

Identification of genetic variation and putative regulatory regions in bovine *CARD15*

Kristen H. Taylor,^{1*} Jeremy F. Taylor,² Stephen N. White,³ James E. Womack¹

¹Department of Veterinary Pathobiology, Texas A&M University, College Station, Texas 77843-4467, USA

²Division of Animal Sciences, University of Missouri, Columbia, Missouri 65211, USA

³Animal Disease Biotechnology Facility, Washington State University, Pullman, Washington 99164, USA

Received: 27 October 2005 / Accepted: 5 April 2006

Abstract

Mutations in *caspase recruitment domain 15* (*CARD15*) are associated with susceptibility to Crohn's disease and Blau Syndrome. We performed comparative analyses of the bovine, murine, and human *CARD15* transcripts to elucidate functionality of bovine *CARD15* and examine its potential role in bovine disease resistance. Comparative analyses of intronic sequence across seven divergent species were performed to identify putative regulatory element binding motifs. High levels of interspecies conservation in sequence, genomic structure, and protein domains were detected indicating common functionality for *CARD15* in cattle, human, and mouse. We identified species-specific regulatory elements in the 5' and 3' untranslated regions, suggesting that modes of regulation may have diverged across species. Thirty-one conserved putative regulatory element binding motifs were identified in the *CARD15* intronic sequence of seven species. To assess the extent of genetic diversity within bovine *CARD15*, 41 individuals from two subspecies were sequenced and screened for polymorphisms. Thirty-six single nucleotide polymorphisms (SNPs) were identified. Finally, 20 subspecies-specific haplotypes were predicted with 7 and 13 unique haplotypes explaining the diversity within *B. taurus taurus* and *B. taurus indicus* animals, respectively. Strong evidence for a simple causal relationship between these SNP loci and their haplotypes with Johne's disease was not detected.

Introduction

Caspase recruitment domain 15 (*CARD15*), first denoted *NOD2*, is an intracellular receptor of pathogen components. Its structure is similar to plant disease resistance genes and comprises an effector domain region (2 caspase recruitment domains), a centrally located nucleotide binding/oligomerization domain (NOD), and carboxy-terminal leucine-rich repeats (LRR). Mutant forms of *CARD15* lacking the LRR have an enhanced ability to activate *nuclear factor kappa-B* (*NF-kB*) (Miceli-Richard et al. 2001), while specific mutations in human *CARD15* have been shown to be associated with Crohn's disease (Hugot et al. 2001; Ogura et al. 2001a) and Blau Syndrome (Miceli-Richard et al. 2001), two distinct granulomatous disorders. Mutations in the human and mouse LRR inhibit recognition or binding ability to bacterial components resulting in a perturbation of *NF-kB* activation (Girardin et al. 2003; Gutierrez et al. 2002; Inohara and Nunez 2001; Ogura et al. 2001b, 2003) which is coupled with the formation of granulomas.

Granulomatous disorders have detrimental consequences to animal health, reduce longevity and productivity, and cost the livestock industry millions of dollars each year. However, a far more significant consequence is the contamination of food products for human consumption and ensuing zoonoses. In particular, cattle are affected by the granulomatous disorder, Johne's disease, which is believed by some to have a similar etiology to Crohn's disease (Chiodini 1989; Cocito et al. 1994; Thompson 1994). The belief that the two diseases are related and the previously discovered associations between mutations in *CARD15* and Crohn's disease make *CARD15* an interesting candidate gene for susceptibility to Johne's disease. To facilitate association studies between bovine *CARD15* and disease resistance, we report the genomic localiza-

*Present address: Department of Pathology/Anatomical Science, University of Missouri, Columbia, Missouri 65212, USA
 Correspondence to: James E. Womack; E-mail: jwomack@cvm.tamu.edu

tion and the naturally occurring variation in the gene within the United States cattle population. We also provide a detailed comparative analysis of *CARD15* intronic and exonic sequence and have identified putative *CARD15* regulatory regions.

Materials and methods

Cloning of bovine *CARD15*. Three partial bovine cDNAs (BF605150, BM032079, and BF601658) with homology to human *CARD15* were obtained from the Children's Hospital Oakland Research Institute. Universal M13 primers were used to sequence these clones and sequence data were obtained for regions with homology to exons 4–12 of the human *CARD15* gene. To acquire sequence data from the 5' and 3' regions of the bovine gene, rapid amplification of cDNA ends (RACE) was performed using the FirstChoice RLM-RACE kit (Ambion, Austin, TX) and *Bos taurus taurus* (*B. taurus taurus*) mRNA extracted from peripheral blood leukocytes as a template. Gene-specific nested primers were designed for the 5' (outer: 5'-AAG CCC TTG AGG CTG AGT TC-3' and inner: 5'-GAA GAG CAG ACT CTG GAC TGA CG-3') and 3' (outer: 5'-CGG CAG AAA CAC AGA TGA GA-3' and inner: 5'-CCC AGC GTG GAG TTG TAA GT-3') ends of bovine *CARD15*. These gene-specific primers and the RACE primers supplied with the kit were used to amplify the 5' and 3' ends of the bovine *CARD15* gene. Exon order in the coding sequence was confirmed by the sequence analysis of a population of cDNAs produced using a bovine *CARD15* gene-specific primer. Overlapping amplicons were generated from this cDNA population using three sets of primer pairs (Amplicon 1, F: 5'-GGC TTG GAG CTC TGT GAG AT-3' and R: 5'-GCA GCT AAA TGG GAA GAC GA-3'; Amplicon 2, F: 5'-GTC CAA GCT GAG GAC CGT TA-3' and R: 5'-ATG CAG CAA AGA AGC ACT GA-3'; Amplicon 3, F: 5'-TGC TGC TAC GTG TTC TCA GC-3' and R: 5'-AAC ACT GGC CTG GAA ACA TC-3'). The amplified cDNA fragments were cloned into plasmid vectors and their complete sequence was determined.

