

Generation and maintenance of diversity in the cattle MHC class I region

James Birch · Lisa Murphy · Niall D. MacHugh · Shirley A. Ellis

Received: 10 April 2006 / Accepted: 31 May 2006 / Published online: 29 June 2006
© Springer-Verlag 2006

Abstract Major histocompatibility complex (MHC) class I genes play a crucial role in the immune defence against intracellular pathogens. An important evolutionary strategy is to generate and maintain a high level of diversity in these genes. Humans express three highly polymorphic classical MHC class I genes (HLA-A, HLA-B and HLA-C). In contrast, some species, for example rat and rhesus macaque, maintain diversity by generation of haplotypes that vary considerably with regard to the number and combination of transcribed genes. Cattle appear to use both strategies. We show that various combinations of six apparently classical genes, three of which are highly polymorphic, are transcribed on different haplotypes. Although additional sequences were identified in both cDNA and gDNA, it was not possible to assign them to any of these defined genes. Most were highly divergent or were non-classical class I genes. Thus, we found little evidence for frequent duplication and deletion of classical class I genes as reported in some other species. However, the maintenance of class I diversity in cattle

may involve limited gene shuffling and deletion, possibly as a result of unequal crossing-over within the class I region.

Keywords MHC · Cattle class I genes · Haplotype diversity

Introduction

The important immunological role played by the major histocompatibility complex (MHC) is largely dependent on maintenance of high levels of polymorphism in these genes at the population level (Parham and Ohta 1996). Recent studies of the MHC in livestock (cattle and sheep) and other species (horse, rat and rhesus macaque) have demonstrated greater genetic complexity than previously predicted (Ellis 2004; Miltiadou et al. 2005; Roos and Walter 2005; Otting et al. 2005). This complexity involves variable gene content in different haplotypes as well as allelic diversity. Much of the groundwork in MHC research was carried out in human, where there are only three classical MHC class I genes consistently present and expressed on all haplotypes. It has been clearly demonstrated that diversity in these genes is generated by point mutations and intralocus recombination, and it appears that the whole process is driven by pathogen selection (Parham et al. 1995). While the driving forces may be the same in other species, the evolutionary strategies used to generate and maintain polymorphism are clearly more diverse than what might have been anticipated by these early studies. To understand and manipulate the immune response to pathogens in economically important species such as cattle, it is therefore necessary to generate a clear picture of MHC diversity at the genomic, haplotype and allele level.

The first two authors made an equal contribution to this work.

J. Birch · S. A. Ellis (✉)
Immunology Division, Institute for Animal Health,
Compton RG20 7NN, UK
e-mail: shirley.ellis@bbsrc.ac.uk

L. Murphy
Division of Animal Production and Public Health,
Institute for Comparative Medicine, Faculty of Veterinary
Medicine, Glasgow University,
Bearsden Road,
Glasgow G61 1QH, UK

N. D. MacHugh
The Centre for Tropical Veterinary Medicine, Royal (Dick)
School of Veterinary Studies, University of Edinburgh,
Easter Bush,
Roslin, Midlothian EH25 9RG, UK

There are currently 61 validated, full-length cattle MHC class I cDNA sequences submitted to IPD-MHC (<http://www.ebi.ac.uk/ipd/mhc/bola/>; Ellis et al. 2006). Previous studies have led to a limited understanding of the relationships between these sequences (Holmes et al. 2003). It has recently become clear that cattle have multiple (six or more) apparently classical class I genes, and limited haplotype analyses suggest that between one and three of these are usually transcribed (Ellis et al. 1999). The classification of these genes as ‘classical’ is based on their apparently ubiquitous transcription and varying levels of polymorphism (Ellis et al. 2005). The classical class I genes in cattle are currently named in a single series (prefixed N*) due to the difficulties in all but a few cases of definitive assignment to a locus (Ellis et al. 2006). The non-classical sequences are similarly named but numbered from N*50001. These sequences demonstrate a range of characteristics consistent with this classification; for example, they tend to show little if any polymorphism and have restricted tissue distribution (Davies et al. 2006).

The origins of modern cattle remain controversial, although it is clear that they result from numerous domestication events from genetically diverse populations (Loftus et al. 1994). Results of attempts to assess MHC diversity using serological reagents suggested that different cattle breeds expressed distinct sets of class I alleles (Stear et al. 1988); however, it is not known whether or not the gene content and usage vary between breeds. The majority of validated MHC sequences are derived from Holstein cattle, which constitute the basis of the commercial dairy industry worldwide; the data described in this paper were generated primarily from this breed.

The aim of this study was to generate sufficient sequence and haplotype data to allow a realistic assessment of the mechanisms that generate and maintain diversity in the cattle MHC class I region. An improved understanding of the derivation of MHC diversity in modern dairy cattle can be used to assess the likely impact of modern breeding practices on the continued ability of these populations to mount effective immune responses after infection or vaccination.

Materials and methods

Animals, genomic DNA extraction, RNA extraction and cDNA synthesis

All cattle used in this study were Holsteins and were part of the Institute for Animal Health herd. They were serologically class I typed where appropriate as described in Ellis et al. (2005). In most cases, breeding records and family members (offspring and siblings) were available. Peripheral blood mononuclear cells were obtained from venous blood

by density gradient centrifugation. Genomic DNA (gDNA) was isolated from 5×10^6 cells using the QiaAmp DNA blood mini kit (Qiagen, Crawley, UK). Poly-adenylated mRNA was isolated from 5×10^6 cells using the Dynal mRNA DIRECT kit (Invitrogen, Paisley, UK). First-strand cDNA was synthesised from the mRNA using an oligo (dT)_{12–18} primer and Superscript II reverse transcriptase (Invitrogen).

