6} and discarded individuals with > 10% missing genotypes. GWAS analysis was done using Fisher's exact test and the dominance gene effect model in the Plink software v1.07 (Purcell et al. 2007) with the parameters "−model −modeldom −Fisher". The GWAS results were visualized using the qqman R package v0.1.4 (https://CRAN.R-project.org/package=qqman).

Genetic differentiation (F_{ST}) between the populations was calculated using a sliding window approach (window size of 50 Kb with step size of 25 Kb) using the VCFtools. We reviewed the plots of average F_{ST} calculated using 50% overlapping windows of variable sizes (1 Kb, 10 Kb, 30 Kb, 50 Kb and 70 Kb) and found that the genome-wide pattern was reasonably smooth for the larger window sizes (30 Kb, 50 Kb and 70 Kb) but relatively noisy for small window sizes (1 Kb and 10 Kb). Therefore, we showed the results of 50 Kb window size from this analysis.

Variant genotyping using the KASP method

For validation of the GWAS signals, we genotyped a panel of 19 SNPs and one indel that showed significant associations with the white color phenotype in the target genomic region. Following the KASP assay guidelines (https://biosearchcdn.azureedge.net/assetsv6/KASP-genotyping-chemistry-User-guide.pdf), the wild-type and mutant-type allele-specific upstream PCR primers and a common downstream primer were designed for each variant (supplementary table 6, Supplementary Material online).

The microfluidic-based IMAP^{TM} platform (CapitalBio Technology, Beijing, China) was used for PCR amplification. The final reaction was done in a total volume of 1 μL, which contained 0.3 μL DNA template (50 ng/μL), 0.14 μL primer mix (12 μM each of the two allele-specific forward primers and 30 μM reverse primer), 0.5 μL 2× universal KASP Master Mix (LGC, UK) and 0.06 μL ddH_2O. PCR thermocycling was done as follows: Initiation at 95°C for 15 min; 10 cycles of denaturation at 95°C for 20 s and touchdown annealing from 61°C (−0.6°C/cycle) for 60 s; followed by 26 cycles of denaturation at 95°C for 20 s and annealing at 55°C for 60 s and finished by an extension at 37°C for 60 s. An end-point fluorescent read of the PCR products was done using the LuxScan^{TM}-10K/D instrument (CapitalBio Technology, Beijing, China).

LD analysis

LD (pairwise r^2 statistic) was calculated and visualized using the Haploview software v4.1 (Barrett et al. 2005). The R package rehh v3.0.1 (Gautier et al. 2017) was used to draw a haplotype bifurcation diagram (Sabeti et al. 2002) that visualizes the breakdown of LD at increasing distances from the focal core allele. The haplotypes used for drawing the bifurcation diagram were phased using the Beagle software v4.1 (Browning & Browning 2007).

Variant annotation in the 1.07 Mb genomic region of GWAS signals

The Annovar program v2018Apr16 (Wang et al. 2010) was applied to annotate each variant, using an annotation file of GTF format prepared for the river buffalo reference genome (UOA_WB_1).

Amplification and Sanger sequencing of *ASIP* exons and flanking regions

Primers (supplementary table 10, Supplementary Material online) were designed based on the river buffalo reference genome (UOA_WB_1) to amplify and sequence the coding exons and flanking regions (2,000 bp upstream of the first coding exon and 1,000 bp downstream of the last coding exon) of *ASIP*. Three DNA pools (Guizhou (white), Guizhou (black) and Yanjin (black)

buffalo breeds) were prepared as templates for PCR amplification. Each pool represented six individuals with genomic DNA equally mixed. PCR products were used for Sanger sequencing.

Structural variation (SV) detection in the 1.07 Mb genomic region

SVs were detected using three software tools – the mrFAST v2.6.1.0 (Alkan et al. 2009), the BreakDancer v1.1.2 (Chen et al. 2009) and the CNVnator v0.3.3 (Abyzov et al. 2011). For the mrFAST analysis, paired-end sequencing reads were first mapped to the river buffalo reference genome (UOA_WB_1) with the parameters "--search -- pe-e 5", followed by calculating read depth to detect segmental duplications and deletions. SVs were detected using the BreakDancer with the default parameters "-q 35-c 3-s 7-b 100-t-d" and using the CNVnator with bin size set to 500 bp (-tree -his 500 -stat 500 -partition500 -call 500). The statistic Vst was used to test the difference in copy numbers at each SV between white buffaloes and black buffaloes: Vst = (Vt-Vs) / Vt, where Vt is the overall variance of copy number and Vs the average variance within populations.

RNA extraction and qPCR

Total RNA was prepared from ear skin samples using the TRIzol reagent (Thermo Scientific, 15596026) in accordance with the manufacturer's recommendation. RNA purity, concentration and integrity were assessed using the LabChip ® GX Touch™ Nucleic Acid Analyzer (PerkinElmer, Waltham, MA, USA). Reverse transcription was done using the PrimeScript™ RT Reagent Kit (TAKARA Bio, Mountain View, CA, USA). Real-time quantitative PCR was done on the LightCycler® 480 Instrument II (Roche Diagnostics, Mannheim, Germany) using the SYBR Green I Master (Roche) kit. The designed primers are listed in supplementary table 17, Supplementary Material online. The 18S rRNA was used as the internal reference gene.

RNA-seq and data analysis

Sequencing libraries were constructed using the NEBNext® UltraTM RNA Library Prep Kit (NEB, Ipswich, MA, USA) following the manufacture's recommendations. The library preparations were sequenced using the Illumina HiSeq X Ten system. After quality control, the paired-end reads were mapped to the river buffalo reference genome (UOA_WB_1) using the HISAT2 software v2.6.1.0 (Kim et al. 2019). Transcripts were assembled and quantified with the StringTie software v2.1.1 (Pertea et al. 2015). The Cuffcompare tool of the Cufflink suite v2.2.1 (Trapnell et al. 2012) was used to compare the alternative transcripts among individuals. The results were visualized using the pheatmap R package v1.0.12 with Ward's hierarchical clustering method (https://CRAN.R-project.org/package=pheatmap).