To capture intronic sequence flanking each exon, primers were developed from the sequence for each exon of bovine *CARD15*, one extending in the 5' direction and the other extending in the 3' direction. The primer located in the 3' region of the first exon was paired with the 5' primer located in second exon to amplify the first intron, and so on. The produced amplicons were then partially sequenced from both directions to confirm their identity and to obtain the sequence flanking each exon. This flanking sequence was next used to develop primer pairs to amplify each

exon and the flanking intronic sequence (Supplementary Table 1). The exon-intron organization of bovine *CARD15* was then determined from the sequence analysis of the amplicons produced using these 14 primer pairs from genomic DNA templates of *B. taurus taurus* and *Bos taurus indicus* (*B. taurus indicus*) individuals. Finally, the obtained genomic sequence was aligned with the cDNA sequence to determine exon-intron boundaries. The complete *B. taurus taurus* coding sequence has been submitted to GenBank as accession number AY518737. The exon and flanking sequence data have been submitted to GenBank as accession numbers AY518738-AY518748 and AY518749-AY518759 for *B. taurus taurus* and *B. taurus indicus*, respectively.

Genomic localization of bovine *CARD15*.

Primers 5'-GCA CAA CCT CCA GAT CAC AG-3' and 5'-GAC ACC GCT GGA CAC AA TC-3' were developed from bovine *CARD15* (GenBank accession No. AY518737), the PCR product was sequenced, and homology to bovine *CARD15* was confirmed. An annealing temperature of 58°C produced amplification of a cattle-specific PCR product in a mouse-hamster background. PCR was performed in 31 somatic and 90 selected radiation hybrid cell lines from a cattle-hamster somatic hybrid cell panel and a 5000-rad whole-genome radiation hybrid (WGRH₅₀₀₀) panel, respectively (Womack and Moll 1986; Womack et al. 1997). Syntenic assignment was made from correlations of marker retention among lines (Chevalet and Corpet 1986). All WGRH₅₀₀₀ typing experiments were performed twice and were independently scored, and data concordant in both experiments were used for RH mapping. Two-point linkage analysis using RHM-APPER (Slonim et al. 1997) was used to assign *CARD15* to a chromosomal region.

Bovine, human, and murine *CARD15* alignments. All alignments were constructed using AlignX in the Vector NTI suite (Informax, Frederick, MD). Alignments for the comparative analysis used published sequence data available for bovine, human, and murine *CARD15* (GenBank Accession Nos. AY518737, AF178930, AF520774). Nucleotide and amino acid identities were computed using pairwise alignments. To identify protein domains within the bovine gene, bovine sequence was analyzed using AnDom (<http://www.bork.embl-heidelberg.de/AnDom>) and ProDom (<http://www.toulouse.inra.fr/prodom.html>). In addition, bovine *CARD15* was aligned to the identified human and murine domains (Iwanaga et al. 2003; Ogura et al. 2001a, 2003) to examine domain conservation.

Identification of regulatory motifs. The human, mouse, and cow sequences were searched for regulatory motifs in the 5' and 3' untranslated regions (UTRs) using UTRscan (<http://www.ba.itb.cnr.it/BIG/UTRScan/>). To identify putative regulatory elements present in intronic sequence, intronic *CARD15* sequences from human, mouse, cow, rat, chimp, rhesus monkey, and dog were aligned using Vector NTI. We were able to determine the intronic sequence for each species by aligning the genomic sequence of each species to the coding sequence of human *CARD15* and identifying the splice acceptor and donor sites flanking each exon. These alignments successfully identified the sequence for introns 2–11 in these species. Intron 1 sizes in chimp, rhesus monkey, rat, and dog are incomplete estimates because complete *CARD15* transcripts have not been reported in these species; therefore, exon 1 could not unequivocally be determined in these species. "Motifs" consisting of at least six nucleotides conserved in all seven species and which included no more than three internal substitutions were identified and considered to be putative regulatory element binding motifs. These motifs were analyzed using NSITE (available through SoftBerry, <http://www.softberry.com/berry.phtml?topic=promoter>) to determine homology to previously identified regulatory binding motifs.

To estimate the probabilities that selectively neutral sequence motifs descended from a common ancestral sequence would be completely or partially conserved among members of a descendent evolutionary group, we used the Jukes-Cantor one-parameter model (Jukes and Cantor 1969). See the Supplementary Information for a complete description of the mathematical model used to estimate these probabilities.

