ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

doi: 10.1111/age.12561

Characterization of a deletion in the *Hsp70* cluster in the bovine reference genome

M. F. Suqueli García*[1], M. A. Castellote[†1], S. E. Feingold[†] and P. M. Corva*

*Facultad de Ciencias Agrarias, Universidad Nacional de Mar del Plata, Unidad Integrada Balcarce, C.C. 276, 7620 Balcarce, Argentina.
[†]Laboratorio de Agrobiotecnología, EEA Balcarce, Instituto Nacional de Tecnología Agropecuaria, Unidad Integrada Balcarce, C.C. 276, 7620 Balcarce, Argentina.

Summary

The 70 kilodalton heat shock proteins (Hsp70) are highly conserved molecular chaperones which have a crucial role in the stress response of the cell. In mammals, the Hsp70 proteins are encoded by a cluster of three genes: *HSPA1A*, *HSPA1B* and *HSPA1L*. In bovines, this cluster is located on chromosome 23 downstream of the major histocompatibility complex (BoLA). We detected inconsistencies in the location of markers on the *Hsp70* genes reported in the literature that pointed to a potential deletion in the bovine reference genome UMD 3.1.1. An *in silico* analysis of the bovine genomic region of the *Hsp70* cluster, using available information from public databases, confirmed the existence of a deletion of 11.1-kb spanning the *HSPA1B* gene and the intergenic region between *HSPA1B* and *HSPA1A*. Although we originally considered this an assembly error, it is most likely a particular condition of L1 Dominette 01449, the cow sequenced in the Bovine Genome Project. Moreover, we suggest a new classification of bovine *Hsp70* sequences reported in NCBI and a reassignment of the location of SNPs from dbSNP that map to the deletion on BTA23. We also compared the location of selected transcription factor binding sites on the promoters of *HSPA1A* and *HSPA1B*. The results generated in the present work could be helpful to refine the reference genome of an important livestock species and also to understand the role and the regulation of the bovine *Hsp70* genes.

Keywords cattle, genome assembly, *HSPA1B*, promoter, SNPs

Introduction

The 70 kilodalton heat shock proteins (Hsp70s) are a family of highly conserved and ubiquitously expressed molecular chaperones that protect cells from stress factors (Kishore *et al.* 2014; Malinverni *et al.* 2015). In mammals, a cluster of three *Hsp70* genes—*HSPA1A*, *HSPA1B*, and *HSPA1L*—is located close to the major histocompatibility complex locus. Although *HSPA1A* and *HSPA1B* are widely expressed and strongly induced by heat shock, *HSPA1L* expression is constitutive and tissue specific (Daugaard *et al.* 2007).

In bovines, Grosz *et al.* (1992) mapped two tandemly arrayed *Hsp70* sequences separated by approximately 8 kb of DNA to chromosome 23 (BTA23), which were designated *HSP70-1* and *HSP70-2* because of their homology with human *Hsp70* genes. Sugimoto *et al.* (2003) confirmed that *HSPA1A* and *HSPA1B* are 9 kb apart on BTA23 and that both intronless genes encode proteins of 641 amino acids with a single sequence difference (methionine/threonine) at the fifth position.

In cattle, the *Hsp70* genes have been associated with variations in several productive traits and with the occurrence of diseases. Picard *et al.* (2014) linked the concentration of Hsp70 to differences in beef tenderness in three cattle breeds. In Japanese Holsteins, polymorphisms in *HSP70-2* were associated with the susceptibility to clinical mastitis (Huang *et al.* 2015), whereas Deb *et al.* (2013) detected an association of promoter variants in *Hsp70-1* with thermal stress response and milk production in Frieswal cows.

We were interested in the genomic analysis of the *Hsp70* cluster in relation to the reproductive performance of Brahman cows in the subtropical area of Argentina. As part of our research with the *Hsp70* genes, we found inconsistencies while developing the routine work to

Address for correspondence

P. M. Corva, Facultad de Ciencias Agrarias, Universidad Nacional de Mar del Plata, C.C. 276, 7620 Balcarce, Argentina.
E-mail: corva.pablo@inta.gob.ar

[1]These authors contributed equally to this work.

Accepted for publication 08 March 2017

confirm the position of SNPs reported in the literature (Rosenkrans *et al.* 2010) in the bovine reference genome UMD3.1.1 and the alternate assembly Btau 5.0.1. A preliminary alignment against the sheep reference genome (Oar_v4.0) suggested the existence of a deletion in the region spanning the *Hsp70* cluster. Moreover, further inspection of bovine *Hsp70* sequences from GenBank showed that they did not clearly specify whether they corresponded to the *HSPA1A* or the *HSPA1B* genes.

Considering the productive relevance of *Hsp70* genes and the importance of a reliable reference genomic sequence, the aims of the present work were: to improve the annotation of the *Hsp70* genomic cluster on BTA23 using available *in silico* information, to compare the promoter sequences of the *HSPA1A* and *HSPA1B* genes across different species, to improve the annotation of *Hsp70* sequences deposited in GenBank and to determine the correct position of SNPs mapped to the *Hsp70* genes. The results of this study will be helpful in performing further gene regulation analyses and to understand the role of the genes within the cluster on productive traits. Moreover, they make a contribution to future sequencing projects through the better definition of a reference genome.

Materials and methods

Analysis of the *Hsp70* cluster in the cow reference genome

Sequences corresponding to BTA23 of the bovine reference genome (UMD3.1.1) and the bovine alternate assembly (Btau 5.0.1) were downloaded from the NCBI site (http://www.ncbi.nlm.nih.gov). Then, the region of interest spanning the *Hsp70* cluster was retrieved from each genomic sequence with standard UNIX commands, using flanking genes *NEU1* and *LSM2* as positional references. Inspection of the corresponding reference genomes of human, mouse, sheep and rat confirmed that the relative location of *NEU1* and *LSM2* with respect to the *Hsp70* cluster is highly conserved.

