



ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA

## ARCHIVIO ISTITUZIONALE DELLA RICERCA

### Alma Mater Studiorum Università di Bologna Archivio istituzionale della ricerca

Application of next generation semiconductor based sequencing for species identification in dairy products

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

Ribani, A., Schiavo, G., Utzeri, V.J., Bertolini, F., Geraci, C., Bovo, S., et al. (2018). Application of next generation semiconductor based sequencing for species identification in dairy products. *FOOD CHEMISTRY*, 246, 90-98 [10.1016/j.foodchem.2017.11.006].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/620213> since: 2018-02-07

*Published:*

DOI: <http://doi.org/10.1016/j.foodchem.2017.11.006>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

Anisa Ribani, Giuseppina Schiavo, Valerio Joe Utzeri, Francesca Bertolini, Claudia Geraci, Samuele Bovo, Luca Fontanesi, *Application of next generation semiconductor based sequencing for species identification in dairy products*, Food Chemistry, Volume 246, 2018, Pages 90-98, ISSN 0308-8146, <https://www.sciencedirect.com/science/article/pii/S0308814617318083>

The final published version is available online at: <https://doi.org/10.1016/j.foodchem.2017.11.006>.

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

**When citing, please refer to the published version.**

## Accepted Manuscript

Application of next generation semiconductor based sequencing for species identification in dairy products

Anisa Ribani, Giuseppina Schiavo, Valerio Joe Utzeri, Francesca Bertolini, Claudia Geraci, Samuele Bovo, Luca Fontanesi

PII: S0308-8146(17)31808-3

DOI: <https://doi.org/10.1016/j.foodchem.2017.11.006>

Reference: FOCH 21983

To appear in: *Food Chemistry*

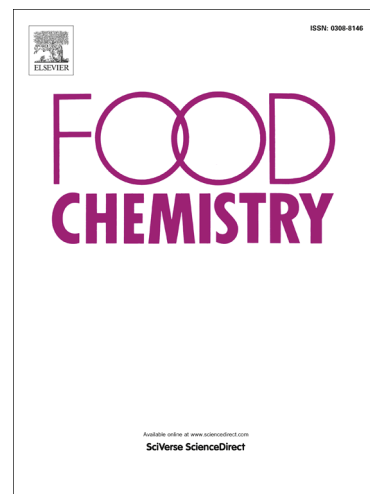
Received Date: 12 May 2017

Revised Date: 26 October 2017

Accepted Date: 2 November 2017

Please cite this article as: Ribani, A., Schiavo, G., Joe Utzeri, V., Bertolini, F., Geraci, C., Bovo, S., Fontanesi, L., Application of next generation semiconductor based sequencing for species identification in dairy products, *Food Chemistry* (2017), doi: <https://doi.org/10.1016/j.foodchem.2017.11.006>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



1 **Application of next generation semiconductor based sequencing for species identification in**  
2 **dairy products**

3

4 Anisa Ribani<sup>a</sup>, Giuseppina Schiavo<sup>a</sup>, Valerio Joe Utzeri<sup>a</sup>, Francesca Bertolini<sup>a,b</sup>, Claudia Geraci<sup>a</sup>,  
5 Samuele Bovo<sup>a,c</sup>, Luca Fontanesi<sup>a,\*</sup>

6

7 <sup>a</sup> Department of Agricultural and Food Sciences (DISTAL), Division of Animal Sciences,  
8 University of Bologna, Viale Fanin 46, 40127 Bologna, Italy

9 <sup>b</sup> Department of Animal Science, Iowa State University, 2255 Kildee Hall, 50011 Ames, Iowa,  
10 USA

11 <sup>c</sup> Biocomputing Group, Department of Biological, Geological, and Environmental Sciences  
12 (BiGeA), University of Bologna, Via San Giacomo 9/2, 40126 Bologna, Italy

13

14

15 \* Corresponding author: Tel: +39 051 2096535

16 Email: [luca.fontanesi@unibo.it](mailto:luca.fontanesi@unibo.it) (L. Fontanesi)

17

18 Running title

19 **Next generation sequencing for dairy species identification**

20 **Highlights**

- 21       • The Ion Torrent platform was used for species identification in dairy products.
- 22       • Sequence analysis of reads from eight libraries detected expected and unexpected species.
- 23       • Next generation sequencing can provide useful information for dairy product authentication.

ACCEPTED MANUSCRIPT

24 **Abstract**

25 In this study, we applied a next generation sequencing (NGS) technology (Ion Torrent) for species  
26 identification based on three mitochondrial DNA (mtDNA) regions amplified on DNA extracted  
27 from dairy products. Sequencing reads derived from three libraries, obtained from artificial DNA  
28 pools or from pooled amplicons, were used to test the method. Then, sequencing results from five  
29 libraries obtained from two mixed goat and cow milk samples, one buffalo mozzarella cheese, one  
30 goat crescenza cheese and one artisanal cured ricotta cheese, were able to detect all expected  
31 species in addition to undeclared species in a few of them. Mining generated reads it was possible to  
32 identify different dairy species mitotypes and the presence of human DNA that could constitute a  
33 potential marker to monitor the hygienic level of dairy products. Overall results demonstrated the  
34 usefulness of NGS for species identification in food products and its possible application for food  
35 authentication.

36  
37 **Keywords:** Dairy product authenticity; Food fraud; DNA analysis; mtDNA; NGS.

## 38 1. Introduction

39 The identification of the species of origin of animal-derived products is an important issue in  
40 food safety and integrity of food production chains, considering the economic impact and the  
41 potential health problems that intentional or accidental admixture and substitutions of products  
42 derived from different species than those declared can create (e.g. Johnson, 2014; Manning, &  
43 Soon, 2016; Spink & Moyer, 2011). For these reasons, food labelling regulations require the precise  
44 identification of the species of origin of commercialized products.

45 Dairy products are among the most important sources of proteins and fat for human nutrition  
46 and the most important agricultural commodity in terms of world economic value. As dairy  
47 products could produce human adverse reactions towards protein allergens of some species, their  
48 substitution or admixture with milk of different species than those declared could create health  
49 problems after consumption (Høst, Jacobsen, Halten, & Holmenlund, 1995; Restani et al., 1999;  
50 Umpiérrez et al., 1999). Moreover, substitution or admixture of cheaper milk, usually of cow origin,  
51 in products labelled as obtained from sheep, goat or buffalo milk (or not declared admixture of milk  
52 from small ruminant species) are the most frequent economically-driven frauds in the dairy industry  
53 (De La Fuente & Juárez, 2005). These fraudulent behaviours are mainly due to lower value of cow  
54 milk compared to that of other species and the seasonality of the milk productions for a few of these  
55 species.

56 Many different methods have been proposed to identify the species of origin of the milk that  
57 is used for the preparation of dairy products, targeting species-specific protein or fat-derived  
58 biomarkers that are present in the milk using various chromatographic, electrophoretic,  
59 immunoenzymatic and mass spectrometric techniques and approaches, among several other  
60 analytical methods (e.g. De La Fuente & Juárez, 2005). Another species-diagnostic molecule  
61 present in the milk is the DNA, contained in the leukocytes and desquamating epithelial cells  
62 originated from the mammary glands of the animals that have produced the milk (i.e. somatic cells).  
63 PCR-based methods have been developed to obtain diagnostic information from the amplification

64 of several targeted sequences of mitochondrial and nuclear DNA (e.g. Mafra, Ferreira, & Oliveira,  
65 2008). Analyses of nucleotide differences contained in the amplified fragments have been obtained  
66 using different approaches (i.e. PCR-RFLP, DNA melting analysis, fragment length analyses,  
67 sequencing, real-time PCR; Dalmasso, Civera, La Neve, & Bottero, 2011; Drummond et al., 2013;  
68 López-Calleja et al., 2007; Maudet & Taberlet, 2001). These methods have been proven to detect  
69 the species of origin of the milk when the potential contaminating species were already known or  
70 expected in advance as the discriminatory analytical methods for their identification have to be  
71 specifically tailored for their detection. In general, PCR based approaches have a poor multiplexing  
72 detection potential and only few studies reported the possibility to identify, at the same time, three  
73 or more species in dairy products (Bottero, Civera, Nucera, Rosati, Sacchi, & Turi, 2003;  
74 Gonçalves, Pereira, Amorim, & van Asch, 2012). To overcome these limits, DNA-based animal  
75 species identification methods have been designed to take advantage from microarray technologies  
76 that, despite the commercial interests on these systems, have not extensively applied in practise  
77 mainly due to the inherent limit determined by their construction (presence or absence of only some  
78 species-specific probes) that cannot give the possibility to detect unexpected or unknown species  
79 (e.g. Chisholm, Conyers, & Hird, 2008; Peter, Brunen-Nieweler, Cammann, Borchers, 2004).