Amplification and sequencing of cattle MHC class I

The primers are detailed in Table 1. The amplification of MHC class I was performed on both cDNA and gDNA using primer pair Bov 7 and Bov 14 with the aim of amplifying a region containing part of exon 2, exon 3 and part of exon 4 (separated by introns 2 and 3 in the case of gDNA) to facilitate gene identification. Additional polymerase chain reaction (PCR) amplification was performed on gDNA using an alternative reverse primer in exon 3 (Bov 11). These primers have been designed to amplify from as many different cattle MHC class I sequences as possible (<http://www.ebi.ac.uk/ipd/mhc/bola/>) and have been used extensively (Pichowski et al. 1996; Ellis et al. 1999). PCR from cDNA template (~20 ng) was carried out in a final volume of 25 μ l containing 1 \times PCR buffer (20 mM Tris-HCl, pH 8.4, and 50 mM KCl; Invitrogen), 2.5 mM MgCl₂, 0.25 mM each dNTP, 1 μ M each primer and 1.25 units of *Taq* polymerase (Invitrogen). For amplification from gDNA template (~100 ng), the same mix was used, with the addition of DMSO to a final concentration of 4%. The thermal cycling profile used for amplification from cDNA was: 95°C for 1 min, 32 cycles of 95°C for 30 s, 55°C for 20 s, 72°C for 30 s and followed by 72°C for 5 min. For gDNA, the same profile was used over 35 cycles, and the initial incubation time at 95°C was increased to 5 min. All thermal cycling were performed on a PTC-200 thermal cycler (MJ Research, Incline Village, NV, USA) set to use calculated reaction temperatures.

For amplification of full-length MHC class I from cDNA, a mixture of primers Bov 21a/g and Bov 21-BSF (forward) and a mixture of Bov 3 and Bov 3-BSF (reverse) were used. These mixed primers were used because there are known polymorphisms at the target sites. The PCR conditions were as described above for cDNA, except that the extension time was increased to 1 min.

The PCR products were purified from 0.9% (full-length class I) or 1.2% agarose gels using the QiaQuick gel extraction kit (Qiagen) and cloned into pGEM-T (Promega, Southampton, UK). A representative number of clones (16) in each case were sequenced using the GenomeLab Dye Terminator Cycle Sequencing Quick Start Kit (Beckman Coulter, Fullerton, CA, USA) and a CEQ 8000 Genetic Analysis System sequencer (Beckman Coulter, USA). The

Table 1 Oligonucleotide primers used for amplification of cattle MHC class I sequences

Name	Direction	Location	Sequence (5'–3')
Bov 7	F	Exon 2	GGCTACGTGGACGACACG
Bov 11	R	Exon 3	CCCTCCAGGTAGTTCCT
Bov 14	R	Exon 4	GAAGGCCTGGTCTCCACAA
Bov 9	F	Exon 3/4	CGTGCTGCGCGCAGACC
Bov 15	F	Exon 2	GGTACTACAACCAGAGC
Bov 21a/g	F	Exon 1	CATGGGGCCGCGARC
Bov 21-BSF	F	Exon 1	TATGAGGCCGCGAAC
Bov 3	R	3' UTR	GGATGAAGCATCACTCAG
Bov 3-BSF	R	3' UTR	GGATGCAGCATCACTTAG
F 1	F	Gene 1, intron 1	TGCGAGGGGACCGCCCGA
R 1	R	Gene 1, intron 3	AGGTGAGAACAGGCCCTTGAGAA
F 2	F	Gene 2, intron 1	GAACRAGCGACCCCGACT
R 2	R	Gene 2, intron 3	CAAGTGGGGCAACTGGTC
F 3	F	Gene 3, intron 1	TCGACCGCTTCCATCTCG
R 3	R	Gene 3, intron 3	GAACAGGCCTTGAGAGAC
F 6	F	Gene 6, intron 1	TCATTGACCCTCCGCCCA
R 6	R	Gene 6, intron 3	GGCGCTGTTYCCACAGGC

vector primers were used for sequencing, together with internal primers Bov 9 and Bov 15 in the case of full-length MHC class I. Exon 3 sequence was not generated after PCR amplification of gDNA using Bov 7 and Bov 14 due to the presence of flanking introns 2 and 3 and no appropriate generic primers. In these cases, only exon 2 sequence is shown. Where Bov 7 and Bov 11 amplification yielded a product from gDNA, exon 1 and 2 sequences were derived. In the case of new full-length cDNA sequences, more than one PCR was performed and a consensus sequence was derived from multiple clones. The new sequences were then submitted to the cattle section of the Immuno Polymorphism Database-MHC (<http://www.ebi.ac.uk/ipd/mhc/bola/>) for validation and assignment of an official name (Ellis et al. 2006).

Reference-strand-mediated conformation analysis

Reference-strand-mediated conformation analysis (RSCA) was performed essentially as described in Arguello et al. (1998). As in human RSCA typing, primer pairs were designed to attempt amplification of cattle MHC class I alleles encoded at discrete loci. The primer design was by reference to a small number of intron 1 and intron 3 sequences (Murphy 2004) from well-characterised cattle MHC class I alleles that had previously been assigned to putative loci based on both sequence and phylogenetic analyses (Holmes et al. 2003). The region amplified contains all of exon 2, intron 2 and exon 3. A specific primer pair was designed for each of the four genes (Table 1): gene 1, gene 2, gene 3 and gene 6.

Each of the four primer pairs was used to amplify gDNA from animals carrying known and unknown MHC class I haplotypes. The presence or absence of a product of the

correct length was determined by separation in a 1% agarose/Tris–acetate–ethylendiaminetetraacetic acid gel containing ethidium bromide; the remaining portion of amplified DNA was stored at -20°C until required for RSCA. PCR was carried out in a final volume of 25 μl containing 1 \times PCR buffer, 1.5 mM MgCl_2 , 0.25 mM each dNTP, 1 μM each primer, 4% DMSO, ~ 100 ng template gDNA and 1.25 units of Taq polymerase. The thermal cycling profile used for gene 1 and gene 6 primers was: 95°C for 5 min, 35 cycles of 95°C for 30 s, 65°C for 20 s, 72°C for 1 min and followed by 72°C for 5 min. For gene 2 and 3 primers, the profile was the same, except that an annealing temperature of 60°C was used.