An NCBI BLAST search was conducted against the non-redundant nucleotide database (nr/nc) selecting *Bos taurus* on the 'Organism' filter to retrieve genomic sequences spanning the *Hsp70* cluster, using as a query the GenBank sequence M98823.1 that corresponds to the promoter and 5′ end of one of the *Hsp70* genes (Rosenkrans *et al.* 2010).

As shown in the Results section, a key finding in the BLAST searches was that the region of interest between *NEU1* and *LSM2* on BTA23 was covered by two overlapping BAC clones: CH240-369N12 (GenBank accession no. FQ482114.7) and CH240-510A19 (GenBank accession no. FQ482128.2). The sequence defined by both overlapping BAC clones was used as a reference throughout the experiment. The integrity of that reference sequence was confirmed through the alignment against short reads from independent next-generation sequencing (NGS) experiments. The same short reads from those experiments were aligned to UMD3.1.1 to confirm the nature of the discrepancies detected in the region of interest.

Sequences from eight independent bovine sequencing projects were downloaded from the Sequence Read Archive (SRA) division of NCBI (http://www.ncbi.nlm.nih.gov/sra). Cattle breeds represented in this data set were Jersey (SRR1262799, SRR1262802 and SRR1262791), Holstein (SRR1262788), Angus (SRR1262656), Simmental (SRR1262806) and Hereford (SRR866420 and SRR1365106). The SRA formatted files were converted to fastq format with the tool FASTQ-DUMP from the SRA ToolKit (http://www.ncbi.nlm.nih.gov/Traces/sra) and the quality check was performed using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

All these runs were mapped to the reference sequence defined by the overlapping BACs using BOWTIE2 v2.2.5 (Langmead & Salzberg 2012). As our reference sequence is considerably smaller than the whole genome sequences of the eight examined projects, the parameters --no-mixed and --no-discordant were included to avoid false positives on the mapping process.

To find evidence of a deletion, the same runs and paired-end reads retrieved from an independent resequencing experiment of cow L1 Dominette 01449, the reference animal in the Bovine Genome Project (SRA accession nos. SRX1177177 through SRX1177186; Whitacre *et al.* 2015) were mapped without discarding the discordantly mapped reads using as reference sequences both the previously described BACs and UMD 3.1.1. Concordantly and discordantly mapped reads within the region of interest were filtered using SAMTOOLS (Li *et al.* 2009).

In order to determine the relative location of the *Hsp70* genes in our reference sequence, paired NCBI BLAST alignments were performed against GenBank sequences AY149618.1 and AY149619.1 (*HSPA1A* and *HSPA1B* genes; Sugimoto *et al.* 2003) and GenBank sequence NM_001167895 (*HSPA1L* gene).

Classification of *Hsp70* sequences retrieved from GenBank

Inspection of bovine *Hsp70* sequences that had been submitted to GenBank showed that many of them did not clearly specify whether they corresponded to the *HSPA1A* or *HSPA1B* genes. To retrieve those sequences, an NCBI BLAST search was conducted against the non-redundant nucleotide database (nr/nc) selecting *Bos taurus* on the 'Organism' filter and using BAC CH240-510A19 (GenBank accession no. FQ482128.2) as the query. The retrieved *Hsp70* sequences were manually inspected to assign each of them to the corresponding gene.

Reassignment of SNP locations on the *Hsp70* genes

To improve the assignment of SNP locations in the *Hsp70* cluster region, three files were downloaded from dbSNP at NCBI (Build 146), corresponding to SNPs mapped to BTA23 ('Chr23' file), SNPs that did not map to the current reference sequence ('Chr Unknown' file) and SNPs on contigs that did not map to any chromosome ('Chr NotOn' file). Each entry in these files contained the SNP and 25–500 bp of flanking sequence on each side.

A command line BLAST alignment was performed between each one of the three SNP files and our reference sequence. The BLAST output was parsed with an ad-hoc R script (R Development Core Team 2012) to establish the right position on *HSPA1A* and *HSPA1B* when possible.

We hypothesized that the inconsistencies in marker locations detected in UMD 3.1.1 could have been originated by a chromosomal rearrangement. Therefore, an NCBI BLAST alignment was run against UMD3.1.1 using the deletion region of the reference sequence as query. Then, a more detailed alignment in selected regions was made using MAUVE software version 2.4.0 (Darling *et al.* 2010).

Analysis of Hsp70 promoters

In order to assess the degree of homology between *HSPA1A* and *HSPA1B* promoters, a 500-bp fragment upstream of the putative transcription start site (TSS) of each gene was retrieved for the following ruminant species: cow, bison, yak, buffalo, sheep and goat. For cow, the promoter sequences were retrieved from BAC CH240-510A19 (GenBank accession no. FQ482128.2), whereas for the other species they were retrieved from the corresponding sequencing projects [GenBank sequences: NC_019477.2 (Oar_v4.0) for sheep, JPYT01740191.1 for bison, AGSK-01139914.1 (*HSPA1A*) and AGSK01192286.1 (*HSPA1B*) for yak, NW_005782253.1 for buffalo and NC_030830.1 for goat]. All these sequences were aligned with PRO-COFFEE, a multiple sequence alignment method specifically designed for promoter regions (Erb *et al.* 2012). Also, the sequences were searched for heat shock factor (HSF) binding sites and a TATA-Box with the MAST tool in MEME SUITE version 4.11.2 (Bailey *et al.* 2009, 2015) using the following motifs from the JASPAR 2016 database (Mathelier *et al.* 2016): MA0486.2 (HSF1), MA0770.1 (HSF2), MA0771.1 (HSF4) and POL012.1 (TATA-Box).