80 Next generation sequencing (NGS) technologies have changed the way to analyse DNA by  
81 combining sequencing and quantification of DNA in a single step (Goodwin, McPherson, &  
82 McCombie, 2016). NGS is considered a common approach in many different studies that requires  
83 the production and analysis of a large number of DNA sequences, including applications in food  
84 sciences. For example, NGS approaches have been used to characterize food microbial communities  
85 or for the identification of pathogenic microorganisms (e.g. Aldrete-Tapia, Escobar-Ramírez,  
86 Tamplin, & Hernández-Iturriaga, 2014; De Pasquale, Di Cagno, Buchin, De Angelis, Gobbetti,  
87 2014; Krawczyk et al., 2015). Despite the power of NGS just few works have applied this  
88 technology for species identification in food authentication and, to our knowledge thus far no  
89 investigation has specifically addressed this question in dairy products. For example, DNA



90 barcoding systems followed by NGS and specific bioinformatics analyses of sequenced reads have  
91 been applied for mammalian species identification in DNA mixtures (Tillmar, Dell'Amico,  
92 Welander, & Holmlund, 2013). Bertolini, Ghionda, D'Alessandro, Geraci, Chiofalo, & Fontanesi  
93 (2015) applied a NGS approach using an Ion Torrent sequencer for the identification of meat  
94 species in DNA mixtures. Other authors reported the application of Ion Torrent sequencing for  
95 species identification in candies, as an example of highly processed food (Muñoz-Colmenero,  
96 Martínez, Roca, & Garcia-Vazquez, 2017).

97 Ion Torrent technology is based on a semiconductor sequencing approach that can detect  
98 small pH modifications on a chip, occurring during the elongation steps in the sequencing process  
99 (Rothberg et al., 2011). Peculiar characteristics of the Ion Torrent platform are i) the relatively low  
100 cost per run compared with other NGS instruments, ii) high speed of the sequencing step (a few  
101 hours), iii) the possibility to use different chips that can allow different scales of sequencing  
102 throughput according to the analytical needs, iv) the possibility to multiplex sequencing analyses by  
103 barcoding different samples that can run on the same chip and v) an error rate that is sequence-  
104 dependent with indels in poly-monomonucleotide stretches (Frey et al., 2014; Quail et al., 2012).

105 In this work, we tested the possibility of using the Ion Torrent next generation semiconductor  
106 based sequencing technology for the identification of dairy species by sequencing targeted mtDNA  
107 fragments obtained from DNA mixtures and DNA extracted from dairy products.

108

## 109 **2. Materials and methods**

### 110 **2.1. DNA isolation**

111 DNA of four dairy species (cattle, *Bos taurus*; sheep, *Ovis aries*; goat, *Capra hircus*; and  
112 buffalo, *Bubalus bubalis*) was isolated from blood using the Wizard® Genomic DNA Purification  
113 kit (Promega Corporation, Madison, WI, USA), following the manufacturer's instructions.

114 DNA was also isolated from: a) two milk mixtures constituted by sterilized milk of goat and  
115 cow, 1) a mixture containing 90% cow milk + 10% goat milk in volume and 2) a mixture containing

116 50% goat milk + 50% cow milk in volume; b) from one buffalo mozzarella (purchased from a  
117 commercial retailer); c) from one goat crescenza cheese (purchased from a commercial retailer and  
118 derived by an industrialized production plant); and d) from one cured ricotta cheese, derived from a  
119 mixture of cow, sheep and goat milk (2:1:1 volume ratio) and purchased from an artisanal  
120 producer. DNA from goat and cow milk was extracted using the NucleoSpin® Tissue kit  
121 (Macherey-Nagel, Düren, Germany) following the manufacturer's protocol for isolation of DNA in  
122 food. For the other dairy products, DNA was isolated using the Wizard® Genomic DNA  
123 Purification kit (Promega Corporation, Madison, WI, USA) following the manufacturer's  
124 instructions for DNA isolation from cell cultures and animal tissues.

125 Nanophotometer P-330 instrument (Implen GmbH, München, Germany) assessment was used  
126 to quantify and to evaluate the quality of the extracted DNA. DNA quality was also evaluated by  
127 visual inspection on 1% agarose gel electrophoresis in TBE 1X buffer after staining with 1X  
128 GelRed Nucleic Acid Gel Stain (Biotium Inc., Hayward, CA, USA).

129 Two artificial DNA pools were constructed using individual DNA samples extracted from  
130 blood: 1) a DNA pool constructed from equimolar DNA content extracted from cattle, sheep, goat  
131 and buffalo; 2) a DNA pool constructed from 10% cattle DNA + 90% buffalo DNA. As the purpose  
132 of this study was to test the potential of the designed approach on species identification, extracted  
133 DNA was used at different defined proportions in DNA pools considering that the mtDNA (the  
134 target DNA in our study) content of milk somatic cells could be similar for each species. Thus,  
135 according to this approximation, the number of reads obtained from each species from artificial  
136 DNA pools containing different proportions of DNA (see below) could provide a preliminary semi-  
137 quantitative estimation of the milk species composition (Ripp et al., 2014), even if this experiment  
138 was not specifically designed to obtain a precise quantitative evaluation. It is also clear that  
139 estimation of milk species composition based on somatic cell DNA evaluation should also assume  
140 similar somatic cell count for each species included in the mixed dairy product (e.g. Russo et al.,  
141 2007).

142

143 **2.2. PCR analyses**

144 PCR amplifications were performed using three primer pairs (Table 1). These primers were  
145 designed on 12S and 16S mitochondrial rRNA genes and were already successfully tested in many  
146 mammalian species (Karlsson & Holmlund 2007; Kitano et al., 2007; Bertolini et al., 2015).  
147 Amplification reactions were performed using a 2720 thermal cycler (Life Technologies, Carlsbad,  
148 CA, USA) in a total volume of 20  $\mu$ L with 2X of the Kapa HiFi HotStart ReadyMix PCR kit (Kapa  
149 Biosystems, Boston, Massachusetts, USA), 50 ng of template DNA and 10 pmol of each primer  
150 with the following cycling profile: initial denaturation of 3 min at 98 °C; then 35 cycles of 20 sec at  
151 98 °C, 15 sec at the primer pairs specific annealing temperature (Table 1), 30 sec at 72 °C; and a  
152 final extension of 1 min at 72 °C. PCR product quality and quantity was estimated by visual  
153 inspection on 2.5% agarose gel electrophoresis in TBE 1X buffer after staining with 1X GelRed  
154 Nucleic Acid Gel Stain (Biotium Inc., Hayward, CA, USA) and on a Nanophotometer P-330  
155 instrument (Implen GmbH).

156

157 **2.3. Sanger sequencing of amplified mtDNA regions and reference sequences**

158 Sanger sequencing was carried out to validate that amplicons obtained by PCR were the  
159 expected targeted mtDNA regions. PCR products were purified with ExoSAP-IT (USB  
160 Corporation, Cleveland, Ohio, USA) and then sequenced with the BrightDye® Terminator Cycle  
161 Sequencing Kit (NIMAGEN, Nijmegen, The Netherlands). Sequencing reactions were loaded on an  
162 ABI3100 Genetic Avant capillary sequencer (Applied Biosystems, Foster City, CA, USA). All  
163 electropherograms were visually inspected and analysed with the help of CodonCode Aligner  
164 (CodonCode Corporation, Dedham, MA, USA). BLASTN ([http://blast.st-](http://blast.st-va.ncbi.nlm.nih.gov/Blast.cgi)  
165 [va.ncbi.nlm.nih.gov/Blast.cgi](http://blast.st-va.ncbi.nlm.nih.gov/Blast.cgi)) was used to confirm the obtained mtDNA sequences against all other  
166 sequences in the nucleotide collection nr/nt constituted by GenBank+EMBL+DDBJ+PDB+RefSeq  
167 sequences (updated on the 30<sup>th</sup> April 2017). Multiple Sequence Comparison by Log- Expectation

168 (MUSCLE) tool (<http://www.ebi.ac.uk/Tools/msa/muscle/>) was used to align obtained sequences  
169 including also the corresponding regions of other mammalian species (retrieved from  
170 EMBL/GenBank; *Cervus elaphus* (deer), accession number KP172593; and *Dama dama* (fallow  
171 deer), accession number JN6326291) that were used for comparison and for subsequent testing  
172 mapping analyses, as described below. To further evaluate the discriminatory potential of the  
173 targeted reference mtDNA regions across dairy species, phylogenetic analyses were obtained using  
174 MEGA v. 6.0.5 (Tamura, Stecher, Peterson, Filipinski, & Kumar, 2013) designing Maximum  
175 Likelihood trees (using the default options) for DNA sequences considering the three different  
176 reference regions separately. The corresponding human mtDNA region was used as outgroup.

