A false single nucleotide polymorphism generated by gene duplication compromises meat traceability

Arianne Sanz a,⁎, Laura Ordovás a, Pilar Zaragoza a, Albina Sanz b, Ignacio de Blas c, Clementina Rodellar a

a Laboratorio de Genética Bioquímica (LAGENBIO), Facultad de Veterinaria, Universidad de Zaragoza, Spain
b Área de Producción, Centro de Investigación y Tecnología Agroalimentaria de Aragón, Spain
c Unidad de Patología Infecciosa y Epidemiología, Facultad de Veterinaria, Universidad de Zaragoza, Spain

A R T I C L E  I N F O

Article history:
Received 26 April 2011
Received in revised form 17 January 2012
Accepted 14 February 2012

Keywords:
SNP
Meat traceability
Genetic identification
False SNP
Duplication gene
Excess of heterozygote

A B S T R A C T

Controlling meat traceability using SNPs is an effective method of ensuring food safety. We have analyzed several SNPs to create a panel for bovine genetic identification and traceability studies. One of these was the transversion g.329C>T (Genbank accession no. AJ496781) on the cytochrome P450 17A1 gene, which has been included in previously published panels. Using minisequencing reactions, we have tested 701 samples belonging to eight Spanish cattle breeds. Surprisingly, an excess of heterozygotes was detected, implying an extreme departure from Hardy–Weinberg equilibrium (P<0.001). By alignment analysis and sequencing, we detected that the g.329C>T SNP is a false positive polymorphism, which allows us to explain the inflated heterozygotic value. We recommend that this ambiguous SNP, as well as other polymorphisms located in this region, should not be used in identification, traceability or disease association studies. Annotation of these false SNPs should improve association studies and avoid misinterpretations.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

In recent years, several food-related crises have shaken the trust of consumers in the safety of their food supply. The outbreak of bovine spongiform encephalopathy (BSE) caused a dramatic decline in the consumption of beef and was one of the factors that spurred the development of traceability systems across several countries (Souza-Monteiro & Caswell, 2004). To quickly isolate a future outbreak of BSE or other food related crisis, it is important to improve traceability systems, ensure food safety and connect all the links in the food chain; to allow consumers to distinguish beef coming from specific regions (Jin, Skripnitchenko, & Koo, 2004). A DNA based methodology is the most reliable way to assure traceability along the meat transformation process and is frequently used as it represents a viable means for verifying the accuracy of physical labels for the identification of cattle and other species (Goffaux, China, Dams, Clinquart, & Daube, 2005; Heaton et al., 2005; Vazquez et al., 2004). DNA technology also allows auditing of the fidelity of conventional traceability systems by identifying violations and removing condemned carcasses from the human food chain, thereby raising the confidence of consumers and the meat industry (Heaton et al., 2005).

Microsatellites have been most commonly used as genetic markers due to their high degree of polymorphism (Vignal, Milan, San Cristobal, & Eggen, 2002). However, single nucleotide polymorphisms (SNPs), the most recent tool for studying DNA variation, have gained more popularity because of their advantages and are rapidly replacing microsatellites for genetic identification purposes. Despite their lower degree of polymorphism, SNPs have lower rates of genotyping errors (Kennedy et al., 2003) and are amenable to automation and high throughput genotyping. SNPs are also associated with some human disorders (Peden et al., 2011) and with several economically important traits in other species (Fan et al., 2011; Sanz et al., 2010).

Several studies have focused on developing an adequate panel of SNPs to carry out genetic identification for paternity testing or traceability in cattle, and multiple SNP panels now exist for this species (Heaton et al., 2002; Karniol et al., 2009; Werner et al., 2004). In our project we have analyzed several SNPs (http://www.unizar.es/lagenbio/servicios/SNPs.pdf) to create a panel for bovine genetic identification and traceability studies. One of the markers proposed by Werner et al. (2004) was the putative transversion g.329C>T (Genbank accession no. AJ496781) located at the seventh intron of the cytochrome P450 17A1 gene (CYP17A1).

CYP17A1 is a protein with 17α-hydroxylase and 17, 20-lyase activities (Steroid 17-α-hydroxylase/17, 20 lyase) that plays a pivotal role in the biosynthesis of steroid hormones. CYP17A1 is involved in the conversion of pregnenolone and progesterone to their respective 17-hydroxy metabolites, as well as the subsequent conversion of...
these intermediates into dehydroepiandrosterone or androstenedi-one, the precursors of estrone and testosterone (Zuber, Simpson, & Waterman, 1986). In humans, the CYP17A1 gene has been associated with both prostate and breast cancer (Kaufman et al., 2011; Sarma et al., 2008; Severi et al., 2008).