The differential expression analysis was performed using the DESeq2 package v1.4.5 (Love et al. 2014). The significance of the difference in gene expression was determined using a Wald test in the DESeq2 package. The results with a false discovery rate (FDR) ≤ 0. 05 were considered noteworthy. A functional annotation of DEGs was conducted through Gene Ontology (GO) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis using the open access WebGestalt tool (http://www.webgestalt.org, Liao et al. 2019). The gene expression level was quantified using TPM with the StringTie software v2.1.1 (Pertea et al. 2015). It normalizes sequencing depth and gene length. To analyze gene expression profiles across different buffalo tissues for the 23 genes in the genomic region of the GWAS signals, RNA-seq data of six skin samples generated in this study were combined with published data of 248 samples from NCBI (PRJEB4351 (30 tissues of a male and a female Mediterranean buffalo calves from Italy) and PRJEB25226 (218 of the 220 tissue and cell samples, except ERX2403664 and ERX2403645 with no runs of data, of six Mediterranean water buffaloes from Italy and four river buffaloes of Pandharpuri and Bhadawari breeds from India), Williams et al. 2017; Li et al. 2019; Low et al. 2019; Young et al. 2019). The gene expression values (TPM) were log2 transformed and then visualized using the pheatmap R package v1.0.12 with Ward's hierarchical clustering method.

5' and 3' RACE

The 5' and 3' RACE PCR experiments were done using the SMARTer® RACE 5'/3' kit (Takara Bio, Mountain View, CA, USA) according to the manufacturer's instructions. The Primer Premier 5.0 software (http://www.premierbiosoft.com) was used to design specific primers (supplementary table 17, Supplementary Material online). The PCR product was cloned into the pClone007 vector using the pClone007 vector kit (Beijing TsingKe Biotech Co., Ltd., China) and multiple individual clones were sequenced.

Genotyping of the LINE-1 insertion

Two pairs of primers were designed based on the assembled white buffalo-specific *ASIP* sequence (supplementary table 17, Supplementary Material online). The PCR was set up in a final volume of 25 ul containing 18 µl dd H_2O, 5 pmol of each primer, 200 ng of genomic DNA, 2.5 µL 10×PCR buffer (TaKaRa, Dalian, China), 200 µM dNTP mixture (TaKaRa) and 1 U *Taq* polymerase (TaKaRa). PCR conditions were: 94 ℃ for 5 min followed by 35 cycles of 94 ℃ for 30 s, 60 ℃ for 1 min, and 72 ℃ for 1 min, and the final extension for 7 min at 72 °C.

Immunohistochemistry and HE staining

The skin samples were embedded in paraffin and then sectioned for Hematoxylin-Eosin (HE) staining and immunohistochemical staining. The rabbit polyclonal antibody anti-TRP1 (ab83774, Abcam, Cambridge, MA, USA) was used for immunohistochemical staining.

Nanopore long-read sequencing

DNA samples from one each of white and one black buffaloes were used for Nanopore long read sequencing in accordance with the standard protocol provided by Oxford Nanopore Technologies (ONT, Oxford, UK).

The FAST5 files containing signal data generated by the Nanopore sequencer were converted into the FASTQ format using the Albacore software in the MinKNOW package (ONT). Clean reads were obtained by removing the adaptor sequences, low-quality sequence reads and short reads (length < 500 bp). The Minimap2 software v2.17 (Li 2018) was used to map clean reads to BBU14 of the river buffalo reference genome (UOA_WB_1). Reads mapped to the target region were extracted using the Samtools software v1.10 (Li et al. 2009). Regional *de novo* assembly was done using the Canu software v1.8 (Koren et al. 2017) with the parameters "correctedErrorRate 0.144; CorOutCoverage 40".

Phylogenetic analyses of LINE-1 repeats

LINE-1 repeat elements were extracted from the reference genomes of river buffalo (UOA_WB_1), swamp buffalo (GWHAAJZ00000000, https://bigd.big.ac.cn/search/?dbId=gwh&q=GWHAAJZ00000000, Luo et al. 2020), taurine cattle (ARS-UCD1.2, GCA_002263795.2, Rosen et al. 2020) and yak (BosGru3.0, GCA_005887515.2), using the RepeatMasker software v4.07 (http://www.repeatmasker.org) with the slow search option, based on the Repbase repeat database v9.04 (http://www.girinst.org/). The ParseRM_GetNesting.pl script was used to filter out the nested LINE-1 elements from the RepeatMasker output. To extract the full-length LINE-1, the resulting non-nested LINE-1s were aligned to the full-length bovine LINE-1 L1-BT transposon element sequence (DQ000238), followed by a filtering to remove elements with length < 7,000 bp, truncation at 5' UTR < 200 bp, and truncation at 3' UTR < 300 bp. Finally, to ensure that the LINE-1 elements were highly homologous, a clustering-based approach was used to keep the LINE-1s with a sequence identity of > 80%, implemented in the CD-HIT software v4.6.8 (Fu et al. 2012) with the parameter "-T 0 -c 0.8 -M 0 -n 5 -p 0".

The Mafft software v7.407 (Katoh et al. 2019) was used for multiple sequence alignment of the qualified full-length LINE-1s with the parameter "mafft --quiet --thread 24 --retree 1". An approximately maximum-likelihood phylogeny was constructed based on the output (aligned.fa) from the Mafft alignment using the FastTree software v2.2.11 (Price et al. 2009) with the default

settings “Nucleotide distances: Jukes-Cantor Joins; balanced Support: SH-like 1000; Search: Normal + NNI + SPR (2 rounds range 10) + ML-NNI opt-each = 1; TopHits: 1.00*sqrtN close = default refresh = 0.80; ML Model: Jukes-Cantor, CAT approximation with 20 rate categories” and visualized using the iTOL online website (https://itol.embl.de/).

For evolutionary analysis within the water buffalo species, the LINE-1 elements homologous with the 2,809 bp long white buffalo *ASIP* LINE-1 insertion were identified using the RepeatMasker software. Sequence homology analysis was done using the cross_match software v1.09 (http://www.phrap.org/consed/) with the parameters “-gap_init -25 -gap_ext -5 -minscore 10 -minmatch 6 -alignments -bandwidth 50 -word_raw”. The minimum spanning trees of LINE-1s were constructed using the COSEG software v0.2.2 (http://www.repeatmasker.org/) to define the subfamilies.