177 The three reference sequences for each species were used to identify mitotypes (mitochondrial  
178 haplotypes) segregating within species. Analysis was carried out using species specific BLASTN  
179 querying targeted species against the nr/nt nucleotide collection (30<sup>th</sup> April 2017). Variant sites  
180 were then listed and counted according to the number of GenBank entries reporting the same  
181 sequence. Only mitotypes already reported in GenBank were considered as reliable in this study.

#### 183 **2.4. Ion Torrent sequencing**

184 Ion Torrent PGM (Thermo Fisher Scientific Inc.) sequencing was performed starting from  
185 PCR products (the same amplification volumes for the different primer pair and DNA target  
186 combinations) purified with ExoSAP-IT<sup>®</sup> (USB Corporation, Cleveland, Ohio, USA). Eight  
187 libraries were obtained and each library included amplicons of the three primer pairs. The eight  
188 libraries were constructed from PCR products produced as described below (see Table 2 for more  
189 details): 1) equimolar pool of PCR products obtained separately from cattle, sheep, goat and buffalo  
190 DNA; 2) amplicons obtained from a DNA pool containing equimolar DNA from the four dairy  
191 species; 3) amplicons obtained from a DNA pool containing 90% buffalo DNA + 10% cattle DNA;  
192 4) amplicons obtained from DNA extracted from the mixture containing 90% cow milk + 10% goat  
193 milk in volume; 5) amplicons obtained from DNA extracted from the mixture containing 50% goat

194 milk + 50% cow milk in volume; 6) amplicons obtained from DNA isolated from buffalo  
195 mozzarella; 7) amplicons obtained from DNA isolated from the goat crescenza cheese; 8)  
196 amplicons obtained from DNA extracted from cured ricotta cheese. For each library, 200 ng of  
197 amplified DNA was end-repaired and ligated with a specific barcode using the Ion Xpress™ Plus  
198 Fragment Library and Ion Xpress™ Barcode Adapters 1-16 kits (Thermo Fisher Scientific Inc.).  
199 Then, each library was quantified with the Ion Library Quantitation kit (Thermo Fisher Scientific  
200 Inc.) by qPCR using a StepOnePlus™ Real-Time PCR System (Thermo Fisher Scientific Inc.).  
201 Finally, the eight barcodes were pooled together at the same concentration, clonally amplified by  
202 emulsion PCR with the Ion PGM™ Hi-Q™ OT2 kit and sequenced following the manufacturer's  
203 instructions using the Ion PGM™ Hi-Q™ Sequencing kit and a Ion 318 v2 chip (Thermo Fisher  
204 Scientific Inc.).

205

## 206 **2.5. Next generation sequencing data analyses**

207 Reads obtained from Ion Torrent sequencing were first processed by the Torrent Suite v.4.6  
208 on the Ion Torrent Server (Thermo Fisher Scientific Inc.). Then, obtained FASTQ files were quality  
209 checked using *fastqc* tool (<http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>). Reads were  
210 subsequently separated according to their barcode and polyclonal and low quality sequences were  
211 eliminated. Adapters and low quality 3'-ends were trimmed from the high quality grouped reads.  
212 After the automatic processes, for each barcode, reads were then trimmed from the primer  
213 sequences at 5' and 3' end using the *trim* function of HOMER. Retained reads were 50-210 bp long  
214 with a quality score equal or higher than Q20 for all bases. Trimmed reads were aligned to pre-built  
215 reference sequences of the four targeted dairy species (*Bos taurus*, *Ovis aries*, *Capra hircus* and  
216 *Bubalus bubalis*) and of the additional sequences (*Dama dama*, *Cervus elaphus*) that were used to  
217 test the alignment algorithms and of *Homo sapiens* (that was used to evaluate potential  
218 contamination). Alignments were based on *bwa* v.0.7.11 (Li et al., 2009) according to the *aln* (Li &  
219 Durbin, 2009) and *mem* (<http://bio-bwa.sourceforge.net/>) algorithms using default parameters.

220 Among the already tested algorithms by Bertolini et al. (2015), *aln* is considered appropriate to  
221 discriminate closely homologous sequences (Li & Durbin, 2009), whereas *mem* has been shown to  
222 have the best mapping rate on short reads (Bertolini et al., 2015). Bam files were obtained using  
223 Samtools software v.0.1.16 (Li et al., 2009). Bioinformatic pipelines were run using Galaxy  
224 platform (Goecks et al., 2010). Identification of different within-species mitotypes from the  
225 produced reads was obtained using *aln* alignments with scripts to automatically retrieve different  
226 mitotype sequence information (as detected by BLASTN analysis as described above) on the  
227 aligned reads. Aligned reads were also visually inspected with Integrative Genomic Viewer (IGV)  
228 (<http://software.broadinstitute.org/software/igv/>) to evaluate if the alignments included the correct  
229 sequences and to calculate the error per sequenced base that was obtained for each target gene by  
230 counting mismatches obtained in the filtered alignments and considering the total number of aligned  
231 nucleotides obtained from Library 1. Reference sequences in these comparisons were those  
232 obtained by Sanger sequencing (as described above) from the DNA of the animals used to construct  
233 Library 1. Next generation sequencing data were deposited in the EMBL-EBI European Nucleotide  
234 Archive (ENA) with the project accession number PRJEB21951.

235

### 236 **3. Results**

#### 237 ***3.1. Sanger sequencing data and analysis of targeted mtDNA regions***

238 Three primer pairs designed on two mtDNA genes (Table 1) were used to amplify DNA from  
239 animals of four dairy species (i.e. cattle, buffalo, sheep and goat). All primer pairs produced the  
240 expected amplicons without any detectable unspecific products (data not shown). Sanger  
241 sequencing of PCR generated products cannot identify the presence of mtDNA heteroplasmy in the  
242 extracted DNA of the amplified animals if the level of heteroplasmy is low (Li et al., 2010). Sanger  
243 sequences obtained for all species and primer pairs were the same as those reported in sequences  
244 already deposited in EMBL/GenBank. Phylogenetic trees produced from the sequenced target  
245 regions confirmed that all three selected mtDNA gene fragments contain sequence specific

246 information useful to distinguish four dairy species included in this study (Fig. S1). The closest  
247 identity was between the goat and sheep 12S\_Ki sequences that differs for two nucleotides only  
248 (Fig. S2 and Table S1). BLASTN analysis showed that for some of these amplified mtDNA  
249 regions, within species variants exist among the database deposited sequences. That means that  
250 some of the targeted regions (buffalo 12S\_Ki, 16S\_Ki and 16S\_KH regions; sheep 12S\_Ki; goat  
251 16S\_KH; Table S2) contain sequence information useful to identify different mitotypes segregating  
252 within species (as estimated based on the number of matched database entries) in three out of four  
253 dairy species.

254

### 255 **3.2. Analysis of Ion Torrent reads: tested mapping algorithms**

256 Ion Torrent reads obtained from Libraries 1, 2 and 3 (prepared using amplicons of the four  
257 different dairy species produced separately or from DNA mixtures from four or two species; Table  
258 2) were used to assess the power of filtering and alignment of two algorithms (*aln* and *mem*). Data  
259 from Library 1 might not be biased by potential unequal PCR amplification efficiency in DNA  
260 pools or complex matrices (all other libraries). As all libraries were produced without any  
261 fragmentation of the obtained amplicons, it is possible to consider all mapped reads on a reference  
262 sequence as unbiased estimators of the number of times in which the targeted fragments are  
263 detected without any adjustments, that multi-reads approaches (for long targeted regions) might  
264 require (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008; Bertolini et al., 2015).

