Annotation of novel transcripts putatively relevant for bovine fat metabolism

**Annett Eberlein · Claudia Kalbe · Tom Goldammer ·
Ronald M. Brunner · Christa Kuehn ·
Rosemarie Weikard**

Abstract Two bovine transcripts encoded by the inter-leukin-1 receptor-associated kinase 1 (*IRAK1*) gene and the locus *LOC618944* predicted as similar to human chromosome 6 open reading frame 52 (*C6orf52*) gene had indicated divergent expression in bovine skeletal muscle containing different amount of intramuscular fat in a pilot screening experiment. However, for both loci any role in the regulation of energy or fat metabolism is not yet described. In this study, we validated and refined gene structure, screened for mRNA splice variants and analyzed the tissue-specific gene expression patterns of both loci as a prerequisite to elucidate their potential physiological function. Based on comparative sequence analysis, a new full-length gene model for the bovine *IRAK1* gene was developed and confirmed experimentally. Expression of *IRAK1* mRNA was found in a variety of tissues, and a splice variant was identified in skeletal muscle caused by an in-frame deleted segment of 210 bp affecting regions of intrinsic disorder in the respective protein. For the locus *LOC618944*, our data contributed to a revised gene model and its assignment to BTA23 (bovine chromosome 23) on the current bovine genome assembly supported by comparative similarity analysis between the bovine and human genomes and experimental data. Furthermore, we identified several splice variants in mammary gland, fat and skeletal muscle tissue and detected a highly similar processed pseudogene on BTA26. All transcript variants of *LOC618944* detected in the analyzed tissues represent noncoding RNAs. For both loci, our results suggest yet undetected physiological functions in tissues relevant for fat or energy metabolism in cattle.

Keywords Cattle · mRNA expression · Splice variant · IRAK1 · *C6orf52* · Noncoding RNA

A. Eberlein · T. Goldammer · R. M. Brunner · C. Kuehn ·
R. Weikard (✉)
Research Unit Molecular Biology, Research Institute for the
Biology of Farm Animals (FBN), 18196 Dummerstorf, Germany
e-mail: weikard@fbn-dummerstorf.de

A. Eberlein
e-mail: eberlein@fbn-dummerstorf.de

T. Goldammer
e-mail: tomgoldammer@fbn-dummerstorf.de

R. M. Brunner
e-mail: brunner@fbn-dummerstorf.de

C. Kuehn
e-mail: kuehn@fbn-dummerstorf.de

C. Kalbe
Research Unit Muscle Biology and Growth, Research Institute
for the Biology of Farm Animals (FBN), 18196 Dummerstorf,
Germany
e-mail: kalbe@fbn-dummerstorf.de

Introduction

Farm animals provide well-suited resource populations for the functional investigation of quantitative traits in mammals, because they exhibit a substantial natural variation in a wide range of target traits and can be monitored under standardized environmental conditions [1]. In cattle, the amount, composition and distribution of intramuscular and depot fat show major inter-individual differences [2]. Thus, this species is a well-suited model to investigate the genetic background of divergent energy and fat metabolism. Different approaches are used to identify putative genetic factors like Quantitative Trait Loci (QTL) studies or functional candidate gene analyses [3–7]. For a systematic approach to dissect loci related to intramuscular fat

metabolism on a molecular level, it is essential to obtain a comprehensive picture of the genes and transcripts that are involved in the physiological pathways responsible for the expression of the target phenotype.

In order to identify functionally relevant transcripts in additional to the well-described major genes encoding enzymes of fat metabolism, an initial expression screening experiment had been performed. This experiment investigating male full-sib F_2 individuals from a Holstein × Charolais resource population [8], which were divergent regarding their intramuscular fat deposition, revealed two loci that showed an indication on differential gene expression, but lacked a functional annotation within known regulatory pathways of energy and fat metabolism (Kalbe et al., unpublished results). While for the *IRAK1* (interleukin-1 receptor-associated kinase 1) gene a role in the regulation of immune response, however not in energy or fat metabolism, was described [9, 10], the second locus *LOC618944* predicted as similar to human chromosome 6 open reading frame 52 (*C6orf52*) was deposited as noncoding RNA without providing any functional annotation.

For a subsequent investigation of the functional relevance of these transcripts, a correct and conclusive structural gene annotation is a prerequisite. Comparative sequence data from well-annotated genomes like human and mouse can aid in this process for species with less developed genomic and functional information. In cattle, more than 4,000 genes of the approximately 22,000 protein coding genes have been manually annotated [11] indicating that by far, not all bovine transcripts are sufficiently annotated in the current bovine genome assembly. Therefore, the aim of our study was the experimental confirmation of the structure and transcription of the loci identified in the initial expression screening experiment, for which a respective functional annotation in energy and fat metabolism was missing as a precondition for the elucidation of the potential physiological function of the transcripts.

Materials and methods

Selection of novel functional transcripts

A previous microarray experiments using the 24K GeneChip® Bovine Genome Array (Affymetrix) had been carried out with male full-sib individuals from a bovine F_2 Charolais × German Holstein resource population [8] kept under carefully monitored identical environmental conditions but exhibiting extreme differences in intramuscular fat content (Kalbe et al., unpublished results). For all probe sets of the array, the respective representative transcripts and presumably related genes were identified according to

the Affymetrix annotation analysis tool 'NetAffx Analysis Center' (https://www.affymetrix.com/analysis/netaffx/) extracting sequence annotation information available from public databases. Within the list of Affymetrix probe sets suggesting differential expression in skeletal muscle tissue (M. longissimus dorsi) of trait-differentiated animals, two probe sets (Bt.5221.1.S1_at and Bt.10616.1.S1_a_at) had deficient or no unequivocal structural and functional annotation information. The selected respective transcripts were suggested as putative candidates with yet undetected function in fat and energy metabolism in cattle.

In silico sequence analyses

To identify sequences highly similar to the target transcripts that may allow to retrieve structural and functional annotation of the genes presumably represented by the probe sets, in silico sequence similarity searches were carried out against the bovine genome sequence assembly available at the National Center for Biotechnical Information (NCBI) [Btau4.0, Btau3.1 http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid=9913), the alternative bovine genome sequence assembly UMD 3.0 (ftp://ftp.cbcb.umd.edu/pub/data/assembly/Bos_taurus/Bos_taurus_UMD_3.0/)], and the nucleotide and protein databases at NCBI using BLAST tools [12, 13]. For further structural and functional analyses, ORF (open reading frame) finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html), CD (conserved domain) search (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) and the Simple Modular Architecture Research Tool (http://smart.embl-heidelberg.de/) were applied.

Gene expression analyses

The qualitative gene expression patterns of the two transcripts putatively representing the *IRAK1* gene and the predicted locus *LOC618944* were determined by reverse transcription (RT)–PCR in bovine tissues relevant for energy and fat metabolism from a lactating F_2 cow belonging to a Charolais × German Holstein resource population.