Haplotype network of the LINE-1 insertion region

SNPs in a genomic region of 10 Kb flanking the LINE-1 insertion point upstream of *ASIP* (BBU14:19,991,854-20,001,504) were used for haplotype analysis. In addition to 63 swamp buffaloes that were whole-genome sequenced in this study, we used 10 river buffaloes for comparative purpose (NCBI Sequence Read Archive SRR4477876-SRR4477880, SRR4477882-SRR4477884, SRR4477888 and SRR4477890, Whitacre et al. 2017). Haplotypes were phased using the Beagle software v4.1 (Browning & Browning 2007). A median-joining network and a maximum likelihood (ML) evolutionary tree based on Tamura-Nei model were constructed using the Network software v5.0.1.1 (Bandelt et al. 1999) and the MEGA7 software (Kumar et al. 2016), respectively. To construct the ML tree, a discrete Gamma distribution (5 categories) was used to model evolutionary rate differences among sites.

Data availability

Whole-genome sequencing data generated in this study have been submitted to the NCBI Sequence Read Archive (SRA) as BioProject ID PRJNA633919.

Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgements

This work was supported by the National Natural Scientific Foundation of China (grant number 31561143010) and the China Agricultural Research System (grant number CARS-36). Linzhao Fang was funded through Health Data Research UK (HDR-UK) (grant number HDR-

9004) and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Actions (MSCA) (grant number 801215). We appreciate the Chinese Government's contribution to the Chinese Academy of Agricultural Sciences (CAAS)-International Livestock Research Institute (ILRI) Joint Laboratory on Livestock and Forage Genetic Resources in Beijing (2018-GJHZ-01) and this article contributes to the Consortium of International Agricultural Research Centers (CGIAR) Research Program on Livestock. We thank Ian J Jackson (University of Edinburgh) for his insightful comments on the manuscript. We also gratefully acknowledge the critical review of our manuscript by three anonymous reviewers.

References

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*. 21:974-984.

Albrecht E, Komolka K, Kuzinski J, Maak S. 2012. Agouti revisited: transcript quantification of the *ASIP* gene in bovine tissues related to protein expression and localization. *PLoS One*, 7:e35282.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet*. 41:1061-1067.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol*. 16:37-48.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*. 21:263-265.

Beck CR, Garcia-Perez JL, Badge RM, Moran JV. 2011. LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet*. 12:187-215.

Boissinot S, Sookdeo A. 2016. The evolution of LINE-1 in vertebrates. *Genome Biol Evol*. 8:3485-3507.

Botchkareva NV, Botchkarev VA, Gilchrest BA. 2003. Fate of melanocytes during development of the hair follicle pigmentary unit. *J Investig Dermatol Symp Proc*. 8:76-79.

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH, Jr. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A*. 100:5280-5285.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*. 81:1084-1097.

Bruders R, Van Hollebeke H, Osborne EJ, Kronenberg Z, Maclary E, Yandell M, Shapiro MD. 2020. A copy number variant is associated with a spectrum of pigmentation patterns in the rock pigeon (*Columba livia*). *PLoS Genet*. 16:e1008274.

Burns KH. 2017. Transposable elements in cancer. Nat Rev Cancer. 17(7):415-424.

Cal L, Suarez-Bregua P, Comesana P, Owen J, Braasch I, Kelsh R, Cerda-Reverter JM, Rotllant J. 2019. Countershading in zebrafish results from an Asip1 controlled dorsoventral gradient of pigment cell differentiation. *Sci Rep*. 9:3449.

Ceinos RM, Guillot R, Kelsh RN, Cerda-Reverter JM, Rotllant J. 2015. Pigment patterns in adult fish result from superimposition of two largely independent pigmentation mechanisms. *Pigment Cell Melanoma Res*. 28:196-209.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*. 6:677-681.

Cieslak M, Reissmann M, Hofreiter M, Ludwig A. 2011. Colours of domestication. Biol Rev *Camb Philos Soc*. 86:885-899.

Criscione SW, Theodosakis N, Micevic G, Cornish TC, Burns KH, Neretti N, Rodic N. 2016. Genome-wide characterization of human L1 antisense promoter-driven transcripts. *BMC Genomics*. 17:463.

Cuthill IC, Allen WL, Arbuckle K, Caspers B, Chaplin G, Hauber ME, Hill GE, Jablonski NG, Jiggins CD, Kelber A, et al. 2017. The biology of color. *Science*. 357:eaan0221.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics*. 27:2156-2158.

De Cecco M, Ito T, Petrashen AP, Elias AE, Skvir NJ, Criscione SW, Caligiana A, Brocculi G, Adney EM, Boeke JD, et al. 2019. L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature*. 566:73-78.

Diehl AG, Ouyang N, Boyle AP. 2020. Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. *Nat Commun*. 11:1796.

Dreger DL, Schmutz SM. 2011. A SINE insertion causes the black-and-tan and saddle tan phenotypes in domestic dogs. *J Hered*. 102:S11-18.

Duhl DMJ, Vrieling H, Miller KA, Wolff GL, Barsh GS. 1994. Neomorphic *agouti* mutations in obese yellow mice. *Nat Genet*. 8, 59-65.

Elbarbary RA, Lucas BA, Maquat LE. 2016. Retrotransposons as regulators of gene expression. *Science*. 351:aac7247.

Fang M, Larson G, Ribeiro HS, Li N, Andersson L. 2009. Contrasting mode of evolution at a coat color locus in wild and domestic pigs. *PLoS Genet*. 5:e1000341.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet*. 41:563-571.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 28:3150-3152.

Furumura M, Sakai C, Potterf SB, Vieira WD, Barsh GS, Hearing VJ. 1998. Characterization of genes modulated during pheomelanogenesis using differential display. *Proc Natl Acad Sci U S A*. 95:7374-7378.

Gautier M, Klassmann A, Vitalis R. 2017. rehh 2.0: A reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol Ecol Resour*. 17:78-90.