Identification of polymorphic *CARD15* sites. We assembled a breed panel of 30 unrelated cattle from nine domestic cattle populations comprising eight *B. taurus indicus* individuals, including six Brahman and two Gir; and 20 *B. taurus taurus* individuals including three Angus, seven Holstein, three Texas Longhorn, two Limousin, three Jersey, and two N'Dama. Two Ankole-Watusi animals, which represent an ancient *B. taurus taurus* × *B. taurus indicus* cross, were also included in the panel. The individuals comprising this panel were not clinically tested for Johne's disease nor were they from herds with known exposure to *Mycobacterium paratuberculosis* (*M. ptb.*).

In addition to the breed panel, a case panel of 11 DNA samples from animals visually examined and clinically diagnosed with Johne's disease was

assembled. Samples from eight animals, including three unrelated Holstein calves experimentally infected with *M. ptb.* at one month of age; one Holstein embryo transfer recipient determined to have the disease after her calf was found to be infected; one Holstein with clinical Johne's that was found to be ELISA positive on repeated assay; one culture positive Holstein; and two Jerseys, one with a low-grade infection based on tissue *M. ptb.* counts and one determined Johne's positive upon necropsy, were purchased from Dr. Michael Collins (University of Wisconsin). Three samples (Holstein, Jersey, and Brahman) from animals that were both ELISA and culture positive for *M. ptb.* were provided by Dr. Allen Roussel (Texas A&M University).

The breed and case DNA panels were screened for *CARD15* polymorphisms by resequencing. Each exon along with a small amount of flanking intronic sequence was amplified in each individual with primers developed for the amplification of individual exons (Supplementary Table 1) and was sequenced in both directions. The sequences acquired for each animal were aligned using Contig Express in Vector NTI and were examined for sequence variation against each other and the consensus *B. taurus taurus CARD15* sequence (Genbank accession No. AY518737). All identified single nucleotide polymorphisms (SNPs) were validated by performing a second PCR and sequencing the produced amplicon in both directions. Sequences including the SNPs located within intronic regions and the 5' UTR and 3' UTR were analyzed using NSITE to identify putative regulatory motifs.

Haplotype estimation. SNPs with missing genotypes were removed, leaving 23 of the original 36 sites, including all but one of the coding SNPs. Haplotyper (Niu et al. 2002) was used to predict haplotypes for each individual from the genotype data. This analysis was performed using all 41 animals from the breed and Johne's case panels simultaneously.

Statistical analysis and association tests. Chi-squared tests were used to test differences in the frequencies of synonymous (sSNP) vs. nonsynonymous (nsSNP) substitutions within each subspecies and also to evaluate differences in the frequencies of each SNP among *B. taurus taurus* and *B. taurus indicus* animals. Association tests were performed within *B. taurus taurus* and within Holsteins. *B. taurus indicus* was not analyzed because only a single animal was diagnosed with Johne's disease and this animal's genotype was atypical of the *B. taurus*

indicus genotypes within the breed panel for many of the SNPs. We first tested SNP allele frequency differences between the breed panel (control group) and the Johne's disease (case group) animals. Associations between *CARD15* polymorphisms and disease status were next tested across genotype classes by examining differences in the frequencies of (1) heterozygotes, (2) homozygotes for the rare allele, and (3) the class of pooled heterozygotes and homozygotes for the rare allele. Tests were also performed to examine differences in allele frequency at each SNP between the case and control groups. These analyses were performed at four levels of stratification: by considering all SNPs, noncoding SNPs (ncSNPs), sSNPs, and nsSNPs. Finally, tests for haplotype associations with Johne's disease were performed within *B. taurus taurus* and within Holsteins alone. Haplotypes for the Texas Longhorn and N'Dama were not typical of those of the European breeds that were represented in the case group, and to avoid population stratification, these breeds were excluded from the control group in the *B. taurus taurus* analyses.

Results and discussion

Localization of bovine *CARD15*. *CARD15* was localized to BTA18 by somatic hybrid cell mapping with 97% concordancy. Integration of *CARD15* into the RH map of BTA18 revealed two alternatives for gene order with high statistical support. We established that the locus order *ADCY7-CARD15-CYLD* in cattle is conserved with human by physically mapping a BAC (data not presented). This gene order is consistent with that contained in the March 2005 bovine sequence assembly.

Comparison of human, murine, and bovine *CARD15*. The bovine *CARD15* transcript is 5105 bp and the protein it encodes comprises 1013 amino acids. This compares to human and murine *CARD15* with transcript lengths of 4485 bp and 4585 bp, respectively. Within all three species, the 12-exon genomic structure of *CARD15* is identical and the size of exons 2–11 and of the coding portion of exon 12 is completely conserved. The size of exon 1 which includes the 5' UTR, of the 3' UTR, and of the intronic regions is variable among the species. Domain analysis of bovine *CARD15* and alignment to murine and human orthologs revealed that the bovine gene also comprises two N-terminal caspase recruitment domains (residues 4–93, 96–191), one centrally located NOD (residues 243–548), and ten tandem LRR at the C-terminus (residues 715–992).