Total RNA was extracted from skeletal muscle (M. semitendinosus, M. longissimus dorsi), heart, intestinal and subcutaneous fat, liver, mammary gland, kidney, adrenal gland, pituitary gland, brain, thyroid gland, duodenum, jejunum, colon, spleen, lymphatic node, and rumen. Generally, the Total RNA Isolation kit NucleoSpin® Extract II (Macherey & Nagel) was used for RNA extraction according to the instructions of the manufacturer. For skeletal muscle, fat, spleen, lymphatic node, duodenum, jejunum, colon, total RNA was extracted using the TRIzol™ reagent (Invitrogen) and consecutively in a

second step, purified with the Total RNA Isolation kit NucleoSpin® Extract II (Macherey & Nagel). Genomic DNA was carefully eliminated from RNA preparations by repeated on-column digestion using twice the amount of RNAse-free DNaseI solution as recommended by the manufacturer. The cDNA was generated from 2 μg total RNA by RT with an oligo(dT)$_{12–18}$ primer and the Super-Script^{TM} First-Strand Synthesis System for RT–PCR (Invitrogen). The cDNA was purified with the High Pure PCR Product Purification kit (Roche). Qualitative PCRs with cDNAs from each tissue were performed in a 20 μl reaction volume containing 0.4 μM dNTPs, 0.5 U GoTaq® DNA polymerase (Promega) and 0.4 μM of a gene specific primer pair (Table 1). Thermocycling was performed with a touch-down protocol from 68°C to the appropriate annealing temperature of the respective primer pair in 1°C steps followed by 30 cycles at the annealing temperature of the respective primer pair and extension at 72°C.

Determination of gene structure

Locus LOC618944

To validate and refine computational gene predictions, to define full-length transcripts, and to identify mRNA splice variants, supportive experiments were carried out for each transcript.

To obtain the complete 5′ end of the *LOC618944* transcript, 5′ rapid amplification of cDNA ends (RACE) was performed using the GeneRacer^{TM} kit (Invitrogen) with the gene-specific primer LOC618944_RACE1 and the nested gene-specific primer LOC618944_RACE2 derived from the predicted bovine mRNA sequence XM_871257.2 (Table 1). For this experiment, total RNA (3 μg) was extracted from skeletal muscle tissue as described above. The 5′ RACE reaction products were separated by agarose gel electrophoresis, and DNA from the two dominant bands was isolated using the NucleoSpin® Extract II kit (Macherey & Nagel). Purified DNA was cloned into the pDrive cloning vector (QIAGEN). From 96 clones, plasmid DNA was isolated with the Perfectprep® Plasmid 96 Vac Direct Bind kit (Eppendorf). The inserts of the plasmids were checked by PCR with gene specific primers (LOC618944_F1/RACE2, Table 1). The inserts of 13 selected clones were sequenced. Sequencing was performed with BigDye© sequencing chemistry on a capillary sequencer (MEGABACE, GE Healthcare) using vector specific primers.

For confirmation of the different splice variants of the *LOC618944* transcript indicated by the initial qualitative expression analysis across tissues (as described in the "Results" section) RT–PCR on skeletal muscle cDNA with primers located in exons 1 and 6 of *LOC618944* (LOC618944_F1/R1, Table 1) were performed. The generated PCR fragments were isolated, cloned and sequenced as described above.

The *LOC618944* transcript revealed high sequence similarity to two bovine genomic sequences: to the transcript XM_871257.2 on BTA23 and to a part of intron 1 of the *NT5C2* gene on BTA26 (Fig. 1). To verify the true genomic origin of the *LOC618944* transcript, a discriminating experiment with liver cDNA and genomic DNA

Table 1 Summary of PCR primers

Primer	Sequence (5′-3′)	Accession no of reference sequence	Amplicon (bp)	Annealing (°C)	Gene region
LOC618944_F1[a]	TCCTTCTTGGTTATGGCTTG	XM_871257.2	624/623[a,b]	55	Exon 2
LOC618944_R1[a]	GGTCTTGTCTGGGATTTCTG	XM_871257.2			Exon 6
LOC618944_RACE1	GGGTCTTGTCTGGGATTTCTGAGT	XM_871257.2			Exon 6
LOC618944_RACE2	GCTGGCAGGACGGGAAGTCTGG	XM_871257.2			Exon 4
IRAK1_F1	GCTTCTACTGCCTTGTCTATG	NM_001040555.1	217[b]	60	Exon 7
IRAK1_R1	GGCATCAGTCTCTCATCCAG	NM_001040555.1			Exon 9
IRAK1_cF1	CGCCCAGCACTTCTTGTA	NM_001040555.1	784[b]	55	Exon 1
IRAK1_cR1	GGCAAAGTCCACGATGTT	NM_001040555.1			Exon 7
IRAK1_cF2	AGCAGCTGTCGCGGTTTC	NM_001040555.1	420[b]	55	Exon 6/7
IRAK1_cR2	GCCTCCCGGTCTTCACGTA	NM_001040555.1			Exon 9
IRAK1_cF3	CCGGCCAGAGCAGTAGTGT	NM_001040555.1	917[b]	58	Exon 9
IRAK1_cR3	CTTCTGCAGCATCTTCCGTC	NM_001040555.1			Exon 13
IRAK1_cF4	GGGCAGCTCCATGTCCTC	BC108132.1	849[b]	62	Exon 13
IRAK1_cR4	GTGTGTGCCAAGCCAAACC	BC108132.1			Exon 14

[a] Primer pair enables simultaneous amplification of the *LOC618944* transcript and of genomic DNA similar to intron 1 of the *NT5C2* gene