265 Performances of the *aln* and *mem* algorithms was evaluated in terms of number of mapped  
266 reads and quality scores of the alignments. All mapped reads were processed without any  
267 preliminary filtering step (raw or unfiltered alignments) and after filtering based on mapping quality  
268 score  $>30$  ( $Q_m > 30$ ) that eliminated reads mapped on more than one reference sequence. Fig. 1 and  
269 Table S3 report the number of mapped reads on the targeted mtDNA reference sequences (12S\_Ki,  
270 16S\_Ki and 16S\_KH) of the four dairy species and the relative proportion of the different species  
271 within targeted mtDNA fragment. Raw alignments obtained with *mem* accounted for a larger



272 number of mapped reads (about 22-29% larger) than those mapped with *aln* in all three libraries.  
273 However, after filtering, *aln* retained a larger number of reads than *mem* in all three libraries (Table  
274 S3). The number of filtered reads with *mem* were about 72%, 41% and 76% the number of  
275 unfiltered reads for the three libraries, respectively. Using *aln*, almost all unfiltered reads (96-98%  
276 in all three libraries) were retained after applying the quality score  $Q_m > 30$ . Based on these results, it  
277 seems that the best performing algorithm for closely related sequences considered in our case  
278 studies is *aln* that was properly designed for the Ion Torrent sequencing platform. The low  
279 efficiency of the *mem* algorithm can be mainly attributed to the absence of  $Q_m > 30$  reads assigned to  
280 goat and sheep for the 12S\_Ki fragment. This bias is evident also from the bioinformatic results  
281 obtained from Library 8 in which both goat and sheep DNA was present. This is due to the very  
282 close identity of these two species in this region (Table S1 and Fig. S2) that prevented the unique  
283 assignment of the reads to one or to the other small ruminant species, considering the very stringent  
284 quality mapping threshold.

285

### 286 **3.3. Analysis of Ion Torrent reads: estimation of sequencing error rates**

287 Errors introduced by the Ion Torrent sequencing system might be important to evaluate the  
288 performances of the mapping algorithm for the assignment of the obtained reads to different  
289 species. Sanger sequence information we obtained from the DNA of the four animals (each of  
290 different species) used to prepare the products sequenced in Library 1 was important to correctly  
291 calculate the error rate of the NGS-based detection system. Error rate was estimated by comparing  
292 Sanger validated sequence data of targeted regions with *aln*  $Q_m > 30$  aligned Ion Torrent reads  
293 obtained from Library 1. All variant positions were then counted as errors, after elimination of  
294 spurious alignments due to the closeness of the investigated species. Error rate was calculated as  
295 fraction of all counted errors on all matched sequence positions, excluding primer positions that  
296 were previously trimmed (Table 3). The error rate was larger for the 16S\_KH fragments (the  
297 shortest targeted mtDNA region) for all species (ranging from 0.046 to 0.057) and lower for the



298 16S\_Ki and 12S\_Ki fragments (the longest targeted mtDNA regions), in all species/region  
299 combinations (ranging from 0.0206 to 0.0243). These errors might be due to the sum of PCR and  
300 Ion Torrent sequencing errors. Most of the errors were observed in short homo-polymeric regions  
301 (data not shown) that are difficult to be correctly sequenced using the Ion Torrent technology  
302 (Rothberg et al., 2011). Estimated error rate for all targeted regions are below the level of  
303 differences between the sequences of the analysed species in all comparisons apart from the sheep  
304 and goat 12S\_Ki sequences whose identity was equal to 0.989 (Table S1). Combining the results of  
305 the mapping algorithms for the 12S\_Ki sheep and goat sequences and the sequencing error rate for  
306 this region, it is possible to understand the reason for the difficulties of the unsupervised  
307 bioinformatic analysis to attribute reads to one or to the other species.

308

### 309 ***3.3. Identification of calling thresholds for species identification from Ion Torrent reads***

310 Table S3 report also the aligned reads to the reference sequences of other two species (deer  
311 and fallow deer) not included in Libraries 1, 2 and 3. Alignments on these two ruminant sequences  
312 were used to explore the possibility to define a false positive alignment rate, considering true  
313 sequences from closely related species. DNA of these deer species has never been analysed in the  
314 laboratory that carried out this experiment, excluding in this way any potential DNA contamination.  
315 Reads from the 16S\_KH primer pair did not align with the reference deer and fallow deer sequences  
316 using both raw and filtered data and both algorithms. Only fallow deer 12S\_Ki reference sequences  
317 aligned with a few unfiltered reads using *mem* (0.03%, 0.07% and 0.08% on the total of the 12S\_Ki  
318 reads for the three libraries, respectively) and none of these reads were retained when filtered for  
319 alignment quality. The *aln* algorithm did not report any aligned reads for both deer species (for both  
320 unfiltered and filtered reads). A larger number of unfiltered reads was aligned with both algorithms  
321 to the deer and fallow deer 16H\_Ki reference sequences. However, aligned filtered reads ranged  
322 from 0 to 0.12% with the *aln* algorithm whereas the *mem* algorithm did not align any reads for both

323 species and all three libraries. Close inspection of the alignment results confirmed that the mapped  
324 reads could not be attributed to deer and fallow deer species.

325 In addition to the evaluation obtained using deer and fallow deer reference sequences, as  
326 Library 3 was derived only from cattle and buffalo amplified DNA, we used mapped reads on goat  
327 and sheep reference sequences to determine the level of spurious alignments due to the  
328 unsupervised performances of the two tested algorithms in addition to true alignments derived by  
329 amplification of potential contaminating DNA. As goat and sheep DNA are currently analysed in  
330 the laboratory that performed the Ion Torrent sequencing experiment, alignments on the goat and  
331 sheep reference sequences coming from Library 3 might be useful to test the problem derived by  
332 potential contaminating DNA from these sources. Both algorithms aligned from 0.0 to 0.05% reads  
333 to the sheep and goat 12S\_Ki and 16S\_KH reference sequences ( $Q_m > 30$ ). For the 16S\_Ki region,  
334 *mem* and *aln* aligned 0.17% and 0.14% of reads on the goat reference sequence and 0.01% and  
335 0.01% of reads to the sheep reference sequence ( $Q_m > 30$ ). These results are in line with the spurious  
336 alignment rates obtained with the deer and fallow deer sequences even if slightly higher than those  
337 obtained with the two previously tested species. If we consider the *aln* algorithm, that seems the  
338 best performing algorithm for closely related sequences, and comparing the true spurious results  
339 obtained for deer and fallow deer sequences, it could be possible that the contribution of the  
340 environmental laboratory contamination derived by goat and sheep DNA could account for about  
341 0.01-0.05% of reads mapped to the reference sequences of these two species. Close inspection of  
342 the aligned reads confirmed that only part of the mapped reads could be attributed to goat or sheep  
343 sequences (data not shown).

344 Based on these evaluations we set as threshold for declaring the presence of a species in a  
345 sequenced library the number of reads matching the species-specific sequence of at least 0.2% of  
346 the total number of reads matching a reference sequence of the same targeted region. Moreover, to  
347 account for potential undisclosed factors causing false positive results from one hand and possible  
348 amplification biases due to different PCR efficiencies in DNA mixtures from the other hands, we

349 declared the presence of one species if at least two out of three targeted mtDNA fragments  
350 accounted for at least 0.2% of the aligned reads on its corresponding species-specific reference  
351 sequence. This level for at least two regions might account for the problem of un-supervised  
352 spurious alignments, environmental laboratory contamination and might consider the need of an  
353 internal confirmation of the results. This threshold of 0.2% defined in the evaluated case studies  
354 should be further validated in other NGS runs.