Results

Characterization of a deletion in UMD3.1.1

A BAC clone retrieved in the BLAST search (CH240-510A19, GenBank accession no. FQ482128.2; Fig. 1a) included the entire *Hsp70* cluster. In turn, this clone was used to retrieve a second overlapping BAC, CH240-369N12 (GenBank accession no. FQ482114.7), which included the gene *NEU1* used as a positional reference in the alignments. Both BACs are part of the CHORI-240 library constructed with DNA from the Hereford bull L1 Domino 99375, which is the sire of L1 Dominette 01449. A reference sequence for further analyses was constructed with both overlapping BAC clones. This reference sequence spanning the interval between genes *NEU1* and *LSM2* comprised from the positions 19 198 to 1 of FQ482114.7 and 51 276 to 23 498 of FQ482128.2 with an overlap of 2001 bases (both BACs are in reverse orientation compared to UMD3.1.1). Two other genomic sequences from a positional cloning experiment in Japanese Holsteins were retrieved from NCBI (GenBank accession nos. AY149618.1 and AY149619.1; Sugimoto *et al.* 2003). The alignment of all those sequences against UMD3.1.1 (Fig. 1a) revealed a deletion with a breakpoint between positions 27 331 772 and 27 331 773 on BTA23 of UMD3.1.1, which correspond to between 27 401 839 and 27 401 840 bp. on Btau 5.0.1. The homozygous deletion had an extension of 11 102 bp and comprised the *HSPA1B* gene and the intergenic region between *HSPA1B* and *HSPA1A*. Surprisingly, sequence AY149619.1 had the same deletion of UMD3.1.1, and it was considered the cause of a disease (hereditary myopathy of diaphragmatic muscles) in Holstein cattle (Sugimoto *et al.* 2003).

In the *Hsp70* cluster, genes *HSPA1A* and *HSPA1L* are 646 bp apart and are arranged in opposite orientation, whereas *HSPA1B* is 9176 bp from *HSPA1A* on the same strand as *HSPA1A* and distal to *HSPA1L* (Fig. 1a).

Independent NGS experiments confirmed the nature of the structural variation detected in UMD3.1.1. Short reads from eight different experiments covered the region between genes *NEU1* and *LSM2* in the reference sequence defined by BACs CH240-369N12 (FQ482114.7) and CH240-510A19 (FQ482128.2). None of those experiments presented the deletion reported in this work, and the genes *HSPA1A* and *HSPA1B* could be correctly distinguished by the two-base substitution on the fifth codon described by Sugimoto *et al.* (2003). We were able to build a consensus sequence between genes *NEU1* and *LSM2* that included the *Hsp70* cluster (data not shown). As an example, pileups of a few selected short reads from a Hereford bull (SRR1365106) are included in Fig. 1.

Paired-end short reads from Dominette (Whitacre *et al.* 2015) aligned concordantly over the deletion breakpoint in UMD3.1.1, whereas the paired reads of the Hereford bull generated discordant alignments (unmapped reads and one-end anchor reads) (Fig. 1b). In turn, when paired-end short reads from Dominette were aligned against BAC CH240-510A19 (FQ482128.2), one-end anchor reads were detected in the region right downstream of *HSPA1A* where the deletion breakpoint is located (Fig. 1d), whereas the same read pairs were aligned in the corresponding region of gene *HSPA1B* (Fig. 1c). When the same comparisons were