[b] Amplicon generated by amplification with the respective reverse primer
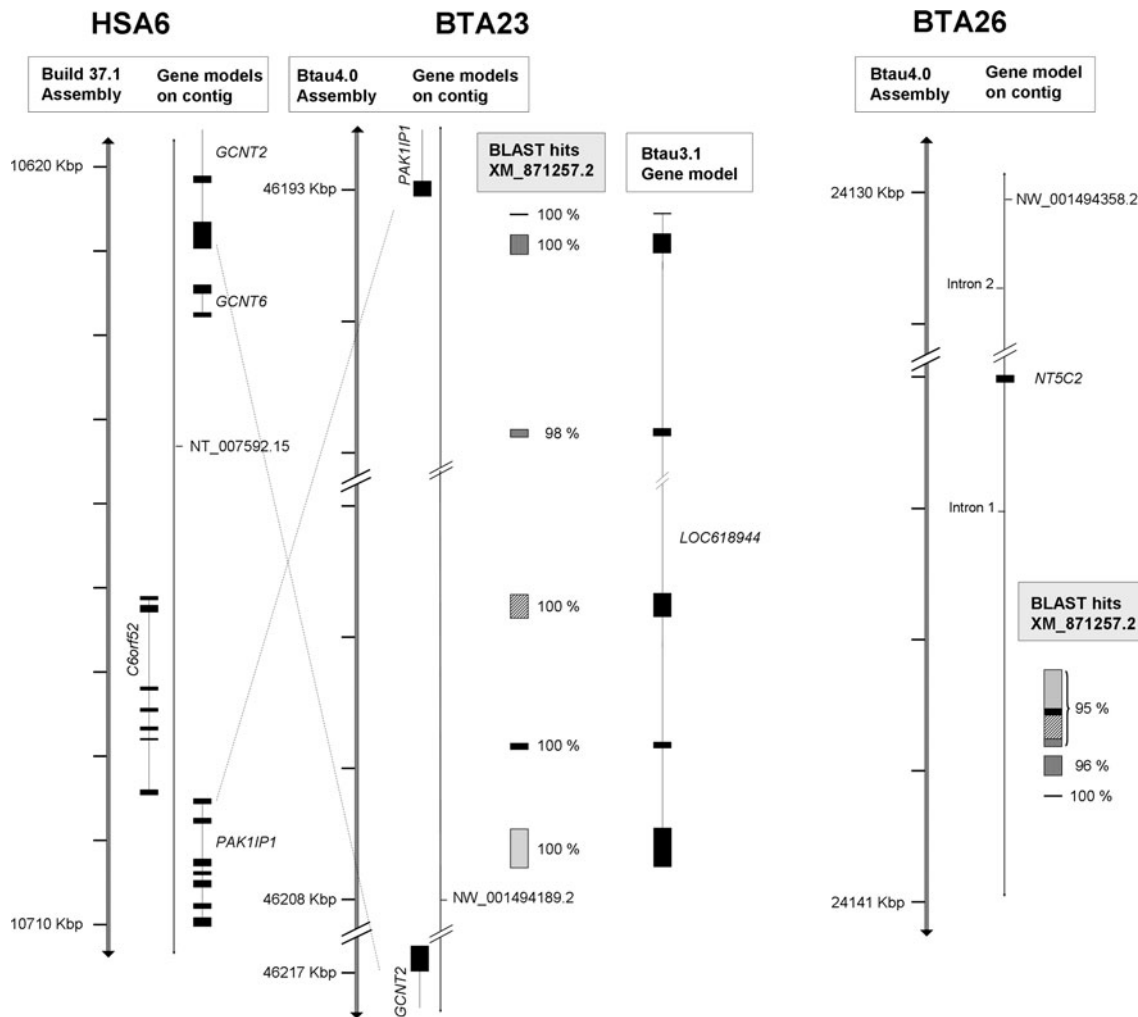
Fig. 1 Structural annotation of the bovine locus *LOC618944* on the bovine genome sequence assembly Btau4.0 based on comparative sequence alignment to the human genome (Hsa37.1) and the previous version of the bovine sequence assembly Btau3.1. XM_871257.2: predicted bovine mRNA for locus *LOC618944* similar to human *C6orf52* gene; NW_001494189.2 and NW_001494358.2: genome sequence contigs on BTA23 and BTA26, respectively; NT_007592.15: genome sequence contig on HSA6. Gene loci are given in *italic*; *black boxes* connected by *black lines*: gene models assigned to chromosomes; *patterned boxes*: XM_871257.2 sequence segments aligned along the chromosomes by sequence similarity search using BLAST tools

from the same German Holstein × Charolais F$_2$ individual was performed. For this purpose, a primer pair enabling the amplification of both loci, from BTA23 and BTA26, was used for amplification of target cDNA and genomic DNA (LOC618944_F1/R1, Table 1). If exclusively the sequence from BTA23 is expressed, we hypothesized that the amplified cDNA fragment should be identical to the putative exonic parts of the genomic sequence from NW_001494189.2 on BTA23 without any sequence variability for those positions that are divergent between NW_001494189.2 and the homologous locus in intron 1 of the *NT5C2* gene on BTA26. When testing the primer pair for amplification with genomic DNA, the generated sequence should be identical to the BTA26 sequence, because due to the large genomic distance between the primers on BTA23 (14,242 bp), the PCR conditions

applied were not appropriate to amplify the genomic BTA23 sequence.

IRAK1 gene

The coding sequence of the bovine *IRAK1* gene is represented only in parts by a cDNA clone (BC108132.1) and the current reference mRNA sequence NM_001040555.1 that is obviously incomplete towards the 3′ end. To experimentally determine the full-length cDNA sequence of the bovine *IRAK1* gene based on available transcript information, overlapping gene fragments were amplified in cDNA from liver, skeletal muscle, mammary gland and fat tissue from a lactating F$_2$ cow (as described above) with four primer pairs (IRAK1_cF1/cR1, IRAK1_cF2/cR2, IRAK1_cF3/cR3 and IRAK1_cF4/cR4, Table 1). Amplicons were isolated from

Fig. 2 Revised gene model and alternative transcript variants of the bovine locus *LOC618944*. **A** Predicted gene model for *LOC618944* (XM_871257.2) according to the bovine genome sequence assembly Btau3.1, **B** revised gene model based on in silico and experimental data representing the longest transcript variant, positions of primers (LOC618944_F1, LOC618944_R1, LOC618944_RACE1, LOC618944_RACE2) are indicated, **C** alternative splice variants identified experimentally, **D** transcript variant NR_026726 according to Smith et al. [18]
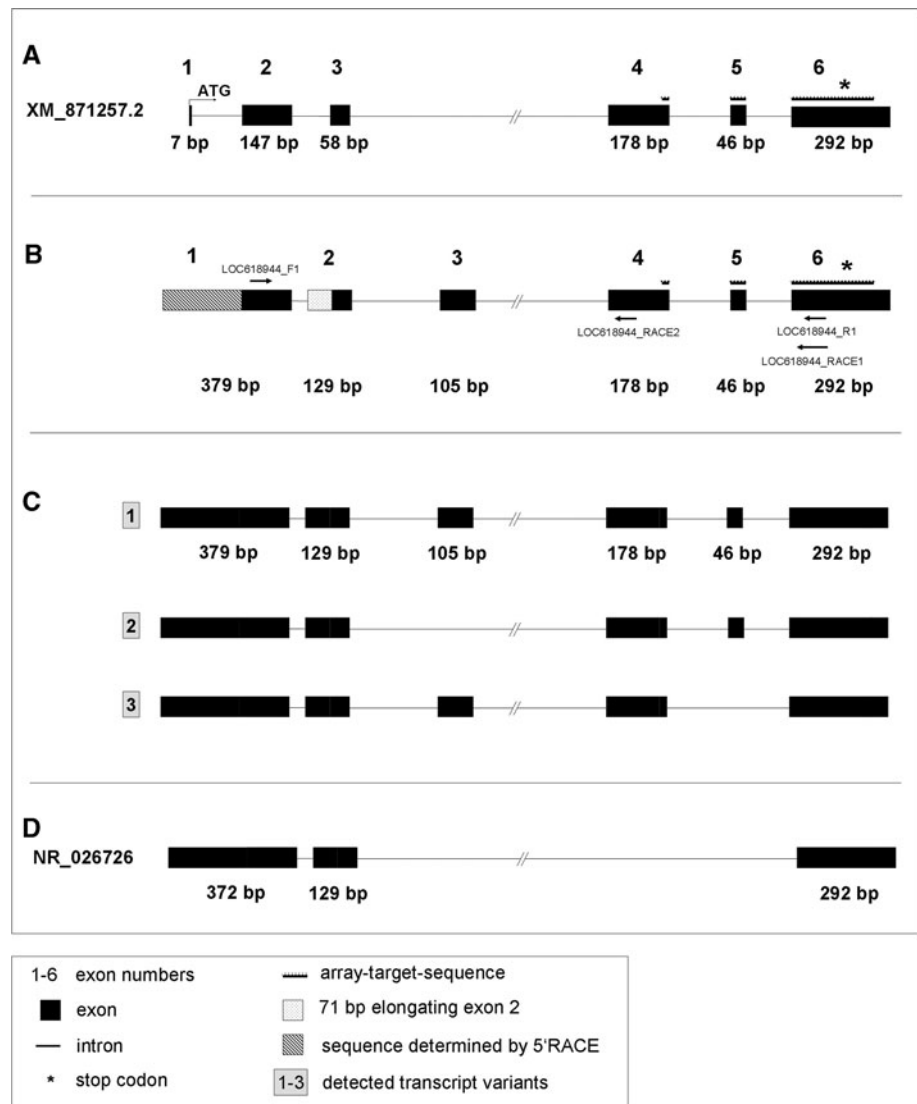
agarose gels using the NucleoSpin® Extract II kit (Macherey & Nagel) and sequenced using the respective PCR primers. The occurrence of the *IRAK1* splice variant detected with the primer pair IRAK1_cF3/cR3 was analyzed in a cDNA panel comprising two skeletal muscles (M. semitendinosus, M. longissimus dorsi), heart, intestine and subcutaneous fat, liver, mammary gland, kidney, adrenal gland, pituitary gland, brain, thyroid gland, duodenum, jejunum, colon, spleen, lymphatic node, and rumen.