Girardot M, Guibert S, Laforet MP, Gallard Y, Larroque H, Oulmouden A. 2006. The insertion of a full-length *Bos taurus* LINE element is responsible for a transcriptional deregulation of the Normande Agouti gene. *Pigment Cell Res*. 19:346-355.

Goodier JL, Kazazian HH, Jr. 2008. Retrotransposons revisited: The restraint and rehabilitation of parasites. *Cell*. 135:23-35.

Haase B, Brooks SA, Schlumbaum A, Azor PJ, Bailey E, Alaeddine F, Mevissen M, Burger D, Poncet PA, Rieder S, et al. 2007. Allelic heterogeneity at the equine *KIT* locus in dominant white (W) horses. *PLoS Genet*. 3:e195.

Hancks DC, Kazazian HH, Jr. 2016. Roles for retrotransposon insertions in human disease. *Mob DNA*. 7:9.

Haupaix N, Curantz C, Bailleul R, Beck S, Robic A, Manceau M. 2018. The periodic coloration in birds forms through a prepattern of somite origin. *Science*. 361:1216-1216.

Henkel J, Saif R, Jagannathan V, Schmocker C, Zeindler F, Bangerter E, Herren U, Posantzis D, Bulut Z, Ammann P, et al. 2019. Selection signatures in goats reveal copy number variants underlying breed-defining coat color phenotypes. *PLoS Genet*. 15:e1008536.

Hida T, Wakamatsu K, Sviderskaya EV, Donkin AJ, Montoliu L, Lynn Lamoreux M, Yu B, Millhauser GL, Ito S, Barsh GS, et al. 2009. Agouti protein, mahogunin, and attractin in pheomelanogenesis and melanoblast-like alteration of melanocytes: A cAMP-independent pathway. *Pigment Cell Melanoma Res*. 22:623-634.

Hoekstra HE. 2006. Genetics, development and evolution of adaptive pigmentation in vertebrates. *Heredity*. 97:222-234.

Hubbard JK, Uy JAC, Hauber ME, Hoekstra HE, Safran RJ. 2010. Vertebrate pigmentation: From underlying genes to adaptive function. *Trends Genet*, 26:231-239.

Inaba M, Jiang TX, Liang YC, Tsai S, Lai YC, Widelitz RB, Chuong CM. 2019. Instructive role of melanocytes during pigment pattern formation of the avian skin. *Proc Natl Acad Sci U S A*. 116:6884-6890.

Jang HS, Shah NM, Du AY, Dailey ZZ, Pehrsson EC, Godoy PM, Zhang D, Li D, Xing X, Kim S, et al. 2019. Transposable elements drive widespread expression of oncogenes in human cancers. *Nat Genet*. 51:611-617.

Jones MR, Mills LS, Alves PC, Callahan CM, Alves JM, Lafferty DJR, Jiggins FM, Jensen JD, Melo-Ferreira J, Good JM. 2018. Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. *Science*. 360:1355-1358.

Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NH, Zody MC, Anderson N, Biagi TM, Patterson N, Pielberg GR, Kulbokas III EJ, et al. 2007. Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat Genet*. 39:1321-1328.

Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform*. 20:1160-1166.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 37:907-915.

Kingsley EP, Manceau M, Wiley CD, Hoekstra HE. 2009. Melanism in *peromyscus* is caused by independent mutations in *agouti*. *PLoS ONE*. 4:e6435.

Klebig ML, Wilkinson JE, Geisler JG, Woychik RP. 1995. Ectopic expression of the agouti gene in transgenic mice causes obesity, features of type II diabetes, and yellow fur. *Proc Natl Acad Sci U S A*. 92:4728-4732.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 27:722-736.

Kratochwil CF, Liang Y, Gerwin J, Woltering JM, Urban S, Henning F, Machado-Schiaffino G, Hulsey CD, Meyer A. 2018. Agouti-related peptide 2 facilitates convergent evolution of stripe patterns across cichlid fish radiations. *Science*. 362:457-460.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 33:1870-1874.

Lavie L, Maldener E, Brouha B, Meese EU, Mayer J. 2004. The human L1 promoter: Variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res*. 14:2253-2260.

Le Pape E, Passeron T, Giubellino A, Valencia JC, Wolber R, Hearing VJ. 2009. Microarray analysis sheds light on the dedifferentiating role of agouti signal protein in murine melanocytes via the Mc1r. *Proc Natl Acad Sci U S A*. 106:1802-1807.

Li H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*. 34:3094-3100.

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

Li W, Bickhart DM, Ramunno L, Iamartino D, Williams JL, Liu GE. 2019. Comparative sequence alignment reveals river buffalo genomic structural differences compared with cattle. *Genomics*. 111: 418-425.

Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. 2019. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*. 47:199-205.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 15:550.

Low WY, Tearle R, Bickhart DM, Rosen BD, Kingan SB, Swale T, Thibaud-Nissen F, Murphy TD, Young R, Lefevre L, et al. 2019. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nature Commun*. 10:260.

Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. *Cell*. 72:595-605.

Luo X, Zhou Y, Zhang B, Zhang Y, Wang X, Feng T, Li Z, Cui K, Zhang Z, Luo C, et al. 2020. Understanding divergent domestication traits from the whole-genome sequencing of swamp- and river-buffalo populations. *Natl Sci Rev*. 7:686-701.

Mallarino R, Henegar C, Mirasierra M, Manceau M, Schradin C, Vallejo M, Beronja S, Barsh GS, Hoekstra HE. 2016. Developmental mechanisms of stripe patterns in rodents. *Nature*. 539:518-523.

Manceau M, Domingues VS, Mallarino R, Hoekstra HE. 2011. The developmental role of Agouti in color pattern evolution. *Science*. 331:1062-1065.

Martin A, Orgogozo V. 2013. The loci of repeated evolution: A catalog of genetic hotspots of phenotypic variation. *Evolution*. 67:1235-1250.

Merenciano M, Ullastres A, de Cara MA, Barron MG, Gonzalez J. 2016. Multiple independent retroelement insertions in the promoter of a stress response gene have variable molecular and functional effects in *Drosophila*. *PLoS Genet*. 12:e1006249.