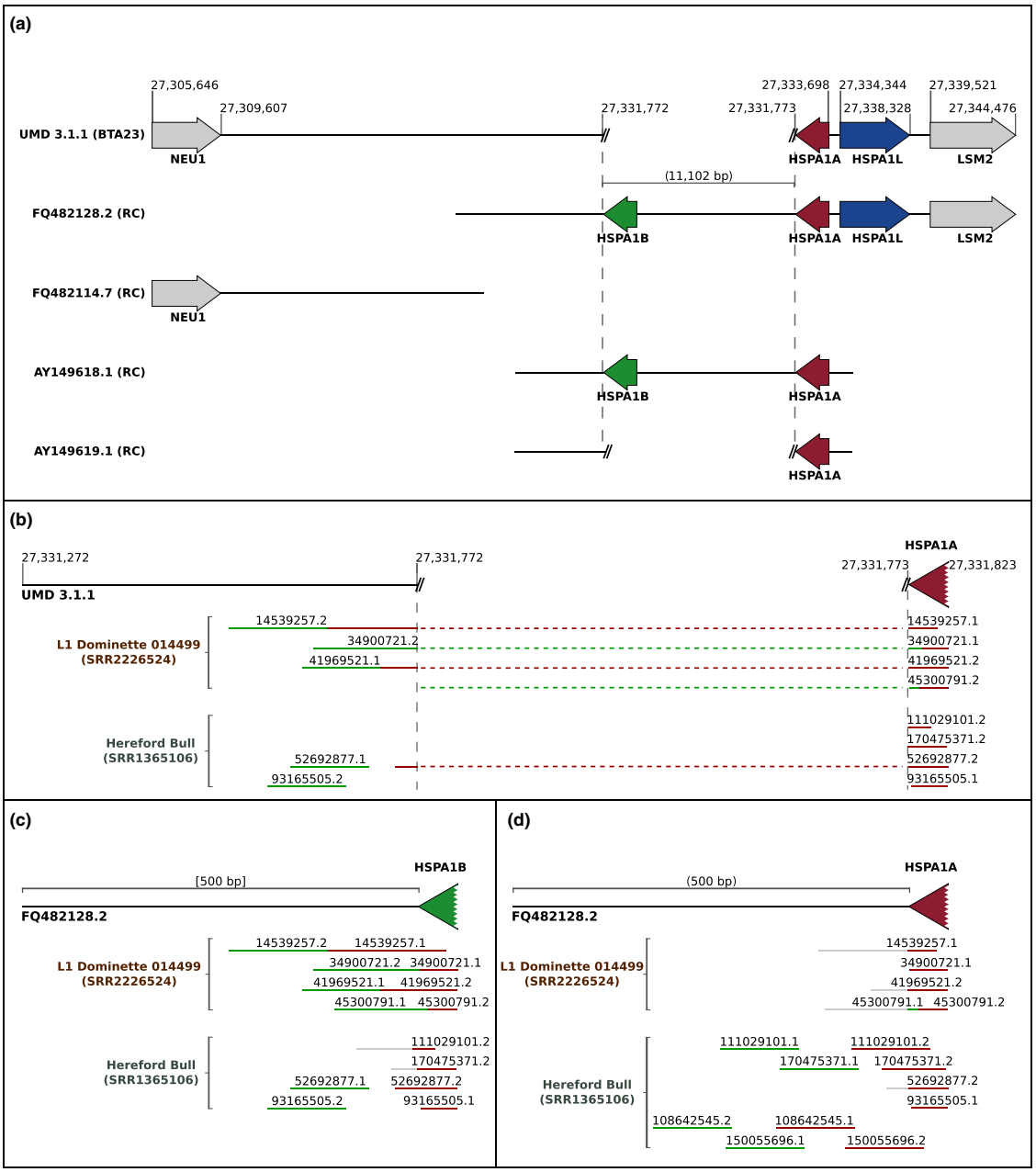

Figure 1 (a) Alignment of the bovine reference genome between genes *NEU1* and *LSM2* on BTA23 with GenBank sequences FQ482128.2, FQ482114.7, AY149618.1 and AY149619.1. The reference genome UMD 3.1.1 has a homozygous deletion of 11 102 bp (region between gray dot lines) spanning gene *HSPA1B* and the intergenic region between this gene and gene *HSPA1A*. The same deletion appears in sequence AY149619.1 (Sugimoto *et al.* 2003). Sequences FQ482128.2 and FQ482114.7 correspond to BAC clones CH240-510A19 and CH240-369N12 respectively from Hereford bull L1 Domino 99375. GenBank accessions and strand orientation (RC: reverse complement) are indicated next to each sequence. Coordinates correspond to UMD3.1.1. (b) Alignment of paired-end short reads from next generation sequencing (NGS) of L1 Dominette 01449 (SRA experiment no. SRX1177177; Whitacre *et al.* 2015) and an unrelated animal (SRA experiment no. SRX582818, run SRR1365106) to the bovine reference genome in the boundaries of the deletion detected in BTA23. Reverse orientated reads are represented with red lines, whereas those forward orientated are in green. Paired-end short reads from Dominette flanking the deletion breakpoint aligned concordantly to the reference genome, whereas paired-end short reads from an independent animal generates one-end anchor reads (e.g. SRR1365106.111029101.2) or discordant pair alignment (e.g. SRR1365106.52692877.1 and 52692877.2). (c-d) Alignment of paired-end short reads from NGS of L1 Dominette 01449 (SRX1177177, Whitacre *et al.* 2015) and an unrelated animal (SRR1365106) to BAC clone CH240-510A19 (FQ482128.2) that includes both *Hsp70* genes (*HSPA1A* and *HSPA1B*). Reverse-orientated reads are represented with red lines, whereas those forward orientated are in green. Gray lines indicate that the read did not align completely. Due to the high sequence homology between both genes, paired-end reads from Dominette are aligned to the 3′ end of gene *HSPA1B* (c), whereas one-end anchor reads from the same pairs aligned to the 3′ end of *HSPA1A* (d). On the contrary, paired-end short reads from SRR1365106 aligned concordantly to the BAC sequence in the boundaries of the 3′ end of both *Hsp70* genes (c–d). However, reads from SRR1365106 that correspond to the coding sequence of one gene could be wrongly assigned to the other.

made with the Hereford bull, paired-end short reads corresponding to *HSPA1A* and *HSPA1B* were indistinguishable if they did not overlap with unique regions either upstream or downstream of the genes. Paired-end short reads from the same animal aligned correctly to the region downstream of *HSPA1A* (Fig. 1d), but due to the high sequence homology between the two genes, short reads corresponding to *HSPA1A* appeared as one-end anchor reads or partially aligned reads that overlapped the 3′ end of *HSPA1B* and vice versa (Fig. 1c, d).

Reassignment of *Hsp70* sequences and SNP locations

The annotations of 25 sequences of *Hsp70* genes retrieved from GenBank were checked. The identity of seven of those sequences that were considered wrongly annotated was corrected, whereas the identities of four of them remain ambiguous (Table 1). Also, a group of SNPs were reassigned to their correct location within the *Hsp70* cluster. However, no SNPs from the 'Chr Unknown' or the 'Chr NotOn' files could be assigned to the Hsp70 cluster region.

From 1393 SNPs mapped to BTA23 that were retrieved from dbSNP, 19 could be unequivocally assigned to the coding sequences of either *HSPA1A* or *HSPA1B* (12 and 7 respectively; Table S1). These reassignments are schematically represented in Fig. 2. Another 124 SNPs matched the sequence of both genes, and their true location could not be solved. As shown below, the promoters of *HSPA1A* and *HSPA1B* are very similar in the proximity of the TSS, and they were not very informative to relocate SNPs at the 5′ end of the genes. According to Sugimoto *et al.* (2003), coding sequences of bovine *HSPA1A* and *HSPA1B* only differ in their fifth codon (ATG and ACA respectively). Given that there is a single *Hsp70* gene in the reference genome, two reported SNPs (rs382492082 and rs385826597) could probably represent the two-base change that supposedly distinguishes one gene from the other, rather than true polymorphisms (Fig. 2).

Table 1 Bovine *Hsp70* sequences retrieved from NCBI using BAC CH240-510A19 (GenBank accession no. FQ482128.2) as a query. The table shows the gene assigned to the sequence in the original submission to GenBank and the corresponding gene (*HSPA1A* or *HSPA1B*), discriminated by their fifth codon (ATG/ACA). Sequences designed as 'ambiguous' do not span the codon that distinguishes the *Hsp70* genes.