Results

Revised gene model and alternative transcripts for *LOC618944*

The computationally predicted locus *LOC618944*, which is homologous to the Affymetrix probe set Bt.10616.1.S1_a_at, is denominated as similar to chromosome 6 open

reading frame 52, and the mRNA sequence XM_871257.2 was predicted in the previous version of the bovine genome sequence assembly Btau3.1. While the *LOC618944* model was placed on bovine chromosome 23 (Fig. 1) in Btau3.1, it has been removed from the current version of the bovine genome assembly (Btau4.0) as a result of standard genome annotation processing. As illustrated in Fig. 2A, according to Btau3.1, the mRNA model comprising 728 bp, is divided into 6 exons and starts directly with the ATG codon of the translation start site. The first predicted exon consists of only seven nucleotides. These facts indicated a deficient 5′ end of the *LOC618944* gene model, which prompted us to perform a 5′ RACE experiment.

Further in silico sequence analysis in the bovine genome detected a second genomic sequence highly similar to the mRNA XM_871257.2 on contig NW_001494358.2 located on BTA26 (Fig. 1). The complete mRNA sequence XM_871257.2 could be aligned to the first intron of the 5′-nucleotidase, cytosolic II gene (*NT5C2*). The aligned

region on the sequence contig NW_001494358.2 on BTA26 is separated into three parts due to an interruption of 148 nucleotides between the first two exons in the predicted mRNA model XM_871257.2 (Fig. 1) and 71 nucleotides located immediately downstream of exon 2 according to the mRNA model. The 148 nucleotides share about 92% similarity to the respective sequence of the first intron predicted on BTA23. The first part aligned on BTA26 revealed a perfect match to the seven nucleotides of exon 1 of the predicted mRNA model XM_871257.2. The second part showed a similarity of 96% to exon 2, and the third part was similar to exons 3–6 of the mRNA model XM_871257.2 by 95% without any interruption. In contrast, the structural alignment of the mRNA sequence XM_871257.2 to BTA23 predicted splicing into 6 exons interrupted by large stretches of intronic sequence regions.

To validate and discriminate the genomic origin of XM_871257.2 expression, PCRs with a primer pair (LOC618944_F1/R1, Table 1) spanning exons 2–6 according to the XM_871257.2 mRNA model sequence were performed with liver cDNA and genomic DNA from the same individual. We received a single fragment from each template comprising 624 bp in liver cDNA and 623 bp in genomic DNA, respectively (Fig. 3A). The results of direct sequencing of these PCR products presented in Table 2 indicate the nucleotides discriminating both loci on BTA23 and BTA26 assemblies in comparison to the respective nucleotides obtained experimentally by sequencing the cDNA and the genomic fragment. A clear assignment of the generated PCR fragments to sequence contig NW_001494189.2 on BTA23 (liver cDNA template) or NW_001494358.2 on BTA26 (genomic DNA

template) is evident. The sequence length obtained in liver cDNA is 71 bp longer than predicted by the XM_871257.2 mRNA model (Fig. 2A) elongating the predicted exon 3 of the XM_871257.2 model in its 5′ orientation. These 71 nucleotides fit perfectly in the sequence contig NW_001494189.2 on BTA23. They are also identical with the 71 nucleotides interrupting the XM_871257.2 mRNA alignment on the current BTA26 assembly. Taken together, we can conclude that the mRNA expression pattern observed in cDNA from bovine liver, skeletal muscle, subcutaneous fat, and mammary gland should be due to the sequence assigned to BTA23, while the BTA26 sequence is most likely not expressed in the tissues investigated.

The qualitative RT–PCR analysis with primer pair LOC618944_F1/R1 spanning exons 2–6 showed expression in cDNA from liver, skeletal muscle, mammary gland and subcutaneous fat. Besides the dominant 624 bp PCR fragment expressed in all analyzed tissues, additional fragments could be detected in cDNAs of all tissues except for liver (Fig. 3A) indicating alternative splice variants for *LOC618944*. To verify this hypothesis and to identify the transcription start site of the *LOC618944*, a 5′ RACE experiment with skeletal muscle cDNA and a gene-specific primer (LOC618944_RACE1) located in exon 6 (Table 1; Fig. 2B) was performed. We obtained two different sequences, which both share an extension of exon 2 by 232 bp in the upstream direction as illustrated in the revised new mRNA model (Fig. 2B, C1–C3). The seven nucleotides of the predicted first exon from the previous mRNA model XM_871257.2 (Fig. 2A) are integrated and together with the newly determined 232 nucleotides form the new exon 1 in the revised model (Fig. 2B, C1–C3). Sequence similarity search with the 232 nucleotides identified by 5′ RACE revealed 99% identity to the equivalent segment on BTA23. According to these experimental data, we adapted the designation of exons in the revised gene model version of the locus (GenBank accession no. GU183135) and will refer to these terms consecutively in the description of our results.

However, the 232 nucleotides identified by 5′ RACE are also similar to the BTA26 sequence by 92%, except for the first 60 nucleotides, suggesting again a continuous alignment of the revised mRNA for the locus *LOC618944* on BTA26 (Fig. 1).