Two well-characterised alleles were chosen from each of the four genes to be used as fluorescently labelled reference strands (FLRs) (Table 2). The PCR amplification was carried out using the appropriate primer pair in each case, and the products were cloned into pGEM-T and confirmed to be correct by sequencing. FLRs were synthesised from these cloned products by further gene-specific PCR, this time using FAM-labelled versions of the forward primers (MWG-Biotech AG, Ebersburg, Germany). The PCR products were separated on an agarose gel; bands of the correct size were

Table 2 Alleles used for production of FLRs

Gene	FLR 1	FLR 2
Gene 1	<i>N*02101</i>	<i>N*02301</i>
Gene 2	<i>N*01801</i>	<i>N*02501</i>
Gene 3	<i>N*01701</i>	<i>N*00201</i>
Gene 6	<i>N*01301</i>	<i>N*01401</i>

excised, and DNA was purified using a QiaQuick gel extraction kit (Qiagen) and stored at -20°C for future use.

After amplification from genomic DNA as described above, each PCR product was annealed to both of the FLRs of the same group, in separate reactions. This was done by mixing 3 μl of PCR product with 1 μl of FLR and performing the denaturing and re-annealing program: 95°C for 4 min, 55°C for 5 min and 15°C for 3 min. Two microlitres of each annealed PCR product/FLR mix was mixed with 2 μl of ficoll loading dye together with 0.5 μl of GeneScan-2500 ROX size standard (Applied Biosystems, Warrington, UK) and loaded onto a 6.5% non-denaturing polyacrylamide gel. Electrophoresis was carried out in an ABI Prism 377 DNA sequencer (Perkin Elmer, Foster City, CA, USA) for 2 h at 35 mA and 51°C (pre-run) followed by 10 h at 60 mA and 40°C . The analysis was performed using the GeneScan 3.1 software (Perkin Elmer). The mobility values were calculated by subtracting the apparent size of the FLR homodimer peak from those of the heterodimer peaks. The alleles were identified by comparing their mobility values to those of previously sequenced alleles. In the case of alleles with no matching mobility value, the initial sequence analysis was performed on products (exon 2, intron 2 and exon 3) amplified from gDNA using gene-specific primers. This allowed a full-length clone in each case to be identified and sequenced after PCR from cDNA generated from the same animal.

Results

Identification of classical class I alleles and provisional gene assignment

A number of new class I alleles were identified during the course of RSCA typing and, in each case, a full-length cDNA sequence was subsequently derived as described above. The new alleles are *N*02601*, *N*02701*, *RSCA2.1*, *N*03801*, *N*01201*, *N*01502* and *N*03101*. These were assigned to genes 1, 2, 3 or 6 based on their amplification by gene-specific primers located within introns 1 and 3, as described in “Materials and methods”. The exception was *N*01201*, which did not amplify with gene-specific primers but was detected during analysis of cDNA using generic primers. Their appropriate gene assignment was supported with full-length cDNA sequence analysis and alignment of 3' sections with other class I alleles (Fig. 1).

Figure 1 shows an alignment of partial coding sequence (predicted amino acids) from all of the new alleles together with a selection of other alleles (for comparison) previously assigned to the putative six classical class I genes, as previously defined by Holmes et al. (2003). The region shown is the transmembrane and cytoplasmic domain, as this area appears to show distinct gene-specific features, in contrast to

the extracellular domains. The cytoplasmic domain in particular is variable, seemingly more so than in other species, for example human or pig (<http://www.ebi.ac.uk/ipd/mhc/sla>; www.ebi.ac.uk/imgt/hla/index.html). There is minor variation in this entire region within genes 1, 2 and 3. In these cases, the sequences appear to fall into two sub-groups; this has been noted previously, as a similar phenomenon is observed when phylogenetic analysis is performed (Holmes et al. 2003), which may suggest that these genes have duplicated. In support of this, one haplotype is shown in Table 3 (A17) that appears to express two alleles encoded by gene 2 (*N*00602* and *N*00802*), and their sequences do differ in the 3' region (Fig. 1). *N*00602* has a deletion at the end of exon 7 that results in the removal of the stop codon; it is unknown how far translation is likely to continue into the 3' UTR, as the primer used to amplify this allele is located shortly after this point. The closely related allele, *N*00601*, identified in a different cattle breed (Garber et al. 1994), is identical in this region and thus demonstrates the same deletion, possibly indicating that these alleles are not functional. Two representative non-classical genes (*N*50001* and *N*50101*) are also included in Fig. 1 to demonstrate that the 3' region of these genes is quite distinct from the classical genes.

MHC class I haplotype configurations

Some of the haplotypes included in this study have been analysed and described previously: A11, A19, w12B, A12 (Ellis et al. 1998), A31, A18 (Ellis et al. 1996), A14 (Ellis et al. 1999), A10/KN104 (Bensaid et al. 1991) and A33 (Ennis et al. 1988). The additional haplotypes analysed by RSCA and sequencing in this study were A10, A13, A15, A17, A20, and RSCA2. In all cases, the alleles were assigned to a haplotype by use of MHC-homozygous animals, by MHC-heterozygotes where one haplotype was already characterised or through reference to the MHC genes inherited by offspring or shared with siblings. In the case of A10, A15 and A20, the haplotypes were analysed from more than one unrelated individual. The alleles from the previously analysed haplotypes were assigned to putative genes by a combination of sequence and phylogenetic analyses (Holmes et al. 2003; Ellis et al. 2006).