Protein alignment of the three species' sequences revealed that human and mouse have two in-frame translation initiation sites, the second of which corresponds to the unique bovine translation initiation site. A recent study by King et al. (2006), using the bovine sequence developed by our group, found a high degree of protein identity across several species including human, mouse, and cow. The consensus sequence of two of the amino acids (G908R and 1007fs), which when mutated are associated with susceptibility to Crohn's disease, are conserved between human, cow, and mouse, while that for a third (R702W) is conserved between human and cow but is variable in mouse (King et al. 2006). Furthermore, these authors also found that the consensus sequences of both amino acids, which when mutated are associated with susceptibility to Blau Syndrome, are conserved across all three species. We found overall amino acid identities of 81.2% between cattle and human, 76.0% between cattle and mouse, and 79.4% between human and mouse with slightly higher levels of conservation within the domains than in the interdomain regions. The high level of conservation of coding sequence and the conservation of protein domains point to a similar function for *CARD15* across human, mouse, and cattle.

Comparison of the human, murine, and bovine 5' UTR and 3' UTR. Because regions of conserved noncoding sequence between divergent species are likely to have a functional role, we aligned the 5' and 3' UTRs in human, mouse, and cow to identify putative regulatory motifs. We chose to include only these species because complete cDNA sequence was available and we did not want introduce error into our analysis because of our need to estimate the sequence comprising the 5' and 3' UTRs in chimp, rhesus monkey, rat, and dog. Pairwise alignments of the 5' UTR revealed only a 35.4% nucleotide identity between human and cow, 42.1% between mouse and cow, and 33.3% between human and mouse. The 3' UTR alignments also revealed low overall levels of nucleotide identity, with only small contiguous regions of homology among the three species.

Next, we examined the individual UTRs in each species for previously identified regulatory motifs. Two motifs were found in the 5' UTRs: a terminal oligopyrimidine tract in cow and mouse which is required for the coordination of translational repression, and in cow an internal ribosome entry site, or internal regulatory sequence (IRES). The IRES is believed to be involved in internal mRNA ribosome binding, which allows for translation to occur

during periods of the cell cycle when the conventional mechanism of translation is ineffective. Three 3' UTR regulatory motifs were detected in human, including an alcohol dehydrogenase 3' UTR down-regulation control element (ADH_DRE) (bases 187–194), a Brd-Box (bases 907–913), and a Gy-Box (bases 1114–1120), all of which are associated with the downregulation of gene expression. No previously documented regulatory motifs were found in the 3' UTR of mouse or cow. The presence of distinct regulatory motifs in these species may suggest that the translation of *CARD15* is regulated by divergent mechanisms in these species.

Identification of short blocks of conserved putative regulatory sequence. Comparative genomics is an effective means of identifying genes and gene regulatory sequences. Currently, assembled genomic sequence is available for vertebrates (human, chimp, rhesus, dog, cow, opossum, chicken, *X. tropicalis*, zebrafish, tetraodon, and fugu), other deuterostomes, insects, nematodes, and yeast. We searched these genomes for sequences orthologous to *CARD15*. Since our interest was in the identification of regions of conservation within intronic *CARD15* sequences, we included in our comparative analysis only those species possessing a similar *CARD15* genomic structure. This resulted in the exclusion of the nonvertebrates, *X. tropicalis* and tetraodon. Furthermore, zebrafish and fugu were excluded from this analysis because the low levels of sequence homology within the coding region of the gene made it difficult to distinguish between intronic sequence and nonhomologous coding sequence. (While the *CARD15* coding sequences for these species have GenBank accession numbers, the sequences are not yet accessible through GenBank.) We also excluded opossum because we were able to identify only a small portion of homologous *CARD15* sequence within Scaffold_15014 from which introns and exons could not accurately be determined. We excluded chicken because no sequence orthologous to *CARD15* was identified by BLAT search and a search of annotated genes for *CARD15* produced only a sequence homologous to *NOD3*. Consequently, we surveyed the introns of human, mouse, cow, rat, chimp, rhesus monkey, and dog *CARD15* orthologs for conserved regions that could potentially be regulatory element binding motifs (Borchert et al. 2003; Brend et al. 2003; Giacopelli et al. 2003; Van Laere et al. 2003). Assuming an average nucleotide mutation rate of $\alpha = 1.8 \times 10^{-8}$ per generation, t^{human} , t^{dog} , t^{cow} , t^{mouse} , t^{rat} , t^{chimp} , and t^{rhesus} of 3.75, 30, 15, 75, 75, 4.67, and 3.57 million generations, respectively, we estimate that

the probability of a selectively neutral base being completely conserved among all seven species to be 0.00725 (see Supplemental Information). Thus, we estimate the probability that a 6-bp selectively neutral "motif" is completely conserved among all seven species to be 1.46×10^{-13} and therefore the expected number of completely conserved 6-bp motifs in a 3-billion-base genome is very close to zero.

We found low levels of overall identity in the multiple alignments of intronic regions ranging from 2.1% to 27.8% (Supplementary Table 2); but the extent of identity is apparently greater than expected from our estimate of the probability of conservation (0.725%) using the Jukes-Cantor approximation. Therefore, these intronic regions may not be selectively neutral, and/or assumptions of the probability calculation such as independence among nucleotides, etc., may be violated. However, one should not overlook the implicit bias toward conservation that is inherent within the process of aligning multiple divergent sequences. The produced consensus sequences were from 1.8% to 17.1% longer than the longest sequence of the seven species (Supplementary Table 2) indicating the extent to which gapping was required to produce alignment between these intronic sequences. While the estimates of conservation may be slightly biased upward, two things seem apparent. First, the probability of conservation that we estimated assuming selective neutrality is probably too small, and second, these introns likely harbor sequences that are conserved between the species because they are under purifying selection due to their regulatory function.