355

### 356 ***3.4. Evaluation of the aligned reads from Libraries 1, 2 and 3 on dairy species reference*** 357 ***sequences***

358 The largest number of reads obtained from Library 1 was mapped on the 16S\_Ki reference  
359 sequences using both algorithms (Table S3). In Library 3, the largest number of reads mapped with  
360  $Q_m > 30$  was obtained from the 12S\_Ki fragment. In Library 2, *mem* reported the largest number of  
361 mapped reads for the 16S\_Ki fragment whereas *aln* showed the largest number of reads mapped on  
362 the 12S\_Ki region. The 16S\_KH mtDNA targeted fragment accounted for the lowest number of  
363 matched reads in all libraries. The lowest number of reads obtained with this primer pair could be  
364 due to a lower amplification efficiency compared to the other two primer pairs (considering that all  
365 libraries were generated by pooling the same amplification volumes for the different primer pair and  
366 DNA target combinations). However, we cannot completely exclude a lower sequencing efficiency  
367 in the sequencing processes for the 16S\_KH reads.

368 Reads produced from all three libraries matched the expected dairy species even if the  
369 proportion of mapped reads was not exactly according to the expected ratios derived by the  
370 constitution of the three libraries (Fig. 2). Reads produced from Library 1 showed the closest  
371 expected proportion according to the equal contribution among species of the amplified fragments  
372 that were included in the sequenced pool of amplicons that originated this library. This proportion is  
373 maintained if reads are counted separately for each targeted region or considering the sum of reads  
374 for each species across the three targeted fragments (Fig. 2). Amplification biases of the sheep DNA

375 seems a possible reason for the overrepresentation of reads of this species observed from Library 2  
376 results. A similar problem might explain the overrepresentation of buffalo reads over cattle reads  
377 reported from Library 3.

378

### 379 **3.5. Analysis of Ion Torrent reads produced from dairy products**

380 All other libraries (Libraries 4, 5, 6, 7 and 8) were obtained from amplified products derived  
381 by extracted DNA from dairy products (two artificial mixtures of goat and cow milk: Libraries 4  
382 and 5; buffalo mozzarella cheese: Library 6; goat crescenza cheese: Library 7; and cured ricotta  
383 cheese: Library 8; Table 2).

384 Results obtained from the two “mixed milk” libraries confirmed the presence of DNA from  
385 goat and cattle (Figs. 1 and 2; Table S3). In particular, results obtained from Library 4 derived by  
386 mixing 90% of cow milk with 10% of goat milk showed about 72-84% of reads of the three targeted  
387 regions mapped on cattle reference sequences and 16-28% of reads attributed to the goat species.  
388 Apart from potential amplification biases that could interfere with the expected 90:10 proportion of  
389 the two species (as in part observed in Libraries 2 and 3) the larger number of goat reads than those  
390 expected could be attributed to different somatic cell count (causing differences in DNA  
391 proportions) of the milk obtained from the two species (higher in goat and lower in cattle). This  
392 latter hypothesis seems confirmed analyzing the results obtained in Library 5 (50:50 of goat and  
393 cattle milk volume) that reported a larger number of goat reads than cattle reads. Moreover, reads  
394 obtained from this library matched sheep mtDNA reference sequences with a proportion for two  
395 targeted regions that trespassed the 0.2% threshold (defined above) suggesting the presence of  
396 sheep milk probably mixed in the original goat milk. The presence of sheep milk could be  
397 confirmed also from the results of Library 4 that reported the presence of 0.25% of 12S\_Ki reads  
398 (above the 0.2% threshold) matching the sheep reference sequence. However, as none of the other  
399 two amplified regions accounted for more than 0.2% of mapped reads on the sheep reference

400 sequences (only the 16S\_Ki accounted for 0.1% of sheep mapped reads) it could not be possible to  
401 have an internal confirmation of the results obtained for the 12S\_Ki region.

402 Ion Torrent reads obtained from mozzarella cheese confirmed that this product derived only  
403 from buffalo milk as no other species trespassed the 0.2% threshold of matched reads (Figs. 1 and  
404 2; Table S3).

405 Goat crescenza cheese produced the vast majority of reads matching the three targeted goat  
406 mtDNA fragments (Figs. 1 and 2; Table S3). However, it seems that this product could contain  
407 traces of sheep DNA probably derived by accidental mixture of sheep milk in goat milk. Two out of  
408 three targeted mtDNA regions showed a number of matched reads on the sheep reference sequences  
409 that trespassed the 0.2% threshold (Table S3).

410 The cured ricotta cheese was declared to be derived from cow, sheep and goat milk in a ratio  
411 of 2:1:1. The producer processes also buffalo milk in his plant. Ion Torrent results confirmed the  
412 presence of milk of the three declared species that accounted for  $\square$ 57%, 25% and 17% of the reads  
413 (considering together all targeted mtDNA fragments) closely matching the declared proportions.  
414 Moreover, traces of buffalo milk were identified from the 12S\_Ki and 16S\_Ki reads that matched  
415 these buffalo reference sequences (0.27% and 0.6% of the total reads of the corresponding regions,  
416 respectively) suggesting a possible accidental inclusion of buffalo milk (Table S3).

### 418 ***3.6. Identification of dairy species mitotypes from Ion Torrent reads***

419 Sequences obtained from mtDNA could reveal within species mitotypes. BLASTN analysis  
420 of the GenBank database using our reference sequences obtained by Sanger sequencing showed that  
421 a few mtDNA region/species combinations could be useful to retrieve information on mitotypes  
422 from sequences produced by other studies. Buffalo 12S\_Ki, 16S\_Ki and 16S\_KH regions showed a  
423 total 2, 3 and 2 different buffalo mitotypes from sequences already deposited in GenBank (with  
424 different occurrence determined by the number of GenBank entries reporting the same sequence;  
425 Table S2). These mitotypes differed from each other in 2, 2 and 2 nucleotide positions, respectively

426 (Table S2). The most frequent mitotype reported in GenBank for the three amplified regions was  
427 also the mitotype sequenced using the Sanger approach in our study. Sheep 12S\_Ki and goat  
428 16S\_KH identified each 2 types of entries distinguishable by one nucleotide position whereas the  
429 bovine mtDNA sequences did not disclose any cattle mitotype. To further reduce the problem  
430 derived by the NGS error rate, we considered informative mitotypes only those that could be  
431 identified by at least two polymorphic sites. In this way, we could define the false positive detection  
432 threshold as the product of the Ion Torrent sequencing error rate determined for each nucleotide  
433 position calculated for each amplified region. Therefore, only *Bubalus bubalis* mtDNA sequences  
434 were considered informative in our study. For example, for the buffalo 16S\_Ki mitotypes that could  
435 be distinguished by two nucleotide positions, its false positive rate (see Table 3) could be defined as  
436  $0.022^2 = 0.000484$  or 0.0484% of the total number of reads matching the buffalo 16S\_Ki fragment.  
437 Based on these evaluations, we reanalyzed the corresponding buffalo reads identified with *aln* from  
438 Libraries 1, 2 and 3, 4 and 8 that reported the presence of buffalo DNA. Reads matching the buffalo  
439 12S\_Ki fragments could be attributed to the two different haplotypes. The less frequent mitotype  
440 (reported in GenBank from only one entry; and detected with 0.18% of the aligned reads from all  
441 libraries from which buffalo read were reported) might be due to mtDNA heteroplasmy present in  
442 the original buffalo sample that was used in Libraries 1, 2 and 3 and then detected with a similar  
443 level from reads of Libraries 4 and 8 derived from dairy products. Only one mitotype (the most  
444 frequent deduced from the number of GenBank entries) was identified for the 16S\_Ki and 16S\_KH  
445 fragments in all buffalo containing libraries.

446

### 447 **3.7. Evaluation of human accidental contamination**

448 The universal primers designed on vertebrate mtDNA sequences amplify also human mtDNA  
449 (Karlsson & Holmlund 2007; Kitano et al., 2007; Bertolini et al., 2015). Therefore, alignments of  
450 Ion Torrent reads, obtained from Libraries 1, 2 and 3 on the corresponding human mtDNA  
451 sequences, were used to evaluate the level of accidental human contamination in the laboratory

452 environment. Next generation sequencing analyses are considered very sensitive to this potential  
453 problem and the quantification of human mtDNA matched reads is important to define the quality  
454 of the laboratory procedures (Tillmar et al., 2013). The number of reads matching the three human  
455 targeted sequences ranged from 0 to 4 using unfiltered reads and from 0 to 1 using filtered reads  
456 (out of a total of about 40,000-160,000 aligned reads) for both algorithms (Table S3). That means  
457 that human DNA contamination in the laboratory environment in which this study was carried out  
458 can be considered almost absent or below the detection limit of the system.