Table 3 details the classical MHC class I alleles transcribed in a number of different haplotypes, chosen to represent different configurations of six putative genes. The majority of alleles are assigned to genes 1, 2 or 3. Gene 5 is represented in Table 3 by a single allele, *N*00301*, which has a unique transmembrane length. In this study, a new allele, *N*03901*, identified from a Holstein animal that was not serologically typed, shared this transmembrane region and most of the cytoplasmic domain with *N*00301* (Fig. 1). *N*03901* is not included in Table 3 as there are insufficient

Fig. 1 Figure shows an alignment of the predicted amino acid sequences derived from 38 cattle class I cDNA sequences, encompassing the entire transmembrane and cytoplasmic domains. Representative sequences are shown for genes 1–6. In addition, two non-classical sequences are shown and, likewise, one sequence (*N*04001*) that is not assigned to a gene. The details of all sequences can be found at <http://www.ebi.ac.uk/ipd/mhc/bola>. Dashes indicate identity; dots represent gaps introduced to maximise alignment. Asterisk indicates stop codon

	Transmembrane domain	Cytoplasmic domain
Gene 1		
N*01901	EPPQTSFLIMGIIIVGLVLLVV..AVVAGAVIWRKKRS	GEKRQTYTQAASGSDSDQGSVDVSLTVPKV*
N*00701	-----	-----
N*00901	-----	-----
N*02001	-----	-----
N*02301	-----	-----
N*03101	-----T-----	-----
N*02101	-----T-----	-----
N*02801	-----T-----	-----
Gene 2		
N*01601	-----T-----	---GRI---S---A-----
N*01801	-----T-----	---GRI---S---A-----
N*02601	-----T-----	---GRI---S---A-----
N*01201	-----T-----	---GRI---S---A-----
N*00802	-----T-----	---GRI---S---T-----
RSCA2.1	-----T-----	---GRI---S---A-----
N*02501	-----T-----	---GRI---S---T---M-----
N*00602	-----L-----	---GRI---S---T-----AV
N*00501	-----L-----	---GRI---S---A-----
N*02201	-----L-----	---GRI---S---A-----
Gene 3		
N*02701	---P---T---TG---V---CM---P---	---G---I---L---S---A-----
N*01701	---P---T---TG---V---CM---	---G---I---S---S---A-----
N*00101	---P---T---TG---V---CM---	---G---I---S---S---A-----
N*01101	---P---T---TG---V---CM---	---G---I---S---S---A-----
N*03601	---P---T---TG---V---CM---	---G---I---S---S---A-----
N*03701	---P---T---TG---V---CM---	---G---I---S---S---A-----
N*03801	---P---T---TG---V---CM---	---G---I---S---S---A-----
N*00201	---P---T---TG---V---VCM---	---G---I---S---S---A-----
N*00401	---P---T---TG---V---CM---	---GGN---I---S---S---A-----
N*01001	---P---T---TG---V---CM---	---GGN---I---SGS---A-----
Gene 4		
N*02401	---T---TG---M---R---	---GGN---I---S---S---A-----
Gene 5		
N*03901	---T---TG---F---M---	---GGN---I---SRS---A-----
N*00301	---T---TG---F---M---	---GGN---I---S---SA---AR-----
Gene 6		
N*03401	---T---M---H---	---R---N---A-----
N*01301	---T---M---H---	---R---N---A-----
N*01401	---T---M---H---	---R---N---A-----
N*01502	---T---M---H---	---R---N---A-----
N*04001	---P---TV---C---	---GR---S---N---A-----
N*50001	---P---VP---LV---S---	---GRI---*
N*50101	---P---IP---M---TG---T---M---H---	---GRG---T---SY---M-----

data concerning the haplotype in which it was found. Gene 4 is represented by only two alleles (*N*02401* and *N*02402*). PCR-sequence-specific primers targeting *N*02401* and PCR amplification from many cDNA samples with generic primers consistently failed to reveal any clearly related sequences (data not shown). The alleles assigned to gene 6 show a reasonably close relationship with gene 1 alleles after phylogenetic analysis (Holmes et al. 2003); however, the cytoplasmic region is quite distinct (Fig. 1), and the introns were sufficiently different to allow discrete amplification.

The data presented in Table 3 demonstrate that almost all haplotypes have a transcribed allele assigned to gene 2, together with either a gene 1- or gene 3-derived allele. Not

all gene combinations were found; genes 1, 2 and 3 occur most frequently, but while 1 and 2, and 3 and 2 occur together, 1 and 3 have never been found. The most obvious explanation for this would be that they are in fact the same gene, but there is no evidence to support this. They appear to have no such relationship by phylogenetic analysis (Holmes et al. 2003), and their 3' regions are quite different (Fig. 1).

The only haplotypes expressing three apparently classical alleles are A17, with an apparently duplicated gene 2, and A14 and A15, which share two of their alleles, *N*02501* and *N*02401*. Only one related allele to *N*02401* has been identified (*N*02402*; Holmes et al. 2003), and it therefore seems that this particular gene is rarely present/expressed and, in this sense, the A14 and A15 haplotypes may be

Table 3 Transcribed classical MHC class I genes in haplotypes selected to show a range of gene content

Haplotype	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6
A19 (A6)		<i>N*01601</i>				<i>N*01401</i>
A17 (A6)		<i>N*00602, N*00802</i>				<i>N*01502</i>
A11		<i>N*01801</i>	<i>N*01701</i>			
A20		<i>N*02601</i>	<i>N*02701</i>			
A33		<i>N*00501</i>	<i>N*00401</i>			
RSCA2 ^a		RSCA2.1 ^b	<i>N*03801</i>			
A10		<i>N*01201</i>	<i>N*00201</i>			
A14	<i>N*02301</i>	<i>N*02501</i>		<i>N*02401</i>		
A15	<i>N*00901</i>	<i>N*02501</i>		<i>N*02401</i>		
W12B (A30)	<i>N*01901</i>	<i>N*00801</i>				
A12 (A30)	<i>N*02001</i>	<i>N*00801</i>				
A31	<i>N*02101</i>	<i>N*02201</i>				
A10/KN104 ^c			<i>N*00101</i>		<i>N*00301</i>	
A13	<i>N*03101</i>					
A18 (A6)						<i>N*01301</i>
A18 (A6)						<i>N*01302</i>

^aA serological specificity has not been determined for this haplotype

^bA formal name has not yet been assigned to this allele

^cThis haplotype was reported and analysed by Bensaid et al. (1991). It is possible that additional genes are transcribed

rather unusual. Despite that, they are found at relatively high frequency in Holstein cattle (data not shown).