Michaud EJ, Van Vugt MJ, Bultman SJ, Sweet HO, Davisson MT, Woychik RP. 1994. Differential expression of a new dominant agouti allele (A^{IAPY}) is correlated with methylation state and is influenced by parental lineage. *Genes Dev*. 8:1463-1472.

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH, Jr. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell*. 87:917-927.

Mort RL, Jackson IJ, Patton EE. 2015. The melanocyte lineage in development and disease. *Development*. 142:620-632.

Nigumann P, Redik K, Matlik K, Speek M. 2002. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics*. 79:628-634.

Norris BJ, Whan VA. 2008. A gene duplication affecting expression of the ovine *ASIP* gene is responsible for white and black sheep. *Genome Res*. 18:1282-1293.

Ostertag EM, Kazazian HH, Jr. 2001. Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res*. 11:2059-2065.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 33:290-295.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol*. 26:1641-1650.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 81:559-575.

Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. 2015. The influence of LINE-1 and SINE retrotransposons on mammalian genomes. *Microbiol Spectr*. 3:MDNA3-0061-2014.

Rife DC. 1962. Color and horn variations in water buffalo: The inheritance of coat color, eye color and shape of horns. *J Hered*. 53:239-246.

Rife DC, Buranamanas P. 1959. Inheritance of white coat color in the water buffalo of Thailand. *J Hered*. 50:269-272.

Robic A, Morisson M, Leroux S, Gourichon D, Vignal A, Thebault N, Fillon V, Minvielle F, Bed'Hom B, Zerjal T, et al. 2019. Two new structural mutations in the 5' region of the *ASIP* gene cause diluted feather color phenotypes in Japanese quail. *Genet Sel Evol*. 51:12.

Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. 2017. Variant Review with the Integrative Genomics Viewer (IGV). *Cancer Res*. 77:31-34.

Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, Rowan TN, Low WY, Zimin A, Couldrey C, et al. 2020. *De novo* assembly of the cattle reference genome with single-molecule sequencing. *GigaScience*. 9:giaa021.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature*. 419:832-837.

Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH. 1997. Many human L1 elements are capable of retrotransposition. *Nat Genet*.16:37-43.

Schiaffino MV. 2010. Signaling pathways in melanosome biogenesis and pathology. *Int J Biochem Cell Biol*. 42:1094-1104.

Speek M. 2001. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol Cell Biol*. 21:1973-1985.

Steel KP, Davidson DR, Jackson IJ. 1992. TRP-2/DT, a new early melanoblast marker, shows that steel growth factor (c-kit ligand) is a survival factor. *Development*. 115:1111-1119.

Sviderskaya EV, Hill SP, Balachandar D, Barsh GS, Bennett DC. 2001. Agouti signaling protein and other factors modulating differentiation and proliferation of immortal melanoblasts. *Dev Dyn*. 221:373-379.

Tanave A, Imai Y, Koide T. 2019. Nested retrotransposition in the East Asian mouse genome causes the classical *nonagouti* mutation. *Commun Boil*. 2:283.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 7:562-578.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 38:e164.

Whitacre LK, Hoff JL, Schnabel RD, Albarella S, Ciotola F, Peretti V, Strozzi F, Ferrandi C, Ramunno L, Sonstegard TS, et al. 2017. Elucidating the genetic basis of an oligogenic birth defect using whole genome sequence data in a non-model organism, *Bubalus bubalis*. *Sci Rep*. 7:39719.

Williams JL, Iamartino D, Pruitt KD, Sonstegard T, Smith TPL, Low WY, Biagini T, Bomba L, Capomaccio S, Castiglioni B, et al. 2017. Genome assembly and transcriptome resource for river buffalo, *Bubalus bubalis* (2n = 50). *Gigascience.* 6: gix088.

Young R, Lefevre L, Bush SJ, Joshi A, Singh SH, Jadhav SK, Dhanikachalam V, Lisowski ZM, Iamartino D, Summers KM, et al. 2019. A gene expression atlas of the domestic water buffalo (*Bubalus bubalis*). *Front Genet*. 10:668.

Zhang Y, Colli L, Barker JSF. 2020. Asian water buffalo: Domestication, history and genetics. *Anim Genet*. 51:177-191.

Zou Z, Zhang J. 2015. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Mol Biol Evol*. 32:2085-2096.

Figure legends

FIG. 1. Whole-genome sequencing identifies that a 1.07 Mb long region including *ASIP* on buffalo chromosome 14 (BBU14) is associated with the white coat phenotype. (A) Manhattan plot showing a single genomic region on BBU14 significantly associated with the white coat phenotype. (B) Zoomed association signals of GWAS and genetic differentiation (F_{ST}) in the significant region. (C) Validation of the strong association of the 1.07 Mb long region including *ASIP* with the white coat phenotype, by genotyping 20 variants in an independent sample and linkage disequilibrium analysis. The location of *ASIP* is indicated by a blue bar. (D) Genes in the significant region. (E) Bifurcation diagram showing the distinct long-range haplotypes in white (red) and black (blue) buffaloes. Moving from the core SNP (vertical dot line) in two directions, the diagram divides if two alleles are present at the SNP. The x-axis is the chromosome location. The thickness of the lines represents the counts of long-range haplotype in the sample.

FIG. 2. Identification of the LINE-1 insertion in *ASIP* of white buffaloes. (A) Significantly higher expression of *ASIP* in white buffalo skin (W1, W2 and W3) than in black buffalo skin (B1, B2 and B3) revealed by RNA-seq and qPCR. Three experimental replicates of qPCR are shown separately (qPCR1, qPCR2 and qPCR3). (B) Distinct transcripts assembled from RNA-seq data of skin samples of white (blue) and black (red) buffaloes. (C) Full-length transcripts of *ASIP* generated using the RACE-PCR and characterization of the LINE-1 insertion in the white buffalo *ASIP*. The structure of a full-length LINE-1 element from cattle (L1-BT; Girardot et al. 2006) is shown as reference. (D) Schematic representation showing the chromosomal position of the LINE-1 insertion determined by soft-clipped reads analysis and the partial sequences of the LINE-1 insertion obtained by *de novo* assembling of the soft-clipped reads. (E) Genotyping the LINE-1 insertion using the allele-specific PCR and its perfect association with white coat phenotype. PCR products of wild allele (Normal; 296 bp) and mutant allele (Ins; 387 bp) are separated for six samples (S1-S6) using agarose gel electrophoresis.