The bovine gene maps to chromosome 26 (Gautier et al., 2001) and its organization is similar to that of the human gene on chromosome 10 (Blaskey et al., 1989). CYP17A1 belongs to the cytochrome P450 (CYP450) superfamily, which consists of 50–80 CYP450 genes in vertebrate genomes (Thomas, 2007), and numerous duplications of members of this superfamily have been previously reported (Lundqvist, Johansson, & Ingelman-Sundberg, 1999; Thomas, 2007).

Ho et al. (Ho, Tsai, Chen, & Lin, 2011) recently reported that more than 10% of human genes have been associated with duplicated loci, which must be taken into account when depending on information from such genes. Gene duplication is considered to be a major contributor to genome evolution (Storz, 2009). Although mutations accumulate gradually in duplicated genes, some recently duplicated regions may retain a higher degree of sequence homology. Consequently, the genotype data derived from these variants located in duplicated genes can resemble the genotyping results of simple nucleotide polymorphisms (Ho et al., 2011). To not identify a single-base difference, located in nearly identical duplicated segments, as a single nucleotide polymorphism, a duplicated gene nucleotide variants database has been recently created for humans (Ho et al., 2011).

To implement a suitable genetic traceability system for protecting public health and assuring food safety, false SNPs should be avoided.

Here, we report that the g.329C>T polymorphism in the P45017A1 gene is an example of an ambiguous SNP that has been previously included in some SNP panels used for traceability studies.

2. Materials and methods

2.1. Samples

A total of 701 DNA samples were obtained from animals from eight cattle breeds reared in Spain (Asturiana de los Valles, n = 23; Frisona, n = 23; Menorquina, n = 22; Limousine, n = 23; Monchina, n = 23; Serrana Negra de Teruel, n = 23; Parda de Montaña, n = 257 and Pirenaïca, n = 307).

2.2. Genotyping

The putative g.329C>T SNP (Genbank accession no. AJ496781) was genotyped by minisequencing reactions in all samples. From the genotype data, we have estimated allelic frequencies and found that the frequency for allele C was 0.5 in all populations, except for Monchina and Parda de Montaña breeds where it was 0.522 and 0.503, respectively. Similar frequencies were reported previously by Werner et al. (Werner et al., 2004) in German Fleckvieh, German Braunvieh and German Holstein dairy breeds. Frequencies close to 0.5 are the most suitable for inclusion in a SNP panel designed for traceability purposes (Capoferrri, Bongioni, Galli, & Aleandri, 2006).

2.3. Statistical calculation

Allelic frequencies were obtained by direct counting using the web version of GenePop ( Raymond & Rousset, 1995; Rousset, 2008). This program was also used to test departures from Hardy-Weinberg equilibrium (Van Laere et al., 2003) using a Markov chain method to estimate the P-values. A P-value lower than 0.05 was considered significant.

2.4. Alignment and DNA sequence analysis

Basic Local Alignment Search Tool (BLAST) (http://www.ncbi.nlm.nih.gov/blast/BLAST.cgi) alignment was performed between the sequence containing the SNP (AJ496781) and the Bos Taurus genome using the megabLAST tool present in the NCBI database.

To determine if a duplicated DNA segment exists, specific fragments were amplified by PCR using primers 4, 5 and 6 designed according to the sequences obtained by BLAST (see Table 1). Amplicons were purified by the enzymatic method ExoSAP-IT (USB, Cleveland, OH, USA); bidirectionally sequenced using the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA) and purified by the BigDye X-Terminator Purification Kit (Applied Biosystems, Foster City, CA, USA). Finally, sequencing reactions were run on an ABI Prism® 3130 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA).

3. Results and discussion

3.1. Genotyping and allelic frequencies

The putative g.329C>T SNP (AJ496781) was genotyped by minisequencing reactions in all samples. From the genotype data, we have estimated allelic frequencies and found that the frequency for allele C was 0.5 in all populations, except for Monchina and Parda de Montaña breeds where it was 0.522 and 0.503, respectively. Similar frequencies were reported previously by Werner et al. (Werner et al., 2004) in German Fleckvieh, German Braunvieh and German Holstein dairy breeds. Frequencies close to 0.5 are the most suitable for inclusion in a SNP panel designed for traceability purposes (Capoferrri, Bongioni, Galli, & Aleandri, 2006).

Table 1

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Primer sequence</th>
<th>Nucleotide position within AJ496781</th>
<th>Nucleotide position within NW_001503883</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Forward</td>
<td>TTGACAAAAGCCACAGCGTGT</td>
<td>218–238</td>
<td>28562–28582</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>47289–47309</td>
</tr>
<tr>
<td>2. Reverse</td>
<td>TGGGATGCAGGAGTCTTAGCA</td>
<td>393–415</td>
<td>28385–28407</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>47112–47134</td>
</tr>
<tr>
<td>3. Minisequencing</td>
<td>(34)GCTTCTTGTCTGCC</td>
<td>309–328</td>
<td>28472–28491</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>47199–47218</td>
</tr>
<tr>
<td>4. Forward 3′ flanking region LOC784256</td>
<td>TACCTGGAAAAACATCCCGG</td>
<td>27129–27148</td>
<td></td>
</tr>
<tr>
<td>5. Forward 3′ flanking region LOC784299</td>
<td>GTCCCTTGTATTTTCCC</td>
<td>45787–45807</td>
<td></td>
</tr>
<tr>
<td>6. Common reverse intron 6–exon 7</td>
<td>GAGGACACATTCGGGACCTTA</td>
<td>28605–28686</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>47332–47313</td>
</tr>
</tbody>
</table>
3.2. Hardy–Weinberg equilibrium