Although cattle MHC class I sequence data are still quite limited, with currently 61 alleles on IPD-MHC (<http://www.ebi.ac.uk/ipd/mhc/bola/>), it appears that genes 1, 2 and 3 are equally polymorphic, and all have minor sequence variants, suggesting that they are under selection pressure (Ellis et al. 2005) and are therefore presumed to be functional.

Additional (unassigned) transcribed class I alleles

Additional class I sequences can be detected in cDNA that cannot be assigned to any of the six putative classical class I genes. In most cases, these are partial sequences consisting of only the extracellular domains, and it is not known if transcription of full-length cDNA occurs. Some of these could be referred to as ‘classical’ class I pseudogenes, because the available sequence includes no divergent substitutions, but they each have one or two nucleotide deletions in alpha 1 or 2 causing a change in reading frame. Examples are *N*03301N* (see Fig. 2), found on the A11 haplotype, and *N*03201N*, found on the A13 haplotype (Davies et al.; <http://www.ebi.ac.uk/ipd/mhc/bola/>). A pseudogene similar to *N*03301N* was found during this study on the A10 haplotype.

To determine the frequency with which such sequences are transcribed, three haplotypes were chosen for closer examination: A14, A18 and A31 (Table 4). This was achieved by PCR amplification from cDNA with two primer pairs designed to amplify partial exon 2 and 3 from all known alleles, followed by cloning and sequencing. In addition to

the characterised classical class I alleles, this exercise revealed two additional sequences in the A18 haplotype, none in the A31 haplotype and three in the A14 haplotype (Table 4; Fig. 2). It is clear that additional transcripts may be present which do not contain the appropriate sequences for primer binding or are at such low levels that they are unlikely to be detected using this method. Figure 2 includes two classical class I alleles (*N*00101* and *N*01801*) and six non-classical alleles (*N*50001–N*50501*) for comparison.

The additional transcribed sequences found in the A18 haplotype were A18.24 and A18.18. A18.24 contains many divergent amino acid substitutions, not seen in classical or non-classical genes (Fig. 2). A18.18 is quite similar to the non-classical gene *N*50101*, and as it has only two unique amino acid substitutions, it is likely to represent a non-classical class I gene. The additional transcribed sequences found in the A14 haplotype consisted of one which was identical to *N*50501* (a non-classical gene) in the region sequenced, A14.02, which is very similar to 18.24 and is therefore assumed to be a related gene, and *N*04001*. While it is possible that many of the additional genes contain deletions or are truncated, a full-length cDNA sequence of *N*04001* has been derived (available on <http://www.ebi.ac.uk/ipd/bola/>), most of which is shown in Figs. 1 and 2. It was first identified (as ‘gene Z’) in a BAC library derived from an A14-homozygous individual, and it has been mapped to a region with many class I genes (Di Palma et al. 2002), including the three expressed classical genes on this haplotype (*N*02301*, *N*02501* and *N*02401*). In addition to the A14 haplotype, two *N*04001* variant partial sequences have been found transcribed on related A15

Fig. 2 Figure shows an alignment of the predicted amino acid sequences derived from 19 cattle class I cDNA and gDNA sequences, encompassing part of the alpha 1 and alpha 2 domains. The new sequences listed in Table 4 are shown together with two classical class I alleles (*N*00101* and *N*01801*), five non-classical alleles (*N*50001*–*N*50501*), one transcribed pseudogene (*N*03301N*) and the unassigned allele *N*04001*. Dashes indicate identity; dots represent gaps introduced to maximise alignment. Amino acids highlighted in grey are only seen in non-classical class I sequences. Amino acids highlighted in black are seen in no other cattle class I sequences