FIG. 3. Functional consequences of *ASIP* over-expression in white buffalo skin. (A) Transcription levels of *ASIP*, *TRY*, *DCT (TYRP2)*, *TYRP1*, *KIT*, *KITLG*, *MITF*, *MC1R* and *EDNRB*. Error bars indicate standard deviation (SD). "**" shows significance level ($P < 0.01$) of a Wald test using the DESeq2 package. (B) Immunohistochemical analysis of skin samples of black (top) and white (bottom) buffaloes. Arrowheads indicate melanocytes with tyrosinase-related protein 1 (TYRP1) expression (Trp+). (C) Haematoxylin Eosin (HE) staining of skin samples of black (top) and white (bottom) buffaloes. Arrowheads indicate melanin pigment near the melanocytes.

FIG. 4. Model of molecular mechanism for swamp buffalo white coat phenotype. The gene is represented 3'-5' from left to right to be consistent with that shown in FIG. 2.

FIG. 5. Evolutionary analyses of LINE-1 elements in the genome and *ASIP* haplotypes. (A) An approximate maximum-likelihood tree of 6,986 full-length LINE-1 elements in buffalo, cattle and yak reference genomes. The buffalo includes both swamp and river buffaloes. The clade on the left are LINE-1 copies that are mixed among species and have long branch lengths, indicating that they are ancient and amplified before the split of these species. Species-specific clades are found on the right, in which the buffalo-specific clade contains both swamp and river buffaloes. One arrow points to the LINE-1 clustering with that of white buffalo *ASIP* and another to the LINE-1 of cattle *ASIP* (L1-BT). The shared branch is shown in grey color. (B) A minimum spanning (MS) tree showing the evolutionary relationship among LINE-1 subfamilies in the water buffalo genomes. The *ASIP* LINE-1 insertion of white buffaloes belongs to the sub20 subfamily (in the grey square), a relatively young subfamily. (C) Median-joining network of water buffalo *ASIP* haplotypes. Colors: green = swamp buffalo; brown = river buffalo; and white = white swamp buffalo.

FIG. 6. Two independent LINE-1 insertions occurred in *ASIP* of water buffalo (*Bubalus bubalis*) and cattle (*Bos taurus*). The genomes, transcripts and LINE-1 elements are

represented 3'-5' from left to right to be consistent with that in **FIG.** 2. The LINE-1 of cattle *ASIP* is based on Girardot et al. (2006).

FIG. S1. The two coat color types in swamp buffaloes.

FIG. S2. Gene expression profile of 23 genes in 56 tissue and cell types of water buffaloes. *ASIP* shows tissue-specific expression in skin. Data are given in supplementary table S19.

FIG. S3. IGV illustration of the mapped RNA-seq reads showing the transcription differences between white and black swamp buffaloes.

FIG. S4. Full-length transcripts of *ASIP* determined using the RACE-PCR followed by Sanger sequencing of their clones. RACE-PCR products were inserted into plasmid vectors for Sanger sequencing. (A) The full-length transcripts of *ASIP* in white and black buffaloes. Length and frequency are shown for six transcripts (B1~B6) of black buffaloes and one transcript (W1) of white buffaloes. (B) cDNA sequence of the white buffalo *ASIP* transcript determined using the RACE-PCR (W1). The 165 bp non-coding exon derived from a LINE-1 insertion was shown in yellow color (L1 exon).

FIG. S5. Positioning of the white buffalo LINE-1 insertion and *de novo* assembling of its partial sequences by soft-clipped read analysis of whole-genome sequencing data. (A) Differences between white (pool of 10 individuals) and black buffaloes (pool of 10 individuals) in soft-clipped reads along the chromosomal positions upstream of *ASIP*. The top signal is highlighted by a red square. (B) IGV screenshot of the mapped reads surrounding the 16 bp genomic region between BBU14:19,996,806 and BBU14:19,996,791 (in the box). (C) *de novo* assembling of the soft-clipped reads generated a contig of 269 bp (blue) and a contig of 257 bp (red). The duplication of the 16 bp region is shown in orange color. The unknown sequence is

represented by a stretch of Ns. (D) Diagram to clarify the position of the insertion and two contigs assembled from soft-clipped reads.

FIG. S6. The approximate maximum-likelihood tree of 6,986 full-length LINE-1 elements extracted from water buffalo, cattle and yak reference genomes. Swamp buffalo and river buffalo are shown in different colors.

FIG. S7. Phylogenetic analyses of *ASIP* haplotypes. (A) Alignment of *ASIP* haplotypes. There are two haplogroups of haplotypes (SW1 and SW2) in swamp buffaloes. The white buffalo-specific haplotypes belong to SW2. The LINE-1 insertion is shown by the red letter L. (B) Maximum likelihood tree of *ASIP* haplotypes. The white buffalo-specific haplotypes are indicated by red triangles.