While the unequivocal identification of regulatory elements is challenging because these elements are typically short (6–15 bp), they tolerate some sequence variation, and rules useful for their recognition are generally unknown, we identified several short-sequence motifs that were completely conserved between these seven species. Considering only motifs with at least 6 bp of conservation and no more than three internal substitutions between the species, we identified 31 conserved motifs ranging from 6 to 23 bp in length (Table 1). Based on our probability calculations, the likelihood of detecting conserved blocks 6–23 bp long should be extremely rare if all bases are free to evolve. Finally, we queried NSITE, a regulatory motif identification database, for these conserved sequences and found that over 75% harbored sequences with homology to previously identified regulatory binding motifs. The most extraordinary observation is the 23-bp motif identified in intron 4. This motif has homology to 23 previously identified regulatory elements and is thus

Table 1. Identity and location of conserved putative regulatory motifs

Motif ^a	<i>p</i> Value ^b	Motif start ^c	NSITE ^d
CTGACC	1.463E-13	E2 (-22)	14
CCTCCC	1.463E-13	E2 (-9)	0
AGAAGgC	1.024E-12	E3 (-1062)	0
AGGTTtA	1.024E-12	E3 (-273)	0
TTCAcTT	1.024E-12	E3 (-182)	5
CCTTCtCACA	5.598E-19	E3 (-38)	0
AAaaaGAACT	1.275E-13	E4 (-1281)	3
CTTCAGA	1.062E-15	E4 (-529)	7
TCTTCTG	1.062E-15	E4 (-399)	2
AGcCAGGA	8.499E-15	E4 (-117)	2
ACCaTGG	1.024E-12	E4 (-42)	1
GAAAGcgGAAG	3.079E-18	E4 (+1689)	0
CAGCCT	1.463E-13	E5 (-2477)	4
GTTTCCATGCCAACcGAAAcCCT	3.033E-43	E5 (-1501)	23
ATGACTT	1.062E-15	E5 (-1316)	15
CAGCaTG	1.024E-12	E5 (-1019)	5
CCaGCTcA	4.098E-12	E5 (-620)	6
CCAGtGTTCTTTAGT	1.693E-29	E5 (-256)	11
GGGTgTcCa	3.824E-14	E5 (-163)	15
TGgGGTgCTC	3.470E-16	E5 (-148)	2
TCACTG	1.463E-13	E6 (-53)	7
TaTGCTtT	4.098E-12	E6 (-43)	0
GGtTgGGGG	3.824E-14	E7 (+37)	0
CTcCTgAA	4.098E-12	E7 (+330)	5
GAACTT	1.463E-13	E8 (+82)	8
GTcATCA	1.024E-12	E11 (-441)	11
ATcTATT	1.024E-12	E11 (+481)	4
AGGGtTCTgAcT	1.232E-17	E12 (-375)	14
CcGGCCTTGaaGAGTCaG	3.453E-27	E12 (-265)	30

^aBovine motifs are presented. Upper-case letters indicate bases conserved among all seven species. Lower-case letters indicate mismatch sites.

^bFor a putative regulatory motif of *n* bases, this *p* value is the estimated probability that *p* bases ($\min(6, n - 3) \leq p \leq n$) are conserved in all seven species (see Materials and methods).

^cThe position of the first base of the motif relative to the nearest exon. For example, E2 (-22) refers to a site 22 bp upstream of exon 2 and E4 (+1689) refers to a site 1689 bp downstream of exon 4.

^dNumber of hits from NSITE demonstrating homology between previously identified regulatory element binding sites and the searched motif.

an interesting candidate for a *CARD15* regulatory element.

Bovine *CARD15* SNPs. In human, polymorphisms within the NOD are associated with Blau Syndrome and polymorphisms in the LRR are associated with Crohn's disease. This prompted our examination of the bovine *CARD15* transcript and flanking intronic sequences to gain insight into the naturally occurring variation within the United States cattle population. Thirty-six SNPs in 6176 bases (Fig. 1) were identified in the 30-animal breed panel and 26 of these were in the 5105-bp transcript (Table 2). This indicates a SNP, on average, every 172 bases within exonic and flanking intronic sequence and every 196 bp within the transcript, which is consistent with previous bovine reports (Heaton et al. 2001). Five polymorphic sites, including one nonsynonymous site, were identified

in each of the NOD and LRR domains. Each SNP in Table 2 is described relative to its position and the reference allele present in the originally generated *B. taurus taurus* sequence (GenBank accession No. AY518737). However, SNPs located within intronic flanking sequences are identified by reference to the nearest coding region. For example, E2(-32) refers to a SNP 32 bp upstream of exon 2, while E8(+12) refers to a SNP 12 bp downstream of exon 8.