One of the novel sequence variants (GenBank accession no. GU183136), we obtained in the 5′ RACE experiment, carried an additional exon, which is now annotated as exon 3 in the revised gene model of the locus. The respective 105 bp of that additional exon fit accurately in the sequence contig NW_001494189.2 on BTA23. No sequence similar to these 105 nucleotides was identified on BTA26. Subsequent experiments with primers (LOC618944_F1/R1) located in exons 1–6
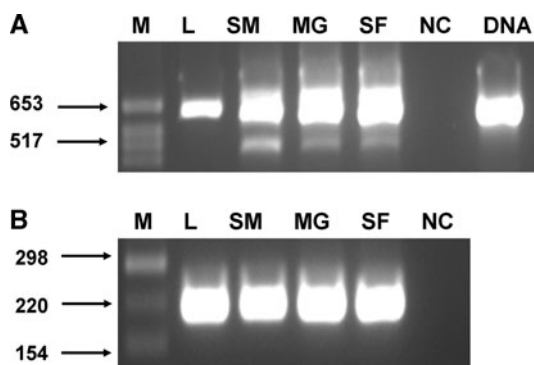
Fig. 3 Electrophoretic expression patterns of the bovine gene loci *LOC618944* (**A**) and *IRAK1* (**B**) in different tissues of a female animal from a F_2 Charolais × Holstein cross. Gene expression was analyzed by RT–PCR using primers spanning the, exons 1–6 for *LOC618944* (624 bp in liver) and exons 7–9 for *IRAK1* (217 bp). *L* liver, *SM* skeletal muscle, *MG* mammary gland, *SF* subcutaneous fat, *NC* negative control, *M* molecular size marker pBR 328 digested with *Hin*fI, *DNA* PCR fragment amplified with genomic DNA and *LOC618944* primers (623 bp; LOC618944_F1/R1) within the same animal

Table 2 Discrimination of similar sequence segments on BTA23 and BTA26 by sequence comparison of the respective PCR fragments amplified from genomic and liver cDNA using the same primer pair

Position in reference sequence[a]	Nucleotide in			
	NW_001494189.2 BTA23	NW_001494358.2 BTA26	Resequenced liver cDNA	Resequenced genomic DNA
84	A	G	A	G
91	G	A	G	A
106	G	C	G	G/C
112	A	G	A	G
140	G	A	G	A
173	T	G	T	G
231	G	T	G	G
241	C	G	C	G
253	A	G	A	G
264	C	Absent	C	Absent
269	A	C	A	C
309	A	C	A	C
345	A	G	A	G
378	G	A	G	A
399	G	A	G	A
419	T	A	T	A
520	A	T	A	T
522	C	T	C	T
543	G	A	G	A
546	A	G	A	G

[a] Positions refer to predicted mRNA sequence XM_871257.1, NW_001494189.2, NW_001494358.2: genomic sequence contigs on BTA23 and BTA26, respectively, according to the bovine genome assembly Btau4.0

according to the revised gene model confirmed the two different transcript variants (Fig. 2C1, C2; GenBank accession nos. GU183135, GU183136) revealed by the 5′ RACE experiment and identified an additional transcript variant (Fig. 2C3). This third variant contains the 105 bp long exon 3 but lacks the 46 bp of the small exon 5 (GenBank accession no. GU183137). All transcript variants of *LOC618944* (Fig. 2C1–C3) represent noncoding RNAs, because no continuous ORF could be detected for any of these sequences.

Comparative analysis of *LOC618944*

The nucleotide sequence of the predicted mRNA XM_871257.2 and also our revised gene model for the *LOC618944* locus have only a moderate similarity of 77% to the 152 amino acids encoding part of human mRNA (NM_001145020.1) of the *C6orf52* gene on HSA6 (human chromosome 6). However, the overall sequence-wide coverage accounted for 54% only. In addition, the nucleotide database search identified sequence similarities of about 77% to mRNA sequences from chimpanzee (XM_518864.2), rhesus monkey (XM_001090396.1), and pig (NM_0011145022.1) (overall sequence query coverage 35, 27 and 37%), respectively.

Although there is only a limited sequence homology of the bovine locus *LOC618944* to the human *C6orf52* mRNA, the organization of that locus appears quite similar in both species. For the human *C6orf52* locus two non-coding (NR_026736.1 and NR_026737.1) and one coding (NM_001145020.1) transcript variants are described. In accordance with the bovine transcript, the longest non-coding human variant consists of six exons. The second noncoding human variant lacks the exons 3 and 5, which are also the variable ones in cattle. The human transcript NM_001145020.1 is encoding because of an 82 bp long alternate exon containing the translational start codon. No similarity between this alternate human exon and any bovine genomic sequence on BTA23 and BTA26, respectively, or elsewhere in the bovine genome assembly was deciphered. In contrast to the highly similar sequence identified on BTA26, no additional sequence for *C6orf52* could be detected in the human genome.

Full-length cDNA and alternative transcripts for *IRAK1*

The target sequence of the Affymetrix probe set Bt.5221.1.S1_at is a part of a cDNA clone (BC108132.1), comprising 3,179 bp and containing the major part of the coding sequence of the bovine *IRAK1* gene. The nucleotides
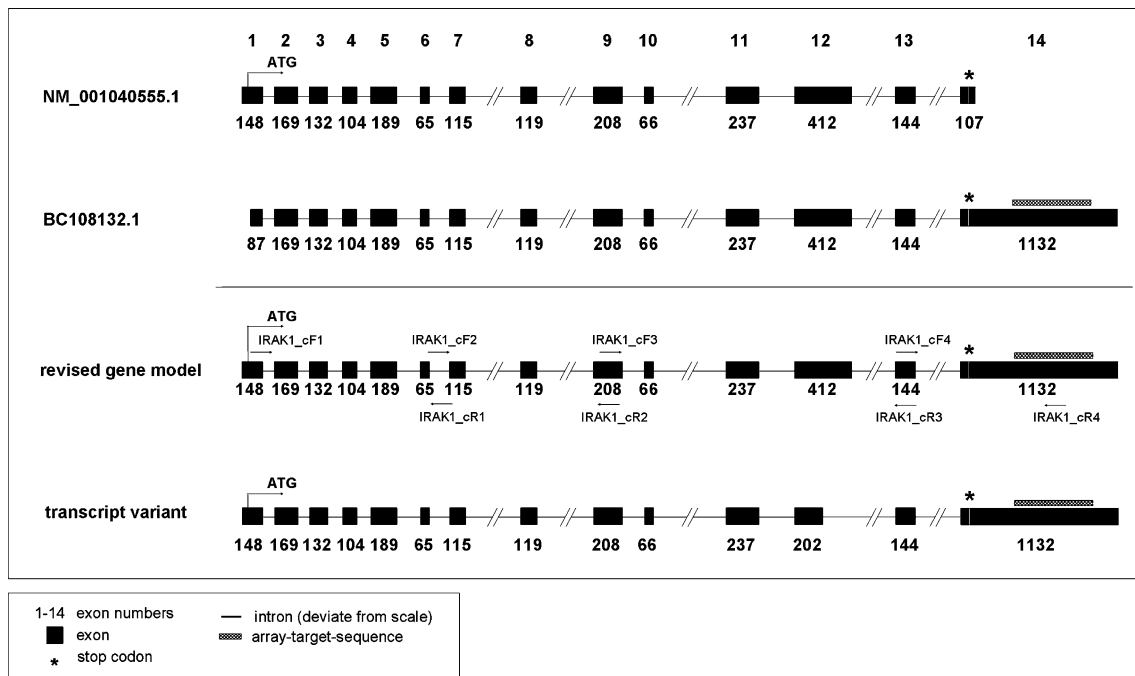
Fig. 4 Revision of the structure of the bovine *IRAK1* gene based on partial bovine mRNA sequences NM_001040555.1 and BC108132.1 and a transcript variant generated by alternative splicing. Positions of primers (IRAK1_cF1, IRAK1_cF2, IRAK1_cF3, IRAK1_cF4, IRAK1_cR1, IRAK1_cR2, IRAK1_cR3, IRAK1_cR4) are indicated on the revised model