	Alpha 1			Alpha 2		
	40	*	60	*	80	100
<i>N*00101</i>	: QFVRFDSADPNRMEPRARWVEQEGPEYWDRETQRAKGNAQFFRVSLNLRGGYINQSEAS					GSHTLQWMSGCYVG
<i>N*01801</i>	: -T-----D---VP-M-----N-RIY-DT-I-AN-TAL-----F-E-Y-----					
18.5	: E-----D-----Q-RK-DA-T-AN-T-----L-Y-----					
18.56	: -----TRD-----N-R-T-DS-LT-AN-----					
31.70	: -----D-----R-QA-T-DTK-T-AN-T-----					
31.69	: E-----D-----Q-RK-DA-T-AN-S-----					
<i>N*04001</i>	: -----K-----E-T-V-N-T-----C-D-----					
<i>N*03301N</i>	: -Q-----N-RN-----					
<i>N*50301</i>	: -A-----P-M-----EAM-RD-KAQ-RL-TG-TI-F-----VL-D-----					
<i>N*50401</i>	: -A-----P-M-----EAM-RD-KAQ-RL-TG-TI-F-----VL-D-----					
14.2	: -IA-----Y-----P-M-----E-EV-GS-EAQRRL-LE-TM-F-----F-VF-D-----					
18.24	: -IA-----Y-----P-M-----E-EV-GS-EAQRRL-LE-TM-F-----F-VF-D-----					
<i>N*50201</i>	: -A-----P-M-----EEM-RD-ESQ-ES-LC-Y-----E-I-V-F-E-----					
14.41	: -A-----P-M-----EEM-RD-ESQ-KS-LC-Y-----					
<i>N*50101</i>	: -----D-I-T-----HQ-----T-DT-----Y-T-----V-E-Y-D-----					
18.18	: -----E-D-I-T-----HQ-----T-DT-----Y-T-----V-E-Y-D-----					
18.14	: -----TRD-----DQ-KT-DT-I-QAN-T-----E-Y-D-----					
<i>N*50001</i>	: -----R-D-----D-----Q-RIQ-E-T-T-AN-T-L-----I-H-G-----					
<i>N*50501</i>	: -----R-D-----D-----Q-RIQ-E-T-T-AN-T-L-----I-H-G-----					
		*	120	*	140	*
<i>N*00101</i>	: PDGRLRRGFMQFGYDGRDYALNEDLRSWTAVETMAQISKRKMEAAAGEAEV					
<i>N*01801</i>	: -----LL-A-----I-----AD-A-T-W-----R					
18.5	: -----LG-YE-Y-E-I-----I-----AD-A-H-V-----D-AR					
18.56	: -----					
31.70	: -----					
31.69	: -----					
<i>N*04001</i>	: -----R-W-----AD-A-----A-S					
<i>N*03301N</i>	: -----					
<i>N*50301</i>	: -E-L-IW-NA-A-----AN-V-T-W-TS-F					
<i>N*50401</i>	: -L-L-IW-NA-A-----AN-V-T-W-TS-F					
14.2	: -E-E-L-W-KA-A-Y-IS-----AN-A-T-RW-TS-DF					
18.24	: -E-L-SW-KA-A-Y-IS-----AN-A-T-RW-TS-DF					
<i>N*50201</i>	: -----L-W-KA-----I-----AD-V-T-WDVS-Q-KI					
14.41	: -----					
<i>N*50101</i>	: -I-RGYDQ-A-----I-----AD-A-VT-HNA-D-AR					
18.18	: -CFL-YD-LA-----I-Q-----AD-A-VT-NA-D-AR					
18.14	: S-CFL-YD-A-----I-N-S-----ADMA-T-H-F-LR-A-R					
<i>N*50001</i>	: S-L-YN-HA-K-----AD-A-T-W-T-A-R					
<i>N*50501</i>	: S-L-YN-HA-K-----AD-A-T-W-----R					

haplotypes (data not shown and AB008645). The *N*04001* sequence has some unusual characteristics. It shares several motifs throughout with gene 3 alleles yet has a TM length typical of groups 1, 2 and 6. It also has a small number of completely unique amino acid substitutions (Figs. 1 and 2).

Additional (unassigned) untranscribed class I alleles

To try and determine how many other class I genes were present, but not transcribed, on each haplotype, PCR was also carried out on genomic DNA from the three selected haplotypes. Previous work established that the number of class I genes (or partial genes) on the A14 haplotype was in the region of 15 (Di Palma et al. 2002) but also demonstrated that the presence of extremely divergent or fragmentary sequences made PCR amplification problematic. In this case, amplification from the A18 haplotype yielded an additional five sequences, A31 yielded an additional three sequences and A14 yielded an additional two sequences (Table 4; Fig. 2). The additional sequences found in the A18 haplotype were A18.5, A18.56 and A18.14, together with two apparently non-classical sequences, one of which demonstrates identity to *N*50301* and *N*50401* and the other to *N*50001* and *N*50501*; it is not possible to distinguish between these genes with the available

sequence (partial exon 2/exon 3). A18.5 and A18.56 have very few unusual amino acid substitutions, while A18.14 has a number of unique substitutions, particularly in the alpha 2 domain.

The additional sequences found in the A31 haplotype were A31.69, A31.70 and a non-classical sequence that demonstrates identity to both *N*50301* and *N*50401* in the amplified region. The additional sequences found in the A14 haplotype were A14.41 which is identical in the alpha 1 domain to the non-classical allele *N*50201* apart from one substitution and a sequence identical to the non-classical alleles *N*50301* and *N*50401* in the alpha 1 domain.

It is clear that this strategy will not reveal all genes, as some of the transcribed genes in both A18 (A18.24 and A18.18) and A14 (*N*02501*, *N*50501* and *N*04001*) were not found in this limited analysis of genomic DNA. However, in each case, several additional genes were detected, some of which appeared to be non-classical, while others could not be categorised with the available data.

Discussion

A range of approaches have been used in this work to verify the variable nature of cattle MHC class I haplotypes

Table 4 Sequences identified in cDNA/gDNA

Haplotype	Sequences identified in cDNA		Sequences identified in gDNA	
	Name	Characteristics	Name	Characteristics
A18	<i>N*01301</i>	Classical	<i>N*01301</i>	Classical
	A18.24		<i>N*50301</i> or <i>N*50401</i>	Non-classical
	A18.18		<i>N*50001</i> or <i>N*50501</i>	Non-classical
			A18.5	
			A18.56	
A31	<i>N*02101</i>	Classical	<i>N*02101</i>	Classical
	<i>N*02201</i>	Classical	<i>N*02201</i>	Classical
			<i>N*50301</i> or <i>N*50401</i>	Non-classical
			A31.69	
			A31.70	
A14	<i>N*02301</i>	Classical	<i>N*02301</i>	Classical
	<i>N*02401</i>	Classical	<i>N*02401</i>	Classical
	<i>N*02501</i>	Classical	<i>N*50301</i> or <i>N*50401</i>	Non-classical
	<i>N*50501</i>	Non-classical	A14.41	Non-classical
	<i>N*04001</i>			
	A14.2			

and to elucidate the underlying mechanisms. We had previously begun to assign alleles to putative genes using a combination of sequence, haplotype and phylogenetic analyses (Ellis et al. 1999; Holmes et al. 2003), together with limited mapping (Bensaid et al. 1991; Di Palma et al. 2002). In this study, we have included seven new full-length cDNA sequences and six new class I haplotypes. This has led to the emergence of a pattern that was not previously clear: there are three common, seemingly equally polymorphic genes that are only expressed in certain combinations. One of these (gene 2) is expressed on nearly all haplotypes investigated. The other genes and alternative haplotype arrangements occur less frequently.