Two intronic SNPs appear to be within potential regulatory regions and may have a functional role. E2(-32) is located within a 7-bp putative regulatory motif (AGAAGcC) identified by our search for conserved motifs present within the introns of human, mouse, cow, chimp, rat, rhesus monkey, and dog. The bolded "G" indicates the SNP site, while the lower-case "c" represents a mismatch in the sequence among the species. E3(-6) is located within 6

Table 2. Characteristics of bovine *CARD15* SNPs

SNP ^a	Alleles ^b	Domain affected ^c	Amino acid ^d	Allele frequency ^e
<i>E2</i> (-32)	G/A	Intron 1 (PRR)	AGAAGcC	T/I
208	A/G	CARD1	T70A	0.5/0.75
363	C/T	CARD2	H121H	0.98/0.5
<i>E3</i> (-6)	G/A/C	Splice site		1/0.75
<i>E4</i> (-58)	C/T	Intron 3		0.8/0.38
<i>E4</i> (-22)	A/G	Intron 3		0.92/0
<i>E4</i> (-16)	T/G	Intron 3		0.92/0
570	T/C	CARD2	A191A	1/0
586	G/A	LBD	V196M	1/0.13
873	C/A	NOD	S291S	1/0.94
969	C/T	NOD	E324E	1/0.88
1194	C/T	NOD	S398S	1/1 ^f
1514	C/A	NOD	T505N	1/0.13
1569	T/C	NOD	L523L	1/0.31
1723	T/C	LBD	C575L	1/0.44
1992	G/A	LBD	A664A	1/0.81
2042	G/A	LBD	R681Q	1/0.25
2066	G/A	LBD	R689H	1/0.13
2145	C/T	LRR	L715L	1/0
2197	T/C	LRR	C733R	1/0
2364	C/T	LRR	G788G	0.98/0.19
<i>E5</i> (-11)	T/A	Intron 4		0.95/1
2481	G/A	LRR	K827K	1/1 ^f
<i>E8</i> (+12)	T/A	Intron 8		0.88/0
2787	C/T	LRR	A929A	1/0.88
<i>E11</i> (-14)	A/G	Intron 10		0.85/1
<i>E11</i> (-8)	C/T	Intron 10		0.97/1
<i>E12</i> (-6)	C/T	Intron 11		0.9/0.14
3020	A/T	LOD	Q1007L	0.88/1
4648	C/A	3' UTR	CTCcCTCcaTTC	0.83/1
4757	G/A	3' UTR	GAATTTAATTAg	0.97/0.25
4798	A/G	3' UTR	CAAGgTAgAAcCGG	0.8/0.17
4801	A/T	3' UTR	CAAGgTAgAAcCGG	0.95/0.17
4911	A/G	3' UTR	GAATTTAtTTtA	1/0.83
5010	A/T	3' UTR		0.93/1
5098	C/T	3' UTR	ATTtGCT	0.93/1

^aSNPs in boldface were used to estimate haplotypes. These include 18 coding and 5 noncoding SNPs. *E2* (-32) identifies a SNP positioned 32 bp upstream of the first base of the second coding exon and *E8* (+12) identifies a SNP positioned 12 bp downstream of the last base of exon 8.

^bAlleles are reference allele (GenBank accession No. AY518737)/nonreference allele.

^cPRR = putative regulatory region identified by homology across all seven species; CARD1 = first caspase recruitment domain; CARD2 = second caspase recruitment domain; LBD = located between domains; NOD = nucleotide oligomerization domain; LRR = leucine-rich repeat; LOD = located outside of last domain; 3' UTR = 3' untranslated region.

^dAmino acid or conserved "motif" harboring polymorphic site. The boldface base within the conserved "motif" represents the polymorphic site. Upper-case letters represent perfect matches across all examined species and lower-case letters represent a mismatch in one or more of the examined species. Bovine motifs are presented.

^eFrequency of the *B. taurus taurus* reference allele (GenBank accession No. AY518737) in *B. taurus taurus* (T)/in *B. taurus indicus* (I).

^fSNPs 1194 and 2481 segregate only in Ankole-Watusi individuals but not in *B. taurus taurus* or *B. taurus indicus*.

bases of the start of the third exon and could potentially alter splicing. No SNPs were found within the 5' UTR, which suggests that this region is under strong purifying selection in cattle. However, the 5' UTR was not conserved between human, mouse, and cow, suggesting that the mechanisms underlying *CARD15* translation may have evolved differently among the species. The bovine 3' UTR contains additional sequence not present in human or mouse and which caused a poor alignment across

species. Despite the poor sequence alignment, six of the seven SNPs identified in the bovine 3' UTR were located within a short sequence that was conserved among human, mouse, and cow.

When compared with the reference sequence, the overall frequency of substitutions was greater in *B. taurus indicus* than in *B. taurus taurus* ($\chi^2 = 84.53$; 2 df). The frequency of synonymous and nonsynonymous substitutions was not significantly different in *B. taurus indicus*, but in *B. taurus taurus*