62–2,215 of the currently annotated 2,215 bp long *IRAK1* mRNA sequence NM_001040555.1 [9] are identical to the partial cDNA sequence BC108132.1, but lack homology to the Affymetrix probe set Bt.5221.1.S1_at. Thus, we assumed that merging both sequences would result in a downstream extension of the NM_00104555.1 sequence by 1,025 bp towards its 3′ UTR and encompass the array-target sequence. This hypothesis was confirmed by RT–PCR in liver, skeletal muscle, mammary gland and subcutaneous fat tissue with primers (IRAK1_cF4/cR4, Table 1) derived from the most downstream region of the NM_001040555.1 sequence and the region of the sequence BC108132.1 encompassing the target sequence on the Affymetrix array. This revised *IRAK1* full-length mRNA model comprises 14 exons and a total length of 3,240 bp (Fig. 4).

The expression of the bovine *IRAK1* gene in liver, muscle, mammary gland and subcutaneous fat was confirmed by RT–PCR products of the expected size (217 bp; Fig. 3B) with primers spanning exons 7–9 (IRAK1_F1/R1, Table 1). While three different alternatively spliced transcript variants are known for human *IRAK1* [14–16], for the orthologous bovine *IRAK1* gene no splice variants are deposited in publicly accessible data bases. RT-PCRs using four primer pair combinations (IRAK1_cF1/cR1, IRAK1_cF2/cR2, IRAK1_cF3/cR3 and IRAK1_cF4/cR4, Table 1) that generated amplification products covering the entire revised *IRAK1* gene model did not only confirm the full-length cDNA sequence suggested by our new gene model (Genbank accession no. GU183133), but also identified an alternative mRNA splice variant in skeletal muscle (Fig. 5). In this transcript variant (Genbank accession no. GU183134), 210 bp are skipped at the downstream end of exon 12 (Fig. 4) encoding for 70 amino acids, which would be lacking in the respective IRAK1 protein. Analysis of expression pattern of this splice variant in cDNA of a variety of 19 different bovine tissues showed that this transcript is found primarily in skeletal muscle but at a relatively low level when compared to the dominant long transcript variant (Fig. 5).

Discussion

Although the main objective of the bovine genome project, the identification and annotation of the whole bovine genome sequence, is almost achieved, there still remain several regions in the current version of the bovine genome assembly (Btau4.0), which are not sufficiently annotated and functionally characterized. Thus, comprehensive and accurate mapping of the complete repertoire of the transcriptional activity across the bovine genome will remain a major future challenge [17]. Refining the bovine genome annotation will require a final structural and functional validation of full-length cDNAs and computationally predicted gene models, the detection of rare transcripts and alternative mRNA splice variants and increasingly, the
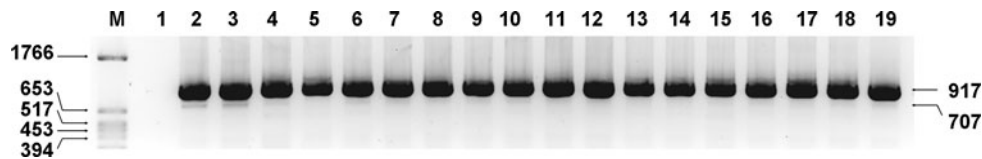
Fig. 5 Detection of an alternative splice variant in the bovine *IRAK1* gene in skeletal muscle. RT–PCR was performed using primers spanning the exons 9–13 of *IRAK1* in tissues from a female animal (F$_2$ Charolais × Holstein cross). Main amplicon: 917 bp, splice variant: 707 bp. *M* molecular size marker pBR 328 digested with *Hin*fI. DNA, *lanes 1–19*: negative control, M. semitendinosus, M. longissimus dorsi, heart, intestine fat, subcutaneous fat, liver, mammary gland, kidney, adrenal gland, pituitary gland, brain, thyroid gland, duodenum, jejunum, colon, spleen, lymphatic node, rumen

identification and characterization of noncoding RNAs. This approach will rely on manual inspection of known gene or transcript sequences and of in silico derived information as well as on generation of new experimental data. Knowledge about specific structural and functional details of the bovine transcriptome is increasingly required, since high-throughput genomic research tools, e.g. expression microarrays, have become standard in bovine genomics research. After the identification of divergently expressed oligo probe sets on the microarrays, the corresponding cognate transcripts, genes and proteins need to be experimentally validated and further characterized as an essential framework for the elucidation of the biological significance of the initial results.

Therefore, we validated and refined computational gene predictions, defined full-length transcripts, screened for mRNA splice variants and analyzed tissue-specific expression patterns of two bovine transcripts from the Affymetrix GeneChip Bovine Genome Array, which were previously assigned to the UniGene clusters Bt.5221 and Bt.10616 representing the bovine *IRAK1* gene and a bovine noncoding RNA (*LOC618944*) similar to human *C6orf52* gene.

Structural annotation and putative
function of *LOC618944*

Our data provide evidence for a revised *LOC618944* gene model and confirm the existence of the respective transcripts in the bovine transcriptome as has been reported by Smith et al. [18]. The assignment of *LOC618944* to BTA23 by in silico sequence similarity analysis on the current bovine genome assembly is supported by comparative similarity analysis between the bovine and human genome assemblies. The syntenic chromosomal region on HSA6 carrying the orthologous human *C6orf52* gene (Fig. 1) and immediately flanking genes *GCNT6* and *PAK1IP1* is completely maintained on the corresponding region on BTA23 carrying the locus *LOC618944*. Hence, the re-entry of this locus on the BTA23 architecture should be considered in upcoming refined versions of the bovine genome sequence assembly.

In addition, our results provide evidence for several splice variants existing at the *LOC618944* locus. A dominant transcript comprising five exons (exons 1–2 and 4–6, Fig. 2C2) was present in all tissues investigated. In bovine skeletal muscle, we experimentally confirmed two additional splice variants. One of these transcript variants essentially corresponds to the dominant transcript, but contains the additional exon 3 (Fig. 2C1). The second variant lacks the small exon 5 (Fig. 2C3) compared to the longest transcript (Fig. 2C1). RNA sequence entries for *LOC618944* NCBI data base, which are supported by EST evidence from construction of the Cattle Gene Index [18] are in agreement with our results for the most parts.