Accession no.	GenBank	FQ482128.2
BC105156.1	*Hsp70-2*	Ambiguous
EU038069.1	*Hsp70*	Ambiguous
HF559382.1	*Hsp70*	Ambiguous
U02891.1	*Hsp70-1*	Ambiguous
AY662497.1	*Hsp70*	*HSPA1A*
BC103083.1	*HSPA1B*	*HSPA1A*
HF559383.1	*Hsp70*	*HSPA1A*
JN604432.1	*Hsp70*	*HSPA1A*
NM_174344.1	*HSPA2*	*HSPA1A*
U09861	*Hsp70*	*HSPA1A*
M98823.1	*HSP70A*	*HSPA1B*

Surprisingly, 1247 SNPs mapped to the region between both *Hsp70* genes, which is missing in the reference genome (data not shown). This unexpected finding is partially justified by the result of a comparative alignment of UMD3.1.1 and the reference sequence defined by the two BACs (GenBank accession nos. FQ482114.7 and FQ482128.2). The alignment showed that some fragments of the intergenic region between *HSPA1A* and *HSPA1B* are reinserted between genes *LSM2* and *VWA7* on BTA23 (Fig. S1). Only a fraction of gene *VWA7* is included at the 3′ end of the BAC CH240-510A19 (FQ482128.2), but it was enough to define the locations of the reinsertions and also to show that this gene is in reverse orientation in UMD3.1.1 compared to the BAC and the reference genomes of human (GRCh38.p7) and mouse (GRCm38.p5) in the Ensembl Genome Browser.

Comparative analysis of *HSPA1A* and *HSPA1B* promoters

Comparative promoter analysis across ruminant species in the first 500 bp upstream of the TSS revealed high sequence homology between both genes, but to a lower extent than the coding regions (Fig. S2). The first 404 bp upstream of the TSS are highly conserved among ruminant species. However, the promoters of *HSPA1A* in cow, bison and yak have a deletion of 31 bp located at –77 bp, not seen in the other species, whereas only the cow had an insertion of four bases at –253 bp. Also, there is an indel of four bases located at –440 bp that distinguishes both genes in all species.

In all promoter sequences, a canonical TATA box is located between –223 and –232 bp (or –190 and –200 bp in those species with the 31-bp deletion described above). When the promoters of both genes were searched for transcription factor binding sites (TFBSs) corresponding to heat shock elements, one TFBS for HSF1 (heat shock factor 1) was identified in the promoters of *HSPA1A* (−391 or −358 bp in those species with the 31-bp deletion described above), whereas the promoters of *HSPA1B* had two TFBSs for HSF2 in all analysed species (approximately −395 and −467 bp respectively) (Fig. 3).

Discussion

All the research reported here was conducted to solve discrepancies detected in the locations of SNPs previously described by Rosenkrans *et al.* (2010) in the promoter of one of the *Hsp70* genes when compared to the bovine reference genome in the Ensembl Genome Browser (http://www.ensembl.org/index.html). Surprisingly, not only was there no agreement in SNP locations but also the entire sequence M98823.1 used by Rosenkrans *et al.* (2010) was

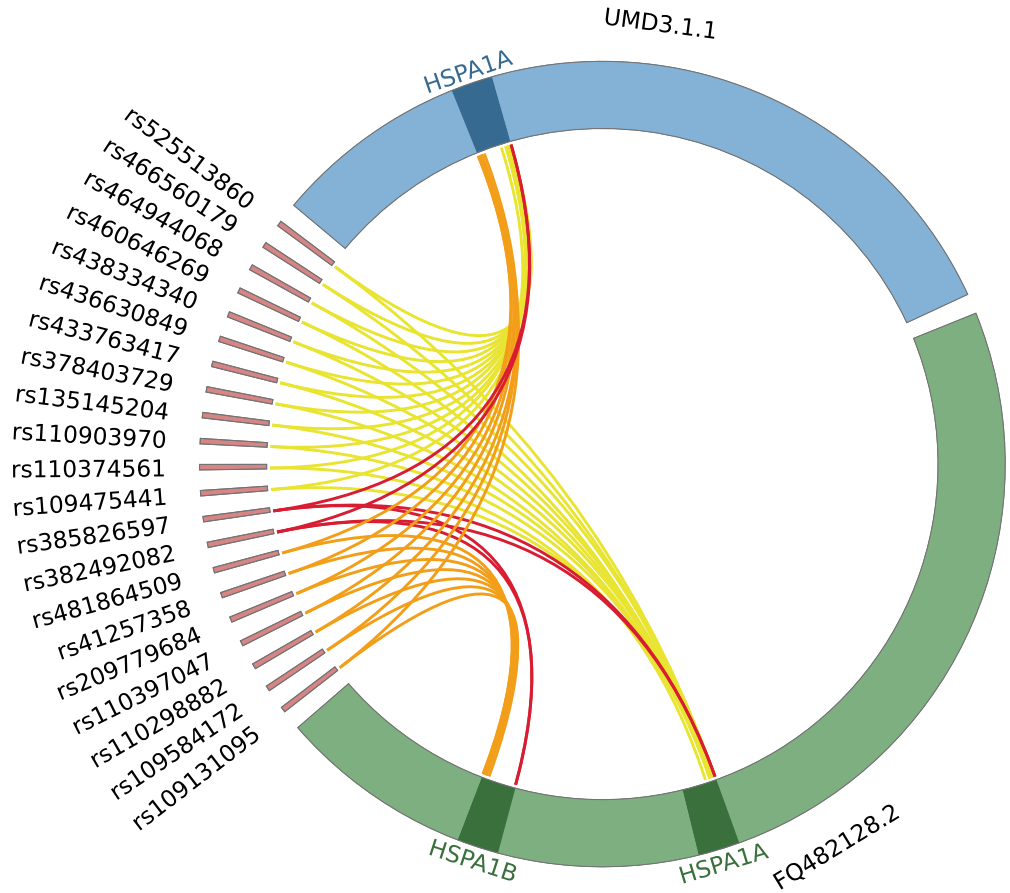

Figure 2 Relocation of SNPs from BTA23 on *HSPA1A* and *HSPA1B* genes. Sequence FQ4892128.2 corresponds to BAC CH240-510A19, which includes both *Hsp70* genes. From 137 SNPs reported in Ensembl on a single *Hsp70* gene, only 19 SNPs could be unequivocally reassigned to either *HSPA1A* (yellow lines) or *HSPA1B* (orange lines) due to the high homology of their sequences. Original and new coordinates for these SNPs are presented in Table S1. The locations of the other SNPs remain ambiguous. Red lines correspond to two SNPs that overlap the fifth codon of *HSPA1A*; their alleles agree with the two-base substitution that allegedly distinguishes both genes (Sugimoto *et al.* 2003). This figure was made using CIRCOS software version 0.69-3 (Krzywinski *et al.* 2009).

missing. It is worth noting that Rosenkrans *et al.* (2010) referred to a single *Hsp70* gene, and also a single gene on BTA23 was found in Ensembl (ENSBTAG00000025441).