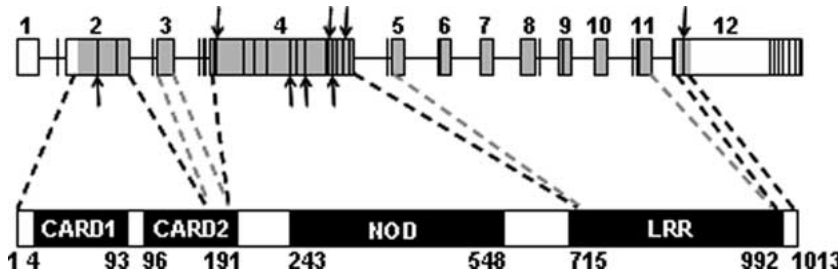


Fig. 1. Location of SNPs within bovine *CARD15*. The top portion of the figure represents the genomic structure of bovine *CARD15* and the location of the identified SNPs. Lines drawn between boxes denote introns that are not drawn to scale. Only portions of the intronic sequence flanking each exon were surveyed for SNPs (see Supplementary Table 1). Bovine *CARD15* comprises 12 exons (numbered boxes), 11 of which are coding. Gray = coding exons; white = UTRs; vertical lines = SNPs; arrows = nonsynonymous SNPs. Dotted lines connect exons to corresponding protein domains in the bottom portion of the figure. Bovine *CARD15* comprises four protein domains (black boxes), two caspase recruitment domains (CARDs), a nucleotide oligomerization domain (NOD), and a leucine-rich repeat region (LRR). Numbers below the boxes indicate amino acid residue positions.

the frequency of nonsynonymous substitution exceeded that of synonymous substitution.

Haplotypes. Twenty-two haplotypes were predicted from a total of 41 animals (Table 3). All haplotypes were subspecies specific, with seven unique haplotypes explaining the diversity within 30 *B. taurus taurus* animals and 13 unique haplotypes explaining the diversity within 9 *B. taurus indicus*

animals. One unique haplotype was predicted in Texas Longhorn, while two unique haplotypes were predicted in N'Dama. Two unique haplotypes were also predicted in Ankole-Watusi individuals which are derived from an ancient *B. taurus taurus* × *B. taurus indicus* cross. The most common taurine haplotype is consistent with the GenBank *B. taurus taurus* reference sequence (AY518737). We derived the consensus haplotype sequence for *B. taurus in-*

Table 3. Bovine *CARD15* haplotypes estimated for 23 SNPs

Haplotype	Taurine breeds ^a						Hybrid ^b	Indicine breeds ^c	
	H	J	A	L	LH	N	AW	B	G
00000000000000000000000000000000	12 (10)	3	4	3	2		1		
00000000000000000000000000000010			1						
00000000000000000000000000001000	2 (2)	1	1	1	1				
00000000000000000000000000000001	(2)	2 (6)			2				
0000000000000000000000000000010000									2
00000000000000000000010100010									2
000000000000000000000100000101					1				
0000000000000000000001000001							2		
00000100000000000000000101							1		
011000110011111100100010									2
011000111111111100100010									3
01100010101111100100010									2
01100011101111100100010									1
01100011100111100100010									
01100011100111000100010									1
111000110011111100100010									1
11100011001111100110000								2	
11100011001111100110000								1	
11100010101111100100000								1	
01001000000011000110000									1
01011000000011000100000									1
11001000111111000100010									(1)
11101010010011100110000									(1)

Numbers in each cell represent the number of haplotypes belonging to control individuals and the number of haplotypes belonging to case individuals with Johne's disease is in parentheses.

^aH = Holstein; J = Jersey; A = Angus; L = Limousin; LH = Texas Longhorn; N = N'Dama.

^bAW = Ankole-Watusi, an ancient cross between *B. taurus taurus* and *B. taurus indicus*.

^cB = Brahman; G = Gir.

dicus by assuming that the most common allele at each SNP locus was present in the consensus sequence. Twelve sites differ between the consensus indicine (0110001110111100100010) and consensus taurine (00000000000000000000) haplotypes. The consensus *B. taurus taurus* haplotype was observed on 34 of 60 chromosomes (56.6%) while the consensus *B. taurus indicus* haplotype was observed once in 18 chromosomes (5.6%).

Association tests. We found no significant associations between variation in *CARD15* and disease status when we examined differences in the frequencies of (1) haplotypes, (2) heterozygotes, (3) homozygotes for the rare allele, or (4) the class of pooled heterozygotes and homozygotes for the rare allele. However, two SNPs differed in allele frequency between the case and control groups ($p < 0.05$) within *B. taurus taurus*, but no allele frequency differences were observed at these sites between the case and control Holsteins. Within *B. taurus taurus*, the frequency of the reference "G" allele for *E2(-32)* was 1 in the case group and 0.5 within the control group ($\chi^2 = 10.8$; 1 df). Also within *B. taurus taurus*, the frequency of the reference allele for *3020* was 0.6 in the case group and 0.875 in the control group ($\chi^2 = 4.43$; 1 df). Of these, the most likely to have a functional consequence is *E2(-32)*. In a case control study, we would expect a causal SNP for a recessive inherited trait to be fixed for one allele in the case group but to be variable, with all possible genotypes present, in the control group since this group would likely consist of both resistant and susceptible but unchallenged individuals. Consequently, if *E2(-32)* plays a role in susceptibility to Johne's disease, the incomplete penetrance within the control group suggests that susceptibility is either polygenic, environmentally influenced, or that the GG control animals had either not been exposed to the mycobacterium or were yet to exhibit signs of Johne's disease. An alternative explanation is that this locus is free to evolve and that small sample size has led to a spurious association.