The predicted *LOC618944* transcript NR_026724 is very similar to our revised full gene model and the longest transcript (GU183135) found in our analysis in bovine skeletal muscle, liver, mammary gland, and subcutaneous fat tissue (Fig. 2C1), but is incomplete towards the 5′ region of the revised model. Compared to the NR_026724 transcript, results from our 5′ RACE experiment provided evidence for additional seven nucleotides in the 5′ upstream region in all transcript variants, which provide more precise data on the transcription start site of the *LOC618944* locus. Smith et al. [18] predicted two further noncoding RNAs for this locus, NR_026725 and NR_026726. Whereas the sequence NR_026725 is almost identical with our transcript variant 2 (GU183136, dominant transcript without exon 3, Fig. 2C2), we have no experimental evidence for the NR_026726 variant (Fig. 2D), which lacks the exons 3–5. In contrast, no data base prediction information does exist in the bovine sequence resources for the transcript variant lacking only the small exon 5 (GU183137, Fig. 2C3) detected in our study.

The biological function of the bovine locus *LOC618944* can not be inferred from comparative gene information because the principal molecular function of the *C6orf52* gene is still unknown, even for the orthologous human gene. The porcine mRNA NM_001145022.1 denoted as similar to the *C6orf52* gene encoding a putative protein of 158 amino acids was derived from three ESTs, one of them was detected at early stages of embryogenesis [19]. However, the detection of this transcript at early developmental stages presumably would suggest a function of the related gene or protein during the embryogenesis in pig or possibly also in other mammals.

Our data and also those from Smith et al. [18] did not provide evidence for bovine transcripts that could serve as a protein-coding template for translation. This is in contrast to the orthologous human locus, where besides two non-coding transcript variants, also a coding transcript is known. However, currently, we cannot exclude the existence of a respective coding transcript in the bovine genome, because we did not perform a complete screening at different developmental stages in a comprehensive tissue panel.

Based on the experimental data available from our study for the *LOC618944* locus assigned to BTA23, two evolutionary scenarios are conceivable. The identified noncoding transcript variants assigned to the corresponding chromosomal region may have arisen as a result of mutation events at a yet unknown functional parent gene, which have caused their non-protein functionality. Thus, these non-coding transcripts may represent variants proceeding towards pseudogene development. Alternatively, it is conceivable that the identified noncoding transcript variants may be due to reactivation of a gene silenced earlier in evolution. Mechanisms known for pseudogene generation [20–22] and the structural features of the sequence on BTA26, which is highly similar to *LOC618944* and let us assume that possibly, this sequence may represent a processed pseudogene derived from a parent mRNA by retrotransposition. The absence of putative introns in the aligned sequence and a missing polyA stretch at the 3′ end in the adjacent chromosomal region aligned to the target sequence on BTA26 support this hypothesis for the molecular evolution of this locus.

Our results provide evidence that the noncoding transcript variants assigned to BTA23 are obviously responsible for the expression signals observed in liver, skeletal muscle, mammary gland and subcutaneous fat tissues. Concomitantly with previous results indicating divergent expression levels in skeletal muscle from animals differing in their intramuscular fat content, this has encouraged us to postulate a potential role of these transcript variants in energy metabolism of cattle. Considering the fact that the highly similar porcine protein-coding transcript variant was detected at embryonic stage of development, it would also be conceivable that a coding transcript variant might possibly exist only at early developmental stages and that presumably, the translation of the gene might be silenced at later stages of development due to the appearance of splice variants. This could be the possible reason that we could not find a respective bovine protein-coding transcript in our study. However, this speculative hypothesis has to be confirmed in prospective experiments. On the other hand, a functional relevance of the observed noncoding transcripts could not be excluded with regard to the increasing number of reports indicating essential roles of noncoding RNAs in tissue-specific and developmental regulation of gene expression in a variety of cells and organisms [23–25]. Gerstein et al. [26] reported that the ENCODE project [27] provided evidence that there is also substantial transcriptional activity in the intergenic space in the human genome due to non-protein-coding RNAs and pseudogenes. The authors found that frequently numerous transcribed pseudogenes and noncoding RNA genes are located within introns of protein-coding genes. These transcripts may possibly influence the expression of their host genes, or they are supposed to be important for particular processes as chromatin accessibility for transcription factor binding.

Indication on a functional role of the *IRAK1* gene in energy metabolism

The bovine *IRAK1* gene is located in the pseudoautosomal region of the X chromosome analogously to the human orthologous gene. In our study we refined the current mRNA (NM_001040555.1) sequence of the bovine *IRAK1* gene by extending it by 1,025 bp towards its 3′ end and thus, could provide a full-length cDNA sequence for the gene.

IRAK1 is known as a protein relevant in immune response [10]. Results of expression analysis in our study in cattle have extended the functional relevance of IRAK1 towards a variety of tissues including bovine liver, skeletal muscle, mammary gland and subcutaneous fat, which are involved in energy and fat metabolism in cattle. Moreover, in an initial expression screening experiment *IRAK1* emerged as differently expressed in animals with divergent intramuscular fat content. Analogously to cattle, in human, the expression of *IRAK1* could be demonstrated for a wide range of cell types [14] supporting our assumption that IRAK1 could have functions exceeding the regulation of immune defense. An additional indication for a putative function of IRAK1 in energy and fat metabolism is provided by the increasing knowledge about relationships between lipid metabolism and immune system [28–30].

For the human *IRAK1* gene, three different alternatively spliced transcript variants are known [14–16]. The human variants 1 and 2 also contain 14 exons as we found for the bovine *IRAK1* mRNA. Compared to transcript variant 1, the variant 2 lacks an in-frame segment at the 5′ site of exon 12, while in the human alternative transcript 3, the entire exon 11, is skipped. Our data provide evidence for a splice variant of the bovine *IRAK1* gene primarily observed in skeletal muscle. This transcript variant is identical to the revised bovine *IRAK1* cDNA sequence except for an in-frame deleted segment of 210 bp of exon 12. In contrast to the human transcript variant 2, the deletion is located at the downstream end of the bovine exon 12.

On protein level, the skipped 70 amino acids affect two regions of intrinsic disorder and a low complexity domain. Intrinsic disorder domains are segments of proteins that lack a well-structured three-dimensional folding, and their functions can be grouped into four broad categories: molecular recognition, molecular assembly/disassembly, protein modification, and entropic chain activities [31]. Numerous studies of protein structure reported that the presence of intrinsically disordered segments has a substantial influence on the function of the proteins, particularly of proteins involved in signaling and regulation processes [32, 33]. Analyzing sequence-function relationships, Dunker et al. [34] reported that signaling sequences and sites of posttranslational modifications are frequently located within regions of intrinsic disorder and that the flexibility of intrinsically disordered segments is important to ensure the capacity for the binding diversity in protein–protein interaction networks. The authors also found that in multicellular eukaryotes intrinsic disorder-based signaling is modulated by alternative splicing and that splicing events more often map to regions of intrinsic disorder than to other structural regions. Thus, they concluded that association of alternative splicing with protein disorder region in proteins can enable spatio-temporal and tissue-specific modulation of protein function needed for cell differentiation and the evolution of multicellular organisms. Trying to interpret a possible consequence of the deleted transcript variant identified in bovine skeletal muscle by considering this knowledge, we postulate that presumably, the omission of the respective 70 amino acids may exert putative effects on the structural and associated biological characteristics of the corresponding protein variant and hence, the regulation of function of the bovine *IRAK1* gene in skeletal muscle. However, this idea merits to be elucidated in detailed analyses of further studies.