459 Therefore, it could be possible to test the level of human contamination on the investigated  
460 dairy products. Among these products, it seems clear that the cured ricotta cheese showed the most  
461 evident signs of human contamination, as deduced from the 0.6% and 1.82% of reads matching the  
462 12S\_Ki and 16S\_Ki corresponding human reference sequences. Buffalo mozzarella cheese reported  
463 0.07% and 0.1% of human matched reads for the two regions respectively. Even if this level might  
464 not trespass the threshold for declaring the presence of a species (that actually was not defined using  
465 human sequences), we could consider this information for further considerations. Milk mixtures and  
466 goat crescenza cheese samples did not show any human contamination as the number of reads  
467 matching the human corresponding regions were similar to those reported for Libraries 1, 2 and 3.

468 Following a similar approach reported for the identification of dairy species mitotypes, it was  
469 possible to trace back the ethnicity of the human DNA contamination in the cured ricotta cheese and  
470 buffalo mozzarella cheese. Human mtDNA sequences corresponded to subgroups of mitotype H,  
471 that is known to be the most frequent mitotype in Mediterranean human populations (Brotherton et  
472 al., 2013). Therefore, as the original production regions of the cured ricotta cheese and of the  
473 buffalo mozzarella cheese are from the South of Italy, the human mtDNA information is compatible  
474 with the origin of human contaminations derived from local people working in the production  
475 chains of these two dairy products. This information could be useful for forensic evaluation of the  
476 analysed products.

477



#### 478 4. Discussion

479 Few other studies applied NGS technologies for species identification in food products. For  
480 example, Park et al. (2012) compared GS Junior Titanium Sequencing with microarray analysis for  
481 fish species identification in fish cake samples. A whole DNA sequencing approach without any  
482 PCR based preselection step was simulated and then applied for species detection in sausages using  
483 Illumina HiSeq 2000 and MiSeq generated data (Ripp et al. 2014). Bertolini et al. (2015) used the  
484 Ion Torrent platform for meat species identification in DNA mixtures testing different species  
485 combinations and proportions. In this study, following the approach of Bertolini et al. (2015), we  
486 applied the Ion Torrent semiconductor based sequencing technology for species identification in  
487 dairy products by sequencing targeted mtDNA fragments amplified with universal primers. One  
488 advantage of NGS approaches for species identification is that it is possible to detect at the same  
489 time all species that are present in a food matrix, including also unexpected species that were not  
490 supposed to contribute to the analysed product. As dairy products are usually derived from milk of a  
491 few species (actually, the most important ones are cattle, buffalo, sheep and goat), NGS for dairy  
492 species identification might be of practical interests if it could be possible to identify other  
493 information in addition to the presence of these species that could be done also using other  
494 approaches. We demonstrated that it is possible to identify, at the same time, other information  
495 useful to further characterize the products, by mining sequence data using different bioinformatic  
496 analyses. In addition, as dairy products are usually not very complex in terms of possible species of  
497 origin of the milk, they can be used as simple case studies for the real application of novel NGS  
498 approaches.

499 We first tested the system by sequencing artificially pooled DNA fragments obtained from  
500 separated species-specific amplifications or from amplicons generated from artificially pooled  
501 whole DNA (Libraries 1, 2 and 3). These artificially controlled mixtures made it possible to  
502 evaluate the efficiency of the bioinformatic pipeline based on different algorithms to map obtained  
503 reads to the corresponding reference sequences. Bertolini et al. (2015) evaluated the performances



504 of three algorithms. As one of them (*SW*) worked similarly to *mem*, in this study we tested two  
505 different algorithms (*mem* and *aln*) that have different procedures for mapping reads. For dairy  
506 species, considering the closeness of a few species (sheep and goat but also cattle and buffalo), *aln*  
507 performed better than *mem*. This latter algorithm discarded at  $Q_m > 30$  a lot of goat and sheep  
508 sequences belonging to the 12S\_Ki fragment for which these two species are different in only two  
509 nucleotide positions. Therefore, we selected *aln* for all subsequent analyses and data interpretations.  
510 Bertolini et al. (2015) selected *mem* instead of *aln* as it performed better in the specific tested  
511 hypotheses. That means the choice of the better bioinformatic approach to analyse NGS data for  
512 food species identification might be according to the problems under investigations, the compared  
513 regions and the closeness of the species-specific sequences. Other algorithms or bioinformatic  
514 approaches could be evaluated (i.e. Pabinger et al., 2014) and this will be a matter of further  
515 investigations. The used NGS technology also conditions the use of the data analysis approach that  
516 is also related to the sequencing error rate.

517 We evaluated the error rate of the Ion Torrent sequenced amplicons, that, according to what  
518 we previously reported (Bertolini et al., 2015), is quite variable and depends on the sequenced  
519 fragment/region. The shortest fragments showed the highest error rate but this might depend also by  
520 the chemistry and kit specificities. Therefore, it is difficult to compare results and performances  
521 across studies. Error rate might affect the bioinformatic detection noise that we estimated  
522 empirically by testing the match rate against reference sequences of species that may not  
523 contaminate the laboratory environment in which the NGS experiment was carried out (deer and  
524 fallow deer) and that, at the same time, are quite close to the ruminant dairy species for which we  
525 expected matches. This is an innovative approach that differentiated the background noise derived  
526 by true DNA contamination of the laboratory environment. The definition of a background noise is  
527 important to identify thresholds that can be used to confidentially declare the presence of a species  
528 in a food product. Next generation sequencing technologies coupled with PCR amplification are  
529 very sensitive in detecting the presence of contaminating DNA and laboratory procedures should be

530 designed to minimize this problem (Tillmar et al., 2013). Our evaluation excluded the presence of a  
531 level of contaminating DNA that prevented the identification of a species from the analysed  
532 libraries and case studies. The limit of detection was defined based on the presence of at least 0.2%  
533 of reads for at least two amplified regions. The use of more than one amplified fragment to attribute  
534 the presence of a species is important to prevent false positive results. An internal confirmation  
535 based on concordance of at least two out of three targeted regions (and not just three out of three) in  
536 the tested case studies is also useful to overcome the problem derived by PCR failure or low  
537 efficiency in the multiplex amplifications or partial inhibitions due to PCR inhibitors that could be  
538 present in food products and that might remain after DNA extraction (i.e. Demeke & Jenkins, 2010;  
539 Hellberg, Kawalek, Van, Shen, & Williams-Hill, 2014).

540 Species identification with the NGS approach tested in our study is not quantitative. The  
541 number of reads attributed to the different species only roughly respect the proportions derived by  
542 the sequenced DNA of Libraries 1, 2 and 3 (derived by the three artificial DNA pool samples).  
543 Different DNA quality of the original samples and quantification precision might be sources of  
544 technical errors. In addition, the sequencing platform could be very sensitive and additional  
545 sequencing biases might be introduced. An extensive evaluation of all variable factors might be  
546 needed to better define the limits for quantitative analysis of the Ion Torrent sequencing data, that  
547 on the other hand seems quite repeatable as correlations of results between runs has been estimated  
548 to be very high (Bertolini et al., 2015).

549 It is well known, however, that quantitative identification of milk of different species based  
550 on DNA analyses is largely affected by somatic cell counts of the mixed milk samples that  
551 contributed to produce the final dairy product. This seems quite evident also from the Ion Torrent  
552 results obtained from Libraries 4 and 5 derived by different volume ratios of cow and goat milk.  
553 Another potential source of quantitative variability in DNA content of the analysed matrixes might  
554 derive by the different number of mtDNA copies contained in the analysed cells (Robin, & Wong,  
555 1988).

556 Even if we could not have a precise evaluation of the original milk samples (sterilization of  
557 the purchased milk largely prevents a precise evaluation of somatic cell count) obtained results are  
558 compatible with higher somatic cell count in the goat milk than in the cow milk used in the  
559 constructed case studies. However, we could not completely exclude other technical problems that  
560 might be derived by the PCR step or by the sequencing procedure.