The reason for the observed limitation on gene combinations is not clear. The fact that genes 1 and 3 are never found together indicates that they are not distinct genes, yet all of the available evidence suggests that they are. For example, phylogenetic analysis shows no particular relationship between the two groups of sequences (Holmes et al. 2003), and their intron differences allow gene-specific amplification. Mapping studies shed little light on this problem. Genes 1, 2 and 4 have been shown to be between 116 and 212 kb apart on the A14 haplotype (Di Palma et al. 2002). In the same study, it was shown that at least nine class I genes/pseudogenes occurred in this area (within 400 kb) and the cluster includes the non-classical gene *N*50001* and the unassigned *N*04001*. In a separate study, genes 3 and 5 were shown to be no more than 210 kb apart (Bensaid et al. 1991), but it is still not known how their position relates to that of genes 1, 2 and 4.

We have previously demonstrated that, in contrast to the situation in human, interlocus recombination (unequal

crossing-over) occurs in the cattle MHC class I region (Holmes et al. 2003). The clearest example is possibly shown by a comparison of *N*02301* (gene 1) and *N*02401* (gene 4), which are found on the same haplotype (A14). They are quite distinct but share almost the whole of exon 3 (alpha 2). As already discussed, these two genes have been mapped (Di Palma et al. 2002) and were shown to be a minimum of 52 kb apart, with no other class I gene in between. As there are nine class I genes in this immediate area, it seems quite probable that unequal crossing-over resulting in such hybrid genes may readily occur. This is likely to involve pseudogenes/gene fragments as well as classical class I genes. This may account for one of the problem areas encountered in this and previous studies. While a number of discrete groups of alleles clearly exist that probably do correspond to well-defined genes, there are others that fall outside of these groups, and it may therefore never be possible to assign all alleles to genes.

Genes 4, 5 and 6 have only been identified on a very small number of haplotypes; thus, it is not possible to comment on their polymorphism. In terms of functional capability, the alleles from five of the six loci have been shown to restrict CD8⁺ T-cell responses (Gaddum et al. 2003; unpublished data). The exception is gene 4; however, this may reflect the relatively small number of studies undertaken. The two alleles assigned to this gene differ at some positions predicted to be important for peptide binding, suggesting selection pressure and a shared peptide-presenting function with the other classical genes. The transcription and expression levels may also relate to function, but these have not been studied in any detail.

The comparative analysis between cDNA and genomic DNA was designed to resolve some of the issues raised above. For example, it might be anticipated that most haplotypes have a full complement of genes but that some of them are inactivated in some way. This does not appear to be the case. Although transcribed pseudogenes were found in cDNA and have been reported by others, there was no indication that they represented the ‘missing’ genes on any haplotype. It is therefore most likely that these ‘classical’ pseudogenes have arisen by relatively recent duplication of an expressed gene followed by mutation resulting in loss of function. Similar phenomena have been reported in other species (Zemmour et al. 1990). It is likely that the accumulation of additional mutations will eventually lead to loss of transcription/expression of these genes. Most additional sequences detected were very divergent and could not be assigned to any of the defined class I loci. As only a small section of these genes was amplified, it is possible that some at least represent gene fragments. In addition, it is possible that some of these sequences represent divergent genes that may be expressed in a tissue-specific manner.

An interesting observation was the number of non-classical genes amplified, mostly from gDNA. There are currently nine class I sequences (*N*50001–N*50501*) categorised as non-classical in the cattle section of IPD-MHC (<http://www.ebi.ac.uk/ipd/mhc/bola>). It has been suggested (Davies et al. 2006) that these are the products of four discrete genes. While an analysis of the sequences does appear to support this (Fig. 2), mapping data are required for confirmation. *N*50001* represents the only mapped gene (Di Palma et al. 2002); it is located in the same region as classical genes 1, 2 and 4. This gene was identified in the A18 and A14 haplotypes in this study, and earlier investigations suggest it is present on most haplotypes (data not shown). At present, there are insufficient data to determine if non-classical class I genes are generally present on all haplotypes.

Rat and rhesus macaque MHC class I haplotypes show considerable diversity with respect to presence or absence of particular genes (Walter and Gunther 2000; Roos and Walter 2005; Otting et al. 2005). In fact, MHC diversity in these species appears to be maintained almost exclusively by duplication, deletion and shuffling of genes (involving unequal crossing-over) rather than by generation of allelic diversity at fixed loci, as in human MHC. Although this does not appear to be the main explanation for the observed haplotype diversity in the cattle MHC class I region, these mechanisms are operating at some level, with evidence for interlocus recombination, gene duplication and deletion.

For example, we have evidence for a duplicated gene 2 in the A17 haplotype (Table 3) and an indication of duplication in genes 1, 2 and 3 by sequence (Fig. 1) and phylogenetic analyses (Holmes et al. 2003). Inactivation of duplicated

genes is suggested on some haplotypes by the presence of pseudogenes that appear to derive from a locus already expressed. An example is *N*03301N* which was found on the A11 haplotype. This pseudogene has characteristics most closely associated with gene 3, and an allele (*N*01701*) assigned to gene 3 is already expressed on the A11 haplotype. The deletion of genes is indicated by the fact that it is not possible to amplify sequences clearly deriving from all of the defined six genes from any given haplotype, using either cDNA or genomic DNA as template.

A mix of strategies seems to be operating within the cattle MHC class I region to maintain and generate diversity and, in some ways, this is similar to the situation seen in the human KIR region (Shilling et al. 2002). The diversity in gene number and combination as well as a reasonable level of allelic diversity, at least in some genes, seem a good way to ensure a range of responses to different pathogens within a population. However, it is not clear that this strategy will continue to operate within the constraints of a highly selective breeding programme, as employed in the dairy industry. The full extent of MHC haplotype variation and the details of the underlying mechanism may only be revealed after complete sequencing of a number of different cattle class I haplotypes. Assuming that all genes have identical functional capability, the particular configurations and their origins may not be important. Further work is needed to confirm that this is, in fact, the case. If it is shown that one or more of the class I genes have particular functional significance, for example, in relation to NK receptor interaction, it might prove desirable to modify current breeding strategies to ensure that all genes are retained at a reasonable level in the population.