We originally attributed the existence of a deletion in BTA23 to assembly errors. Assemblies can collapse around repetitive sequences (Salzberg & Yorke 2005), and this is one of the major challenges of genome sequencing (Treangen & Salzberg 2011). In the present case, the three *Hsp70* genes have a high degree of homology and genes *HSPA1A* and *HSPA1B* are almost identical. However, further analysis based on NGS of the cow L1 Dominette 01449 suggested that rather than an assembly error, the deletion could be a particular condition of the genome of this animal (Fig. 1). Following a positional strategy, Sugimoto *et al.* (2003) mapped the locus responsible for hereditary myopathy of diaphragmatic muscles (HMDM) in Holstein cattle to the Hsp70 region on BTA23. Comparative DNA sequencing revealed a deletion common to affected animals (GenBank accession nos. AY149619.1 for cases and AY149618.1 for controls). One of the most surprising results of this work was that the breakpoint for the deletion in the genome of L1 Dominette 01449 was in the same location as the breakpoint in sequence AY149619.1 (Fig. 1); in fact, the alignment of AY149619.1 (8,566 bp) to UMD 3.1.1 resulted in 96 percent coverage and 99 percent identity. Possible explanations for the repeated appearance of a deletion at a given position are the existence of a hotspot for breakage and recombination in this region and/or the highly frequent occurrence of unequal crossovers. The frequency of the deletion reported in the present work in cattle populations is currently unknown and warrants further research.

It appears that this deletion of one *Hsp70* paralog is not necessarily associated with any discernible disease in cattle. Deletions are indeed a cause of productive and reproductive problems in cattle (for example, a 500-kb deletion in BTA23 is a strong candidate for stillbirth; Sahana *et al.* 2016). However, there is no reported evidence of any health

 doi: 10.1111/age.12561

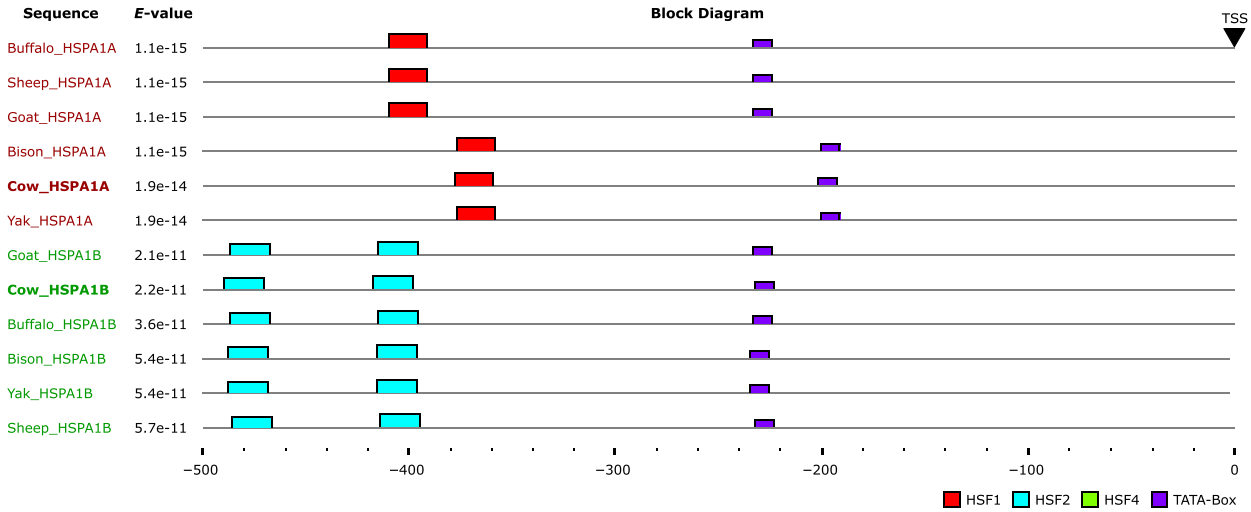

Figure 3 Comparative analysis of transcription factor binding sites in the promoters of *HSPA1A* and *HSPA1B* using the MAST tool of the MEME SUITE (Bailey *et al.* 2015) with motifs from the JASPAR 2016 public database: MA0486.2 (HSF1), MA0770.1 (HSF2), MA0771.1 (HSF4) and POL012.1 (TATA-Box). The promoters of *HSPA1A* and *HSPA1B* genes showed different composition of HSF, whereas these differences are conserved across ruminant species.

problems in the Hereford cow L1 Dominette 01449. Inconsistencies in the association between the deletion in the *Hsp70* cluster and the HMDM phenotype described by Sugimoto *et al.* (2003) could be due to background effects in different breeds. Alternatively, HMDM in Holstein could be caused by a mutation other than the reported deletion.

Evidence from knock-out mice suggests that the loss of *HSPA1A* has more severe effects over that of *HSPA1B*, for which only a higher sensitivity to heat stress has been reported (Daugaard *et al.* 2007). Lan & Pritchard (2016) proposed a model in which duplicated genes can undergo down-regulation to match required expression levels, generating 'asymmetrically expressed duplicates'. In any case, whether the loss of one *Hsp70* gene, particularly *HSPA1B*, has noticeable effects on cattle health still remains to be elucidated.