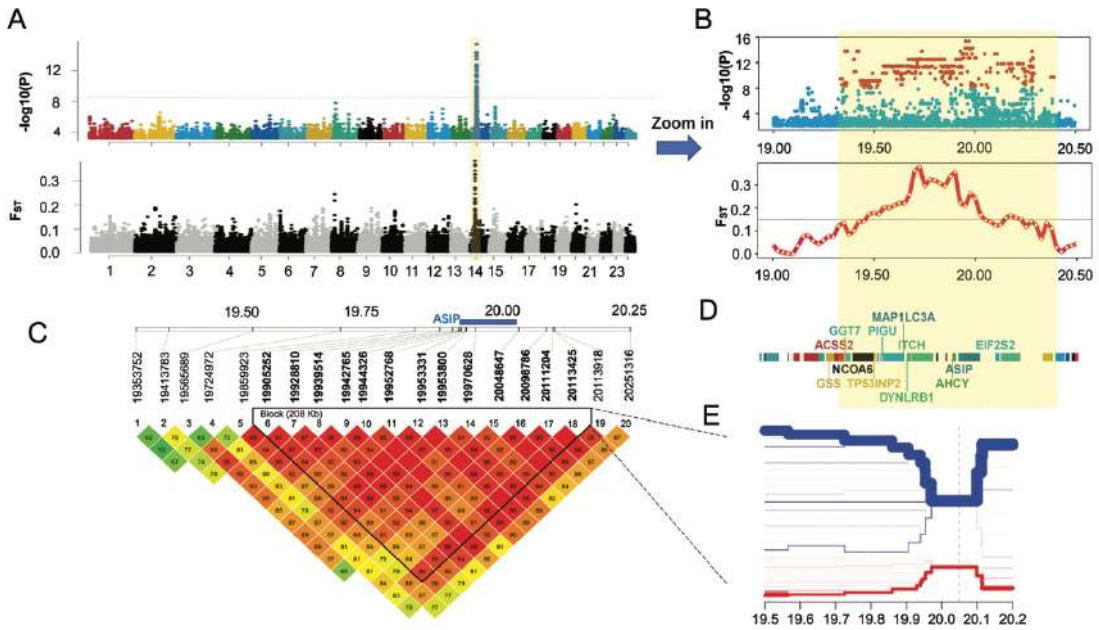
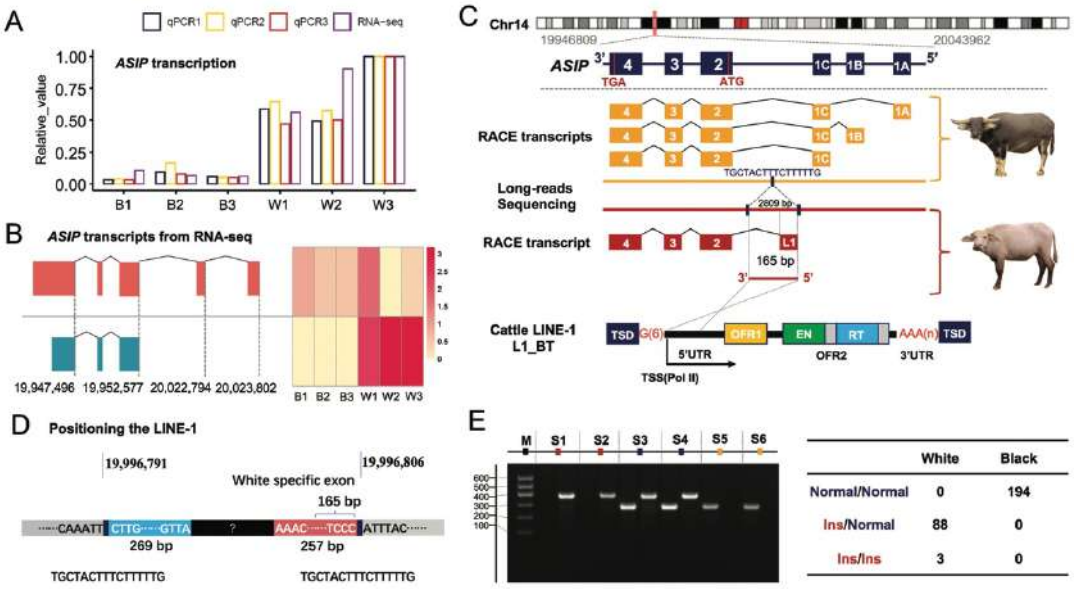