For a profound transcript or gene analysis in deficiently annotated regions of the bovine genome, comparative structural and functional sequence information from sequence-ready genomes, like human, is a beneficial source. In turn, results obtained by detailed elucidation of orthologous genes in other species, like bovine, also may help to decipher yet unknown functions for known or unknown transcripts and thus, may contribute to a better understanding of their principal physiological role in the organisms.

References

1. Andersson L, Georges M (2004) Domestic-animal genomics: deciphering the genetics of complex traits. Nat Rev Genet 5:202–212

2. De Smet S, Raes K, Demeyer D (2004) Meat fatty acid composition as affected by fatness and genetic factors: a review. Anim Res 53:81–92

3. Abe T, Saburi J, Hasebe H, Nakagawa T, Kawamura T, Saito K, Nade T, Misumi S, Okumura T, Kuchida K, Hayashi T, Nakane S, Mitsuhasi T, Nirasawa K, Sugimoto Y, Kobayashi E (2008) Bovine quantitative trait loci analysis for growth, carcass, and meat quality traits in an F-2 population from a cross between Japanese Black and Limousin. J Anim Sci 86:2821–2832

4. Thaller G, Kuhn C, Winter A, Ewald G, Bellmann O, Wegner J, Zuhlke H, Fries R (2003) DGAT1, a new positional and functional candidate gene for intramuscular fat deposition in cattle. Anim Genet 34:354–357

5. Barendse W, Bunch RJ, Kijas JW, Thomas MB (2007) The effect of genetic variation of the retinoic acid receptor-related orphan receptor C gene on fatness in cattle. Genetics 175:843–853

6. Barendse W, Bunch RJ, Thomas MB, Harrison BE (2009) A splice site single nucleotide polymorphism of the fatty acid binding protein 4 gene appears to be associated with intramuscular fat deposition in longissimus muscle in Australian cattle. Anim Genet 40:770–773

7. Fortes MRS, Curi RA, Chardulo LAL, Silveira AC, Assumpcao MEOD, Visintin JA, de Oliveira HN (2009) Bovine gene polymorphisms related to fat deposition and meat tenderness. Genet Mol Biol 32:75–82

8. Kuehn C, Bellmann O, Voigt J, Wegner J, Guiard V, Ender K (2002) An experimental approach for studying the genetic and physiological background of nutrient transformation in cattle with respect to nutrient secretion and accretion type. Arch Anim Breed 45:317–330

9. Connor EE, Cates EA, Williams JL, Bannerman DD (2006) Cloning and radiation hybrid mapping of bovine toll-like receptor-4 (TLR-4) signaling molecules. Vet Immunol Immunopathol 112:302–308

10. Gottipati S, Rao NL, Fung-Leung WP (2008) IRAK1: a critical signaling mediator of innate immunity. Cell Signal 20:269–276

11. Elsik CG, Tellam RL, Worley KC, Bovine Genome Sequencing Anal Cons (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science 324:522–528

12. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

13. Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–214

14. Cao ZD, Henzel WJ, Gao XO (1996) IRAK: a kinase associated with the interleukin-1 receptor. Science 271:1128–1131

15. Jensen LE, Whitehead AS (2001) IRAK1b, a novel alternative splice variant of interleukin-1 receptor-associated kinase (IRAK), mediates interleukin-1 signaling and has prolonged stability. J Biol Chem 276:29037–29044

16. Rao N, Nguyen S, Ngo K, Fung-Leung WP (2005) A novel splice variant of interleukin-1 receptor (IL-1R)-associated kinase 1 plays a negative regulatory role in toll/IL-1R-induced inflammatory signaling. Mol Cell Biol 25:6521–6532

17. Tellam RL, Lemay DG, Van Tassell CP, Lewin HA, Worley KC, Elsik CG (2009) Unlocking the bovine genome. BMC Genomics 10:193

18. Smith TPL, Grosse WM, Freking BA, Roberts AJ, Stone RT, Casas E, Wray JE, White J, Cho J, Fahrenkrug SC, Bennett GL,

Heaton MP, Laegreid WW, Rohrer GA, Chitko-McKown CG, Pertea G, Holt I, Karamycheva S, Liang F, Quackenbush J, Keele JW (2001) Sequence evaluation of four pooled-tissue normalized bovine cDNA libraries and construction of a gene index for cattle. Genome Res 11:626–630

19. Whitworth K, Springer GK, Forrester LJ, Spollen WG, Ries J, Lamberson WR, Bivens N, Murphy CN, Mathialigan N, Green JA, Prather RS (2004) Developmental expression of 2489 gene clusters during pig embryogenesis: an expressed sequence tag project. Biol Reprod 71:1230–1243

20. Vanin EF (1985) Processed pseudogenes—characteristics and evolution. Annu Rev Genet 19:253–272

21. Torrents D, Suyama M, Zdobnov E, Bork P (2003) A genome-wide survey of human pseudogenes. Genome Res 13:2559–2567

22. Gerstein M, Zheng DY (2006) The real life of pseudogenes. Sci Am 295:48–55

23. Amaral PP, Mattick JS (2008) Noncoding RNA in development. Mamm Genome 19:454–492

24. Mattick JS, Amaral PP, Dinger ME, Mercer TR, Mehler MF (2009) RNA regulation of epigenetic processes. Bioessays 31:51–59

25. Mattick JS (2009) The genetic signatures of noncoding RNAs. PLoS Genet 5:e1000459

26. Gerstein MB, Bruce C, Rozowsky JS, Zheng DY, Du J, Korbel JO, Emanuelsson O, Zhang ZDD, Weissman S, Snyder M (2007) What is a gene, post-ENCODE? History and updated definition. Genome Res 17:669–681

27. Zheng DY, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu YT, Denoeud F, Antonarakis SE, Snyder M, Ruan YJ, Wei CL, Gingeras TR, Guigo R, Harrow J, Gerstein MB (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res 17:839–851

28. Cabral GA (2005) Lipids as bioeffectors in the immune system. Life Sci 77:1699–1710

29. Yaqoob P (2004) Fatty acids and the immune system: from basic science to clinical applications. Proc Nutr Soc 63:89–104

30. Moyes KM, Drackley JK, Morin DE, Bionaz M, Rodriguez-Zas SL, Everts RE, Lewin HA, Loor JJ (2009) Gene network and pathway analysis of bovine mammary tissue challenged with streptococcus uberis reveals induction of cell proliferation and inhibition of PPAR-gamma signaling as potential mechanism for the negative relationships between immune response and lipid metabolism. BMC Genomics 10:542

31. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z (2002) Intrinsic disorder and protein function. Biochemistry 41:6573–6582

32. Dunker AK, Obradovic Z (2001) The protein trinity—linking function and disorder. Nat Biotechnol 19:805–806

33. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6:197–208

34. Dunker AK, Oldfield CJ, Meng JW, Romero P, Yang JY, Chen JW, Vacic V, Obradovic Z, Uversky VN (2007) The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics 9:S1