All groups of samples showed a significantly high departure from Hardy–Weinberg equilibrium ($P<0.001$) as a consequence of 100% heterozygosity in most of the analyzed breeds, except for Monchina and Parda de Montaña breeds with 95.6% and 99.6% heterozygosity, respectively. To elucidate the high deviation from HWE, different assays were utilized because false SNPs located in duplicated regions can inflate the heterozygosity of SNPs (Ho et al., 2011).

3.3. BLAST

Taking into account this significant disequilibrium, we performed a BLAST search between the described sequence containing the putative g.329C>T SNP (AJ496781) and the Bos Taurus genome database. BLAST results showed matching of AJ496781 fragment on five contigs (NW_001503883.1, NW_001503027.1, NW_001494364.1, NW_001494358.2 and NW_003104582.1). We identified the duplication of AJ496781, which aligned twice on the contig with GenBank accession number NW_001503883, (Region A from 28164 to 28799 and Region B from 46891 to 475226). Regions A and B showed either a C or T at positions g.28527 and g.47254, respectively (Fig. 1).

These two regions fit in loci LOC784256 and LOC784299, which are 6017 bp and 1161 bp long, respectively. These loci are 98% identical to each other and both are similar to the bovine CYP17A1 gene located on the minus chain of bovine chromosome 26.

The adjacent region of the two loci (LOC784256 and LOC784299) showed only 32% of query coverage, with 89% identity. We took advantage of these differences to design specific primers to amplify the two loci to demonstrate the duplication phenomenon.

3.4. Sequencing analysis

To confirm the duplication observed by BLAST, we designed two locus-specific forward primers (primers 4 and 5, Table 1) in the non-homologous 3’contiguous region to loci LOC784256 and LOC784299. A common reverse primer (primer 6, Table 1) was designed in the analogous region. Fig. 2 shows the alignment between the two regions of contig NW_001503883 containing AJ495781, the specific primers used and the relative position of the false SNP.

As a result of the PCR strategy employed, two different fragments were amplified showing a high identity with contig NW_001503883. Furthermore, a C or T base at positions g.28527 and g.47254, also seen in NW_001503883, were found. These results demonstrate that a duplicated segment of DNA is present on bovine chromosome 26.

The possible conflicting results derived from the matching of AJ496781 with the other contigs could be explained by the size of contigs and by some inconsistencies in the Btau 4.2 and UMD3.1 alternate assemblies (Partipilo, D’Addabbo, Lacalandra, Liu, & Rocchi, 2011).

Taken together, our results support the existence of two fragments, containing either a C or T, at the relative position of the putative SNP. These results suggest that the g.329C>T (AJ496781) SNP reported is actually a single base difference located in nearly identical duplicated segments and not a SNP.

As there is no database on domestic animals containing this kind of information, as ambiguous results or duplication, we promote the

---

**Fig. 1.** Evidence of alignment between AJ496781 and contig with the accession number NW_001503883. Image from NCBI Map Viewer shows identity with two regions corresponding to loci LOC784256 and LOC784299.
establishment of an informative website where researchers could add their own examples (http://www.unizar.es/lagenbio/investigacion/AmbiguousSNPs.pdf). Annotating these results could prevent the misinterpretations of false positive SNPs, avoiding that such information be forgotten.

Given the characteristics and complexity of the CYP450 superfamily and the fact that numerous members are known to be duplicated in vertebrate genomes (Lundqvist et al., 1999; Thomas, 2007), further investigations will be required to elucidate the origin of the LOC784256 and LOC784299 loci, and it will be essential to determine their functional roles.
whether these duplicates exist in humans. Our findings also explain the high deviation from HWE shown in our data as a consequence of a putative heterozygote excess.

4. Conclusions

As ambiguous SNPs could jeopardize meat traceability, we strongly recommend that the putative transversion g.329C>T (AJ496781) or any presumed SNP located within this region should not be used in identification, traceability studies or disease associations due to the gene duplication. Because there is currently no database for annotating ambiguous SNPs in domestic species, as there is in humans, it may be important to publish findings of this kind. Preventing the use of false positive SNPs should improve the association studies of these variants and, as a consequence, improve human health and safety.

Acknowledgments

This work was supported by projects PET2007-01-C07-04, PET2007-05-C03-01 and RZ2006-00003-C02-02.

References