561 Analysing the results we obtained from the investigated dairy products (the same mixed milk  
562 samples, the buffalo mozzarella cheese, the goat crescenza cheese and ricotta cheese) interesting  
563 information could be obtained. We detected the presence of sheep DNA in the milk mixtures,  
564 probably derived by its presence in the goat milk. This is mainly evident from the 50:50 volume of  
565 goat and cattle milk library preparation. The level of sheep DNA is low (just above the 0.2%  
566 thresholds for two out of three targeted mtDNA fragments) and could be due to accidental  
567 contamination of sheep milk in the goat production chain or milk processing plant. We carefully  
568 checked the aligned sheep reads and we confirmed that these reads are not due to misalignments of  
569 closely related sequences coming from goat DNA. In this specific case, both the 12S\_Ki (the closest  
570 region between goat and sheep among the three analysed regions) and the 16S\_Ki (quite different  
571 between sheep and goat) fragments confirmed this result. A similar low level of sheep DNA was  
572 detected in the goat crescenza cheese for which the same two amplified regions confirmed the  
573 presence (probably accidental) of sheep milk in the analysed dairy product.

574 In the buffalo mozzarella cheese, the presence of cattle DNA was below the defined  
575 thresholds and might be derived by spurious sequence alignments. Interestingly, buffalo DNA could  
576 be detected in the cured ricotta cheese for which its presence, even if not declared, could not  
577 represent a surprise due to the artisanal production in a processing plant that uses also buffalo milk.  
578 Its level is however low and probably due to accidental mixture. This artisanal product showed also  
579 another interesting signature derived by quite a high level (0.6%-1.8%) of reads assigned to two  
580 human mtDNA regions. As we excluded that this result could be derived by accidental laboratory  
581 contamination from human DNA, we could use these indicators as a possible source of information

582 of hygienic conduction of this artisanal production. Among the other investigated dairy products,  
583 only the buffalo mozzarella cheese showed a certain level of human reads. The identification of  
584 human mtDNA sequences could be useful to define the level of hygienic conditions in processed  
585 dairy products. In our study, we were able to distinguish the level of human DNA contamination  
586 derived by laboratory procedures from that derived by dairy production contaminated environments.  
587 As human DNA contamination in laboratory practices might depend by many factors that cannot be  
588 controlled always, it is important that NGS experiments are properly designed to evaluate this  
589 potential problem that might prevent the possibility to use information of human DNA for other  
590 monitoring purposes.

591 Another interesting information that is possible to obtain by analyzing dairy species reads is  
592 the presence of different mitotypes. Mitochondrial haplotypes segregating within species or derived  
593 by individual heteroplasmy (Li et al., 2010) can add another level of signature to the obtained dairy  
594 products. For example, the different mitotypes identified from the mozzarella cheese DNA could be  
595 derived by the contribution of different animals to the buffalo milk that was used to produce the  
596 analysed sample, providing a sort of mtDNA fingerprinting of the product. Therefore, milk  
597 produced by different animals or obtained from some animal populations can be disclosed by  
598 detecting mitotypes.

599

## 600 **5. Conclusions**

601 This study reports several methodological aspects on the use of the Ion Torrent semiconductor  
602 based sequencing for species identification in dairy products and tested this technology in a few  
603 case studies based on real dairy products for which we knew in advance few information or that we  
604 disclosed by analyzing data contained in thousands of sequenced reads. The bioinformatic mining  
605 of the generated reads constitutes a key part of this NGS approach, in particular when sequences are  
606 very similar between species. By using universal primers to amplify several regions, it is possible to  
607 obtain an internal validation of the results derived by the concordance of the assigned reads to a

608 species. The sensitivity of the platform needs to define the appropriate level of background signal of  
609 the analytical steps. However, results we have obtained are promising and might constitute a first  
610 step to design of robust analytical systems that could become a reference method for species  
611 identification not only for dairy products but for many other food products for which this  
612 information is important to identify or prevent frauds and for food authenticity.

613

#### 614 **Conflict of interest**

615 The authors declare no conflicts of interest.

616

#### 617 **Acknowledgements**

618 The authors thank Dr. Luca Buttazzoni (CREA) who provided buffalo DNA and Marco  
619 Ghionda (University of Bologna) who collaborated in this study. This work was developed as part  
620 of the GRIFFA research program and was supported by University of Bologna RFO funds.

621

#### 622 **References**

623 Aldrete-Tapia, A., Escobar-Ramírez, M. C., Tamplin, M. L., & Hernández-Iturriaga, M. (2014).

624 High-throughput sequencing of microbial communities in Poro cheese, an artisanal Mexican  
625 cheese. *Food Microbiology*, *44*, 136-141.

626 Bertolini, F., Ghionda, M. C., D'Alessandro, E., Geraci, C., Chiofalo, V., & Fontanesi, L. (2015). A

627 next generation semiconductor based sequencing approach for the identification of meat  
628 species in DNA mixtures. *PLoS ONE*, *10*, e0121701.

629 Bottero, M. T., Civera, T., Nucera, D., Rosati, S., Sacchi, P., & Turi, R. M. (2003). A multiplex

630 polymerase chain reaction for the identification of cows', goats' and sheep's milk in dairy  
631 products. *International Dairy Journal*, *13*, 277-282.

632 Brotherton, P., Haak, W., Templeton, J., Brandt, G., Soubrier, J., Jane Adler, C., Richards, S. M.,

633 Sarkissian, C. D., Ganslmeier, R., Friederich, S., Dresely, V., van Oven, M., Kenyon, R., Van

- 634 der Hoek, M. B., Korlach, J., Luong, K., Ho, S. Y., Quintana-Murci, L., Behar, D. M., Meller,  
635 H., Alt, K. W., Cooper, A., Genographic Consortium, Adhikarla, S., Ganesh Prasad, A. K.,  
636 Pitchappan, R., Varatharajan Santhakumari, A., Balanovska, E., Balanovsky, O., Bertranpetit,  
637 J., Comas, D., Martínez-Cruz, B., Melé, M., Clarke, A. C., Matisoo-Smith, E. A., Dulik, M.  
638 C., Gaieski, J. B., Owings, A. C., Schurr, T. G., Vilar, M. G., Hobbs, A., Soodyall, H., Javed,  
639 A., Parida, L., Platt, D. E., Royyuru, A. K., Jin, L., Li, S., Kaplan, M. E., Merchant, N. C.,  
640 John Mitchell, R., Renfrew, C., Lacerda, D. R., Santos, F. R., Soria Hernanz, D. F., Spencer  
641 Wells, R., Swamikrishnan, P., Tyler-Smith, C., Paulo Vieira, P., & Ziegler, J. S. (2013).  
642 Genographic Consortium. (2013). Neolithic mitochondrial haplogroup H genomes and the  
643 genetic origins of Europeans. *Nature Communications*, *4*, 1764.
- 644 Chisholm, J., Conyers, C. M., & Hird, H. (2008). Species identification in food products using the  
645 bioMerieux FoodExpert-ID (R) system. *European Food Research & Technology*, *228*, 39-45.
- 646 Dalmaso, A., Civera, T., La Neve, F., & Bottero, M. T. (2011). Simultaneous detection of cow and  
647 buffalo milk in mozzarella cheese by real-time PCR assay. *Food Chemistry*, *124*, 362-366.
- 648 De La Fuente, M. A., & Juárez, M. (2005). Authenticity assessment of dairy products. *Critical*  
649 *Reviews in Food Science and Nutrition*, *45*, 563-585.
- 650 De Pasquale, I., Di Cagno, R., Buchin, S., De Angelis, M., & Gobbetti, M. (2014). Microbial  
651 ecology dynamics reveal a succession in the core microbiota involved in the ripening of pasta  
652 filata caciocavallo pugliese cheese. *Applied and Environmental Microbiology*, *80*, 6243-6255.
- 653 Demeke, T., & Jenkins, G. R. (2010). Influence of DNA extraction methods, PCR inhibitors and  
654 quantification methods on real-time PCR assay of biotechnology-derived traits. *Analytical*  
655 *and Bioanalytical Chemistry*, *396*, 1977-1990.
- 656 Drummond, M. G., Brasil, B. S. A. F., Dalsecco, L. S., Brasil, R. S. A. F., Teixeira, L. V., &  
657 Oliveira, D. A. A. (2013). A versatile real-time PCR method to quantify bovine contamination  
658 in buffalo products. *Food Control*, *29*, 131-137.