Acknowledgements This work was supported by the Biotechnology and Biological Sciences Research Council, UK. We would like to thank the staff of the IAH farm, Helen Prentice for help with sampling and Alison Burrells for help with sequence analysis. The experiments carried out in this study comply with UK law.

References

- Arguello JR, Little A-M, Bohan E, Goldman JM, Marsh SGE, Madrigal JA (1998) High resolution HLA class I typing by RSCA. *Tissue Antigens* 52:57–66
- Bensaid A, Kaushal A, Baldwin CL, Clevers H, Young JR, Kemp SJ, MacHugh ND, Toye PG, Teale AJ (1991) Identification of expressed bovine class I MHC genes at 2 loci and demonstration of physical linkage. *Immunogenetics* 33:247–254
- Davies CJ, Eldridge JA, Fisher PJ, Schlafer DH (2006) Evidence for expression of both classical and non-classical major histocompatibility complex class I genes in bovine trophoblast cells. *Am J Reprod Immunol* 55:188–200
- Di Palma F, Archibald SD, Young JR, Ellis SA (2002) A BAC contig of approximately 400 kb contains the classical class I MHC genes of cattle. *Eur J Immunogenet* 29:65–68

- Ellis SA (2004) The cattle major histocompatibility complex: is it unique? *Vet Immunol Immunopathol* 102:1–8
- Ellis SA, Staines KA, Morrison WI (1996) cDNA sequence of cattle MHC class I genes transcribed in serologically defined haplotypes A18 and A31. *Immunogenetics* 43:156–159
- Ellis SA, Staines KA, Stear MJ, Hensen EJ, Morrison WI (1998) DNA typing for BoLA class I using PCR-SSP. *Eur J Immunogenet* 25:365–370
- Ellis SA, Holmes EC, Staines KA, Smith KB, Stear MJ, McKeever DJ, MacHugh ND, Morrison WI (1999) Variation in the number of expressed MHC genes in different cattle class I haplotypes. *Immunogenetics* 50:319–328
- Ellis SA, Morrison WI, MacHugh ND, Birch J, Burrells A, Stear MJ (2005) Serological and molecular diversity in the cattle MHC class I region. *Immunogenetics* 57:601–606
- Ellis SA, Bontrop RE, Antczak DF, Ballingall KT, Davies CJ, Kaufman J, Kennedy L, Robinson J, Smith DJ, Stear MJ, Stet RJ, Waller M, Walter L, Marsh SGE (2006) ISAG/IUIS-VIC Comparative MHC Nomenclature Committee report, 2005. *Immunogenetics* 57:953–958
- Ennis PD, Jackson AP, Parham P (1988) Molecular cloning of bovine class I MHC cDNA. *J Immunol* 144:642–651
- Gaddum RM, Cook RS, Furze JM, Ellis SA, Taylor G (2003) Recognition of BRSV proteins by bovine CD8⁺ T lymphocytes. *Immunology* 108:220–229
- Garber TL, Hughes AL, Watkins DI, Templeton JW (1994) Evidence for at least three transcribed BoLA class I loci. *Immunogenetics* 39:257–263
- Holmes EC, Roberts AFC, Staines KA, Ellis SA (2003) Evolution of MHC class I genes in Cetartiodactyls. *Immunogenetics* 55:193–202
- Loftus RT, MacHugh DE, Bradley DE, Sharp PM, Cunningham P (1994) Evidence for 2 independent domestications of domestic cattle. *Proc Natl Acad Sci U S A* 91:2757–2761
- Miltiadou D, Ballingall KT, Ellis SA, Russel GC, McKeever DJ (2005) Haplotype characterization of transcribed ovine major histocompatibility complex (MHC) class I genes. *Immunogenetics* 57:499–509
- Murphy L (2004) Variations in immune response as determined by MHC class I polymorphism in cattle. Ph.D. thesis, University of Edinburgh
- Otting N, Heijmans CMC, Noort RC, de Groot NG, Doxiadis GGM, van Rood JJ, Watkins DI, Bontrop RE (2005) Unparalleled complexity of the MHC class I region in rhesus macaques. *Proc Natl Acad Sci U S A* 102:1626–1631
- Parham P, Ohta T (1996) Population biology of antigen presentation by MHC class I molecules. *Science* 272:67–74
- Parham P, Adams EJ, Arnett KL (1995) The origins of HLA-A, B, C polymorphism. *Immunol Rev* 143:141–180
- Pichowski JS, Ellis SA, Morrison WI (1996) Sequence of 2 cattle MHC class I cDNAs associated with BoLA specificity A10. *Immunogenetics* 43:253–254
- Roos C, Walter L (2005) Considerable haplotype diversity in the RT1-CE class I gene region of the rat MHC. *Immunogenetics* 56:773–777
- Shilling HG, Guethlein LA, Cheng NW, Gardiner CM, Rodriguez R, Tyan D, Parham P (2002) Allelic polymorphism synergizes with variable gene content to individualize human KIR genotype. *J Immunol* 168:2307–2315
- Stear MJ, Pokorny TS, Muggli NE, Stone RT (1988) Breed differences in the distribution of BoLA-A locus antigens in American cattle. *Anim Genet* 19:171–176
- Walter L, Gunther E (2000) Physical mapping and evolution of the centromeric class I gene-containing region of the rat MHC. *Immunogenetics* 51:829–837
- Zemmour J, Koller BH, Ennis PD, Geraghty DE, Lawlor DA, Orr HT, Parham P (1990) HLA-AR, an inactivated antigen-presenting locus related to HLA-A. Implications for the evolution of the MHC. *J Immunol* 144:3619–3629