FIG. 1



	White	Black
Normal/Normal	0	194
Ins/Normal	88	0
Ins/Ins	3	0

FIG. 2

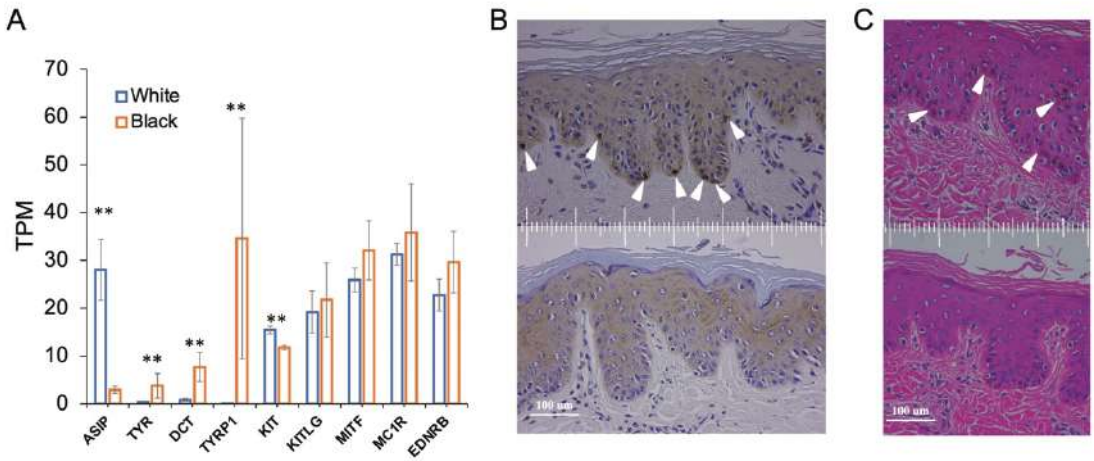


FIG. 3

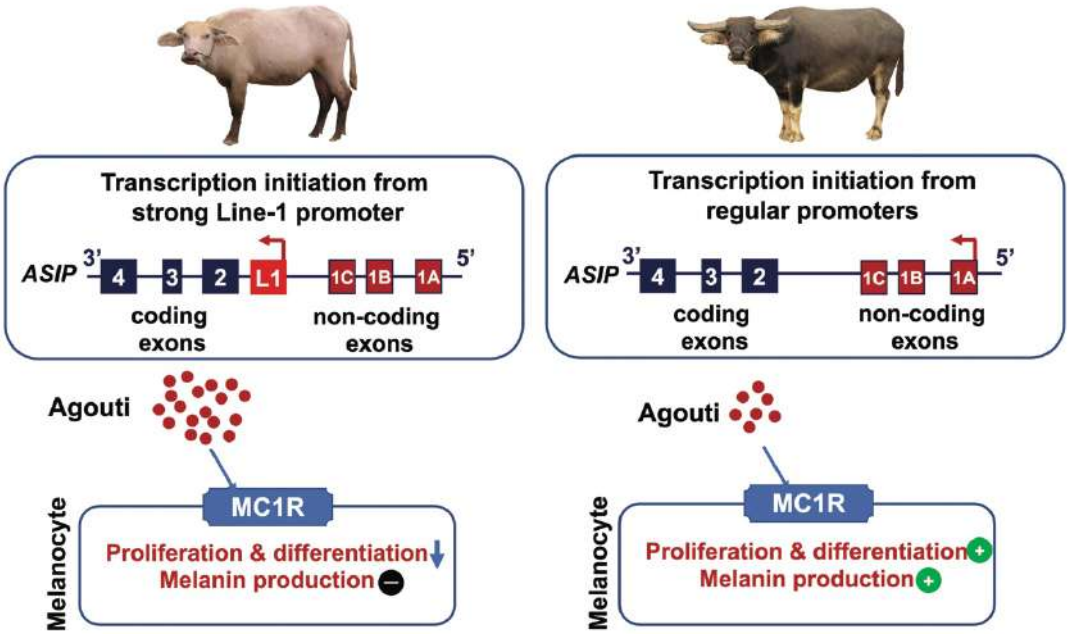


FIG. 4

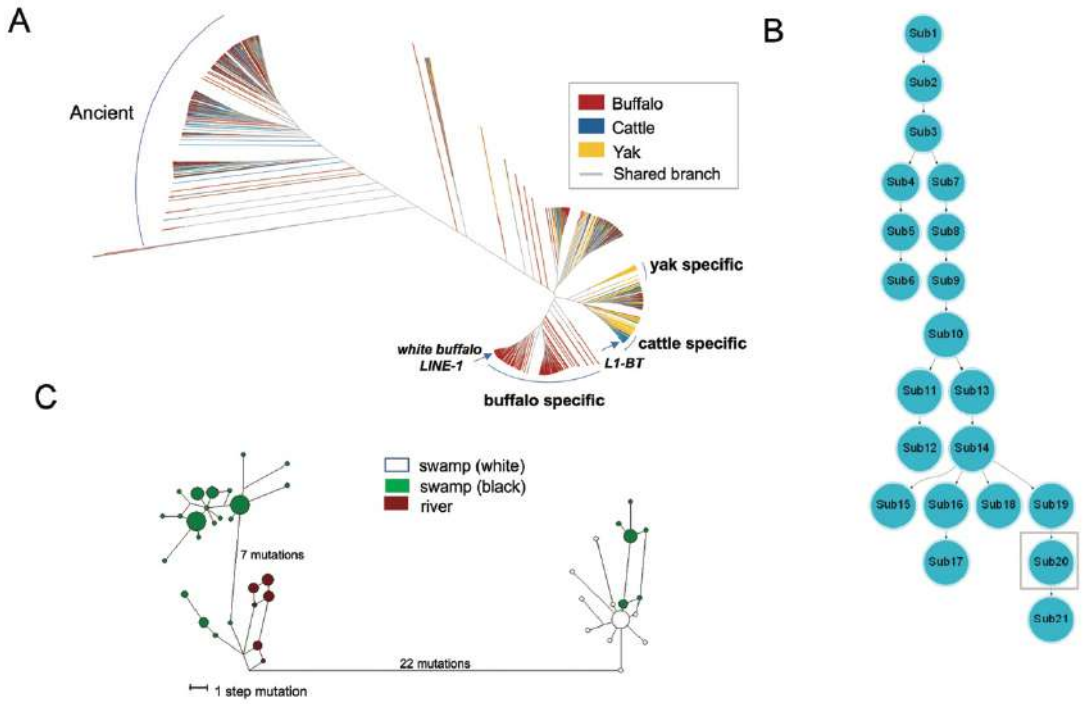


FIG. 5

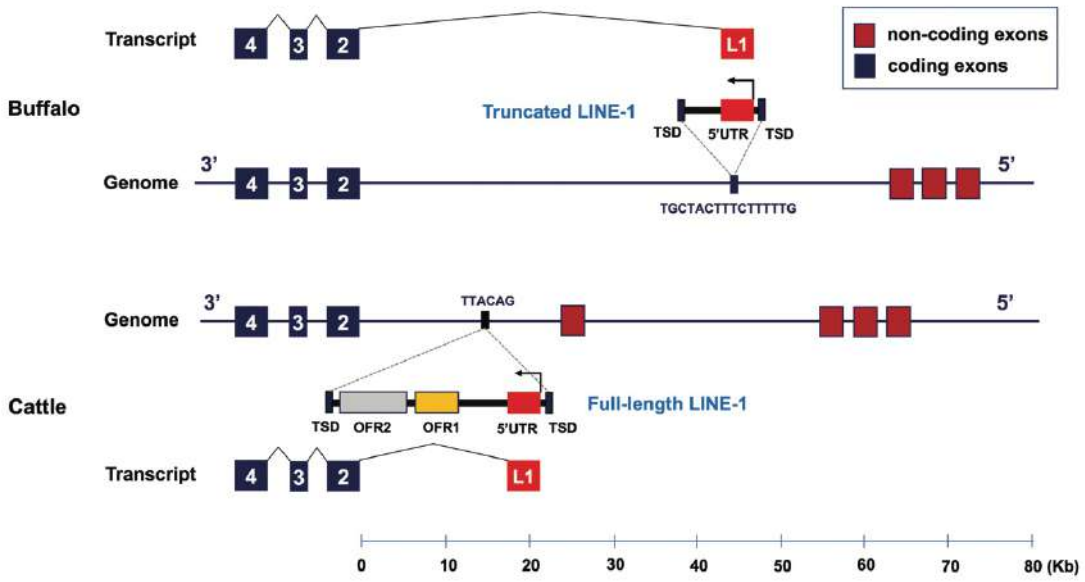


FIG. 6