The existence of incorrectly collapsed repeats complicates the identification of true polymorphisms (Salzberg & Yorke 2005). This effect could be extended to the loss of one of two highly homologous genes, as it seems to be the present case. We were able to reassign a group of SNPs, already mapped to BTA23, to their right location in the *Hsp70* genes (Fig. 2, Table S1). Moreover, the results of the present work showed that the gene *VWA7* (which is downstream of the *Hsp70* cluster) is in reverse orientation in the reference genome UMD 3.1.1 when compared to other cattle and also to human, mouse, sheep and goat (Fig. S1). Although a misassembly in UMD3.1.1 could not be ruled out, this result would suggest that an extensive region of BTA23 beyond the *Hsp70* cluster was part of a chromosomal rearrangement.

As mentioned above, variations in the *Hsp70* genes have been associated with multiple production traits (Brown *et al.* 2010; Deb *et al.* 2013) and diseases (Han *et al.* 2009). However, in some cases, it is not possible to identify which one of the genes (*HSPA1A* or *HSPA1B*) was the subject of study. For example, Rosenkrans *et al.* (2010) reported SNPs significantly associated with reproductive traits in cows in the promoter of a *Hsp70* gene that turned out to be *HSPA1B* (GenBank sequence M98823.1; Table 1). Also, Deb *et al.* (2013) conducted an association study to assess the effect of a promoter indel on the expression of an *Hsp70* gene in blood cells of Frieswal crossbred cattle. Based on the sequences of the corresponding primers, the genotyped indel mapped to the promoter of *HSPA1B*, whereas the RT-PCR system designed to quantify expression targeted both genes. In the same work, the deletion of a cytosine in the promoter of *HSPA1B* was associated with higher body temperature and lower milk production.

We found conserved differences between the promoter sequences of *HSPA1A* and *HSPA1B* across ruminant species (Fig. S2). Furthermore, promoters showed different kinds of HSFs for both genes in all the analysed species (HSF1 in *HSPA1A* and HSF2 in *HSPA1B*); also, *HSPA1A* had only one HSF binding site, whereas *HSPA1B* had two of them (Fig. 3). These results justify the analysis of differential regulation of expression between the two genes. However, our findings do not agree with those from a study conducted by Garbuz *et al.* (2011) in which the comparison of the regulatory regions of *Hsp70* promoters in a more diverse group of mammal species, including the bovine, showed HSFs located in a similar position with respect to the TSS in both *HSPA1B* and *HSPA1A* promoter regions, suggesting a similar regulation under heat stress conditions.

Some *Hsp70* expression analyses have been conducted in bovine tissues (e.g. blood cells; Parmar *et al.* 2015 and

embryos; Khan *et al.* 2016), but these studies did not discriminate between *HSPA1A* and *HSPA1B* expression. In fact, based on the sequence of the primers that were used, in some cases they could have potentially targeted both genes.

Brocchieri *et al.* (2008) estimated the relative expression of *Hsp70* genes in different human tissues by expressed sequence tag counting in the NCBI UniGene database; in most cases, the expression of *HSPA1A* was higher than that of *HSPA1B*. Maugeri *et al.* (2010) discriminated between the expression of *HSPA1A* and *HSPA1B* in human lymphoblastoid cell lines following heat shock. They identified a strong cis-acting eQTL, affecting *HSPA1B* only, located at least 62 kb telomeric to the gene. All these results point to a possible differential expression of *Hsp70* genes due to factors other than heat stress, which could also be tissue specific.

Given the extent of linkage disequilibrium in cattle (McKay *et al.* 2007), a marker on any of the *Hsp70* genes could probably tag haplotypes spanning *HSPA1A* and *HSPA1B*, but it would be of interest to clarify whether one or both genes are responsible for potential significant effects detected in association studies, with different populations and varying experimental conditions.

In summary, we have improved the annotation of a small but very important region of the bovine genome. Inconsistencies in a reference genome have consequences downstream in the research process, given that it is an essential resource for genome sequencing of new individuals. Also, the conserved organization of the *Hsp70* cluster in mammals and its significant role in animal health and production warrant further research. The information reported here could be helpful on that regard.

Acknowledgements

M.F.S.G. gratefully acknowledges receipt of a fellowship from CONICET (Buenos Aires, Argentina). This research was conducted in partial fulfilment of the requirements for the PhD degree at the Universidad Nacional de Mar del Plata (UNMdP). The assistance of the Bioinformatics Unit (Instituto de Biotecnología, CICVyA, INTA) is greatly appreciated.

References

Bailey T.L., Boden M., Buske F.A., Frith M., Grant C.E., Clementi L., Ren J., Li W.W. & Noble W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* **37**, W202–8.

Bailey T.L., Johnson J., Grant C.E. & Noble W.S. (2015) The MEME SUITE. *Nucleic Acids Research* **43**, W39–49.

Brocchieri L., de Macario E.C. & Macario A.J.L. (2008) hsp70 genes in the human genome: conservation and differentiation patterns predict a wide array of overlapping and specialized functions. *BMC Evolutionary Biology* **8**, 19.

Brown A.H. Jr, Reiter S.T., Brown M.A., Johnson Z.B., Nabhan I.A., Lamb M.A., Starnes A.R. & Rosenkrans C.F. Jr (2010) Effects of heat shock protein-70 gene and forage system on milk yield and composition of beef cattle. *The Professional Animal Scientist* **26**, 398–403.

Darling A.E., Mau B. & Perna N.T. (2010) PROGRESSIVEMAUVE: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* **5**, e11147.

Daugaard M., Rohde M. & Jäättelä M. (2007) The heat shock protein 70 family: highly homologous proteins with overlapping and distinct functions. *FEBS Letters* **581**, 3702–10.

Deb R., Sajjanar B., Singh U., Kumar S., Brahmane M.P., Singh R., Sengar G. & Sharma A. (2013) Promoter variants at AP2 box region of Hsp70.1 affect thermal stress response & milk production traits in Frieswal cross bred cattle. *Gene* **532**, 230–5.

Erb I., González-Vallinas J.R., Bussotti G., Blanco E., Eyras E. & Notredame C. (2012) Use of ChIP-Seq data for the design of a multiple promoter-alignment method. *Nucleic Acids Research* **40**, e52.