Overall, these data provide comprehensive sequence information for bovine *CARD15*. The SNPs identified within bovine *CARD15* will assist QTL or candidate gene studies targeted at identifying genes associated with disease resistance or susceptibility. Our comparative analysis highlights genomic regions that warrant further investigation to identify polymorphisms with effects on the regulation of *CARD15*. Finally, the naturally occurring variation present in bovine *CARD15* in *B. taurus taurus* and *B. taurus indicus* is reported. A SNP survey was performed in 11 diseased animals and SNP charac-

teristics were compared with those of animals within a breed panel. We found no evidence for a simple causal relationship between variation in bovine *CARD15* and Johne's disease. However, the small sample size, the admixture of breeds in the control group, and the fact that no animal within the control group was experimentally challenged with *M. ptb.* could easily have obscured a causal genetic relationship between *CARD15* and disease. Consequently, it is important that the SNPs identified in this study be used to screen larger populations of animals in which the experimental control of breed and phenotype definition has been possible.

Acknowledgments

The authors gratefully acknowledge the assistance and support of Janice Elliott. This work was supported by a Programs of Excellence grant from the Life Sciences Task Force of Texas A&M University, USDA-CREES NRI grant 99-35205-8534, US DHS BAA-ONR grant N00014-04-1-0 from the Department of Homeland Security, and grant 517-0186-2001 from the State of Texas Advanced Technology Program.

References

1. Borchert A, Savaskan NE, Kuhn H (2003) Regulation of expression of the phospholipid hydroperoxide/sperm nucleus glutathione peroxidase gene. Tissue-specific expression pattern and identification of functional cis- and trans-regulatory elements. *J Biol Chem* 278, 2571–2580
2. Brend T, Gilthorpe J, Summerbell D, Rigby PW (2003) Multiple levels of transcriptional and post-transcriptional regulation are required to define the domain of Hoxb4 expression. *Development* 130, 2717–2728
3. Chevalet C, Corpet F (1986) Statistical decision rules concerning synteny or independence between markers. *Cytogenet Cell Genet* 43, 132–139
4. Chiodini RJ (1989) Crohn's disease and the mycobacterioses: a review and comparison of two disease entities. *Clin Microbiol Rev* 2, 90–117
5. Cocito C, Gilot P, Coene M, de Kesel M, Poupart P, Vannuffel P (1994) Paratuberculosis. *Clin Microbiol Rev* 7, 328–345
6. Giacomelli F, Rosatto N, Divizia MT, Cusano R, Caridi G, et al. (2003) The first intron of the human osteopontin gene contains a C/EBP-beta-responsive enhancer. *Gene Expr* 11, 95–104
7. Girardin SE, Boneca IG, Carneiro LA, Antignac A, Jehanno M, et al. (2003) Nod1 detects a unique muropeptide from gram-negative bacterial peptidoglycan. *Science* 300, 1584–1587

8. Gutierrez O, Pipaon C, Inohara N, Fontalba A, Ogura Y, et al. (2002) Induction of Nod2 in myelomonocytic and intestinal epithelial cells via nuclear factor-kappa B activation. *J Biol Chem* 277, 41701–41705
9. Heaton MP, Grosse WM, Kappes SM, Keele JW, Chitko-McKown CG, et al. (2001) Estimation of DNA sequence diversity in bovine cytokine genes. *Mamm Genome* 12, 32–37
10. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, et al. (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 411, 599–603
11. Inohara N, Nunez G (2001) The NOD: a signaling module that regulates apoptosis and host defense against pathogens. *Oncogene* 20, 6473–6481
12. Iwanaga Y, Davey MP, Martin TM, Planck SR, DePriest ML, et al. (2003) Cloning, sequencing and expression analysis of the mouse NOD2/ CARD15 gene. *Inflamm Res* 52, 272–276
13. Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: *Mammalian Protein Metabolism*, Munro HN (ed.) (New York: Academic Press), pp 21–132
14. King K, Sheikh MF, Cuthbert AP, Fisher SA, Onnie CM, et al. (2006) Mutation, selection, and evolution of the Crohn disease susceptibility gene CARD15. *Hum Mutat* 27, 44–54
15. Miceli-Richard C, Lesage S, Rybojad M, Prieur AM, Manouvrier-Hanu S, et al. (2001) CARD15 mutations in Blau syndrome. *Nat Genet* 29, 19–20
16. Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet* 70, 157–169
17. Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, et al. (2001a) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 411, 603–606
18. Ogura Y, Inohara N, Benito A, Chen FF, Yamaoka S, et al. (2001b) Nod2, a Nod1/Apaf-1 family member that is restricted to monocytes and activates NF-kappaB. *J Biol Chem* 276, 4812–4818
19. Ogura Y, Saab L, Chen FF, Benito A, Inohara N, et al. (2003) Genetic variation and activity of mouse Nod2, a susceptibility gene for Crohn's disease. *Genomics* 81, 369–377
20. Slonim D, Kruglyak L, Stein L, Lander E (1997) Building human genome maps with radiation hybrids. *J Comput Biol* 4, 487–504
21. Thompson DE (1994) The role of mycobacteria in Crohn's disease. *J Med Microbiol* 41, 74–94
22. Van Laere AS, Nguyen M, Braunschweig M, Nezer C, Collette C, et al. (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* 425, 832–836
23. Womack JE, Moll YD (1986) Gene map of the cow: conservation of linkage with mouse and man. *J Hered* 77, 2–7
24. Womack JE, Johnson JS, Owens EK, Rexroad CE III, Schlapfer J, et al. (1997) A whole-genome radiation hybrid panel for bovine gene mapping. *Mamm Genome* 8, 854–856