Garbuz D.G., Astakhova L.N., Zatsepina O.G., Arkhipova I.R., Nudler E. & Evgen'ev M.B. (2011) Functional organization of hsp70 cluster in camel (*Camelus dromedarius*) and other mammals. *PLoS One* **6**, e27205.

Grosz M.D., Womack J.E. & Skow L.C. (1992) Syntenic conservation of HSP70 genes in cattle and humans. *Genomics* **14**, 863–8.

Han J., Li Q., Wang C., Wang H., Li J., Zhong J. & Pan Q. (2009) A new SNP in coding region of HSP70 gene and the association of polymorphism with heat stress traits in Chinese Holstein cattle. *Journal of Agricultural Science and Technology* **11**, 56–63.

Huang P., Lu C., Li J. *et al.* (2015) Mutations in *HSP70-2* gene change the susceptibility to clinical mastitis in Chinese Holstein. *Gene* **559**, 62–72.

Khan I., Lee K.-L., Fakruzzaman M. *et al.* (2016) Coagulansin-A has beneficial effects on the development of bovine embryos *in vitro* via HSP70 induction. *Bioscience Reports* **36**, 1186–97.

Kishore A., Sodhi M., Kumari P. *et al.* (2014) Peripheral blood mononuclear cells: a potential cellular system to understand differential heat shock response across native cattle (*Bos indicus*), exotic cattle (*Bos taurus*), and riverine buffaloes (*Bubalus bubalis*) of India. *Cell Stress and Chaperones* **19**, 613–21.

Krzywinski M., Schein J., Birol I., Connors J., Gascoyne R., Horsman D., Jones S.J. & Marra M.A. (2009) CIRCOS: an information aesthetic for comparative genomics. *Genome Research* **19**, 1639–45.

Lan X. & Pritchard J.K. (2016) Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science* **352**, 1009–13.

Langmead B. & Salzberg S.L. (2012) Fast gapped-read alignment with BOWTIE 2. *Nature Methods* **9**, 357–9.

Li H., Handsaker B., Wysoker A. *et al.* (2009) The sequence alignment/map format and SAMTOOLS. *Bioinformatics* **25**, 2078–9.

Malinverni D., Marsili S., Barducci A., Rios P.D.L. & de Los Rios P. (2015) Large-scale conformational transitions and dimerization are encoded in the amino-acid sequences of hsp70 chaperones. *PLoS Computational Biology* **11**, 1–15.

Mathelier A., Fornes O., Arenillas D.J. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **44**, D110–5.

Maugeri N., Radhakrishnan J. & Knight J.C. (2010) Genetic determinants of HSP70 gene expression following heat shock. *Human Molecular Genetics* **19**, 4939–47.

McKay S.D., Schnabel R.D., Murdoch B.M. *et al.* (2007) Whole genome linkage disequilibrium maps in cattle. *BMC Genetics* **8**, 74.

Parmar M.S., Madan A.K., Rastogi S.K. & Huozha R. (2015) Expression of *heat shock protein 70-1* gene in bovine peripheral blood mononuclear cells. *Indian Journal of Animal Research* **49**, 325–7.

Picard B., Gagaoua M., Micol D., Cassar-Malek I., Hocquette J.-F. & Terlouw C.E.M. (2014) Inverse relationships between biomarkers and beef tenderness according to contractile and metabolic properties of the muscle. *Journal of Agricultural and Food Chemistry* **62**, 9808–18.

R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*. Vienna: The R Foundation for Statistical Computing.

Rosenkrans C., Banks A., Reiter S. & Looper M. (2010) Calving traits of crossbred Brahman cows are associated with *heat shock protein 70* genetic polymorphisms. *Animal Reproduction Science* **119**, 178–82.

Sahana G., Iso-Touru T., Wu X., Nielsen U.S., de Koning D.-J., Lund M.S., Vilkki J. & Guldbrandtsen B. (2016) A 0.5-Mbp deletion on bovine chromosome 23 is a strong candidate for stillbirth in Nordic Red cattle. *Genetics Selection Evolution* **48**, 35.

Salzberg S.L. & Yorke J.A. (2005) Beware of mis-assembled genomes. *Bioinformatics* **21**, 4320–1.

Sugimoto M., Furuoka H. & Sugimoto Y. (2003) Deletion of one of the duplicated *Hsp70* genes causes hereditary myopathy of diaphragmatic muscles in Holstein-Friesian cattle. *Animal Genetics* **34**, 191–7.

Treangen T.J. & Salzberg S.L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13**, 36–46.

Whitacre L.K., Tizioto P.C., Kim J. *et al.* (2015) What's in your next-generation sequence data? An exploration of unmapped DNA and RNA sequence reads from the bovine reference individual *BMC Genomics* **16**, 1114.

Supporting information

Additional supporting information may be found online in the supporting information tab for this article:

Figure S1 Alignment of the bovine reference genome between genes *NEU1* and *VWA7* (27 305 646–27 365 002 bp) on chromosome 23 (bottom) against a reference sequence defined by BAC clones CH240-369N12 (GenBank accession no. FQ482114.7) and CH240-510A19 (GenBank accession no. FQ482128.2) (top), using MAUVE software version 2.4.0. Homologous regions are indicated by boxes of the same colour. In the reference genome, fragments corresponding to the intergenic region between *HSPA1A* and *HSPA1B* are reinserted between *LSM2* and *VWA7*, whereas gene *HSPA1B* is missing. Note that *VWA7* is in reverse orientation in the reference genome UMD 3.1.1.

Figure S2 Alignment of ruminant *HSPA1A* and *HSPA1B* promoters (500 bp upstream of the TSS) made with PROCOFFEE (Erb *et al.* 2012). The red boxes highlight the major differences. Numbers indicate the position with respect to the TSS.

Table S1 Location of the SNPs that could be unequivocally assigned to *HSPA1A* or *HSPA1B* coding regions on BTA23.