- 659 Frey, K. G., Herrera-Galeano, J. E., Redden, C. L., Luu, T. V., Servetas, S. L., Mateczun, A. J.,  
660 Mokashi, V. P., & Bishop-Lilly, K. A. (2014). Comparison of three next-generation  
661 sequencing platforms for metagenomic sequencing and identification of pathogens in blood.  
662 *BMC Genomics, 15*, 96.
- 663 Goecks, J., Nekrutenko, A., Taylor, J., & The Galaxy Team (2010). Galaxy: a comprehensive  
664 approach for supporting accessible, reproducible, and transparent computational research in  
665 the life sciences. *Genome Biology, 11*, R86.
- 666 Golinelli, L. P., Carvalho, A. C., Casaes, R. S., Lopes, C. S., Deliza, R., Paschoalin, V. M., & Silva,  
667 J. T. (2014). Sensory analysis and species-specific PCR detect bovine milk adulteration of  
668 frescal (fresh) goat cheese. *Journal of Dairy Science, 97*, 6693-6699.
- 669 Gonçalves, J., Pereira, F., Amorim, A., & van Asch, B. (2012). New method for the simultaneous  
670 identification of cow, sheep, goat, and water buffalo in dairy products by analysis of short  
671 species-specific mitochondrial DNA targets. *Journal of Agricultural and Food Chemistry, 60*,  
672 10480-10485.
- 673 Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-  
674 generation sequencing technologies. *Nature Reviews Genetics, 17*, 333-351.
- 675 Hellberg, R. S., Kawalek, M. D., Van, K. T., Shen, Y., & Williams-Hill, D. M. (2014). Comparison  
676 of DNA extraction and PCR setup methods for use in high-throughput DNA barcoding of fish  
677 species. *Food Analytical Methods, 7*, 1950-1959.
- 678 Høst, A., Jacobsen, H. P., Halken, S., & Holmenlund, D. (1995). The natural history of cow's milk  
679 protein allergy/intolerance. *European Journal of Clinical Nutrition, 49*, Suppl. 1, S13-S18.
- 680 Karlsson, A. O., & Holmlund, G. (2007). Identification of mammal species using species-specific  
681 DNA pyrosequencing. *Forensic Science International, 73*, 16-20.
- 682 Kitano, T., Umetsu, K., Tian, W., Osawa, M. (2007). Two universal primer sets for species  
683 identification among vertebrates. *International Journal of Legal Medicine, 121*, 423-427.



- 684 Krawczyk, A. O., de Jong, A., Eijlander, R. T., Berendsen, E. M., Holsappel, S., Wells-Bennik, M.  
685 H. J., & Kuipers, O. P. (2015). Next-generation whole-genome sequencing of eight strains of  
686 *Bacillus cereus*, isolated from food. *Genome Announcements*, 3, e01480-15.
- 687 Li, H., & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler  
688 transform. *Bioinformatics*, 25, 1754-1760.
- 689 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., &  
690 Durbin, R. (2009). 1000 Genome Project Data Processing Subgroup. The Sequence  
691 alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25, 2078-2079.
- 692 Li, M., Schönberg, A., Schaefer, M., Schroeder, R., Nasidze, I., & Stoneking, M. (2010). Detecting  
693 heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA  
694 genomes. *American Journal of Human Genetics*, 87, 237-249.
- 695 López-Calleja, I., González, I., Fajardo, V., Martín, I., Hernández, P. E., García, T., & Martín, R.  
696 (2007). Real-time TaqMan PCR for quantitative detection of cows' milk in ewes' milk  
697 mixtures. *International Dairy Journal*, 17, 729-736.
- 698 Mafra, I., Ferreira, I. M. P. L. V. O., & Oliveira, M. B. P. P. (2008). Food authentication by PCR-  
699 based methods. *European Food Research & Technology*, 227, 649-665.
- 700 Manning, L., & Soon, J. M. (2016). Food safety, food fraud, and food defense: A fast evolving  
701 Literature. *Journal of Food Science*, 81, R823-R834.
- 702 Maudet, C., & Taberlet, P. (2001). Detection of cows' milk in goats' cheeses inferred from  
703 mitochondrial DNA polymorphism. *Journal of Dairy Research*, 68, 229-235.
- 704 Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and  
705 quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5, 621-628.
- 706 Muñoz-Colmenero, M., Martínez, J. L., Roca, A., & Garcia-Vazquez, E. (2017). NGS tools for  
707 traceability in candies as high processed food products: Ion Torrent PGM versus conventional  
708 PCR-cloning. *Food Chemistry*, 214, 631-636.



- 709 Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B.,  
710 Speicher, M.R., Zschocke, J., & Trajanoski, Z. (2014). A survey of tools for variant analysis  
711 of next-generation genome sequencing data. *Brief Bioinformatics*, *15*, 256-278.
- 712 Park, J. Y., Lee, S. Y., An, C. M., Kang, J. H., Kim, J. H., Chai, J.C., Chen, Y. Y., Kang, J. S., Ahn,  
713 J. J., Lee, Y. S., & Hwang, S. Y. (2012). Comparative study between Next Generation  
714 Sequencing Technique and identification of microarray for Species Identification within  
715 blended food products. *Biochip Journal*, *6*, 354-361.
- 716 Peter, C., Brunen-Nieweler, C., Cammann, K., & Borchers, T. (2004). Differentiation of animal  
717 species in food by oligonucleotide microarray hybridization. *European Food Research &*  
718 *Technology*, *219*, 286-293.
- 719 Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A.,  
720 Swerdlow, H. P., & Gu, Y. (2012). A tale of three next generation sequencing platforms:  
721 comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC*  
722 *Genomics*, *13*, 341.
- 723 Restani, P., Ballabio, C., Di Lorenzo, C., Tripodi, S., & Fiocchi, A. (2009). Molecular aspects of  
724 milk allergens and their role in clinical events. *Analytical and Bioanalytical Chemistry*, *395*,  
725 47-56.
- 726 Ripp, F., Krombholz, C. F., Liu, Y., Weber, M., Schäfer, A., Schmidt, B., Köppel, R., & Hankeln,  
727 T. (2014). All-Food-Seq (AFS): a quantifiable screen for species in biological samples by  
728 deep DNA sequencing. *BMC Genomics*, *15*, 639.
- 729 Robin, E. D., & Wong, R. (1988). Mitochondrial DNA molecules and virtual number of  
730 mitochondria per cell in mammalian cells. *Journal of Cellular Physiology*, *136*, 507-513.
- 731 Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H.,  
732 Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W.,  
733 Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Hube, M.,  
734 Branciforte, J. T., Stone, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M.,

- 735 Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E.,  
736 Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J.A., Namsaraev, E., McKernan, K. J.,  
737 Williams, A., Roth, G. T., & Bustillo, J. (2011). An integrated semiconductor device  
738 enabling non-optical genome sequencing. *Nature*, *475*, 348-352.
- 739 Russo, V., Fontanesi, L., Scotti, E., Tazzoli, M., Dall'Olio, S., & Davoli, R. (2007). Analysis of  
740 melanocortin 1 receptor (MC1R) gene polymorphisms in some cattle breeds: their usefulness  
741 and application for breed traceability and authentication of Parmigiano Reggiano cheese.  
742 *Italian Journal of Animal Science*, *6*, 257-272.
- 743 Spink, J., & Moyer, D. C. (2011). Defining the public health threat of food fraud. *Journal of Food*  
744 *Science*, *76*, R157-R163.
- 745 Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular  
746 Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution*, *30*, 2725-  
747 2729.
- 748 Tillmar, A. O., Dell'Amico, B., Welander, J., & Holmlund, G. (2013). A universal method for  
749 species identification of mammals utilizing next generation sequencing for the analysis of  
750 DNA mixtures. *PLoS ONE*, *8*, e83761.
- 751 Umpiérrez, A., Quirce, S., Marañón, F., Cuesta, J., García-Villamuza, Y., Lahoz, C., & Sastre, J.  
752 (1999). Allergy to goat and sheep cheese with good tolerance to cow cheese. *Clinical and*  
753 *Experimental Allergy*, *29*, 1064-1068.

754 **Figures' legends**

755

756 **Fig. 1**

757 Number of mapped reads assigned to different species by *aln* in the eight libraries and for the three  
758 amplified mtDNA fragments.

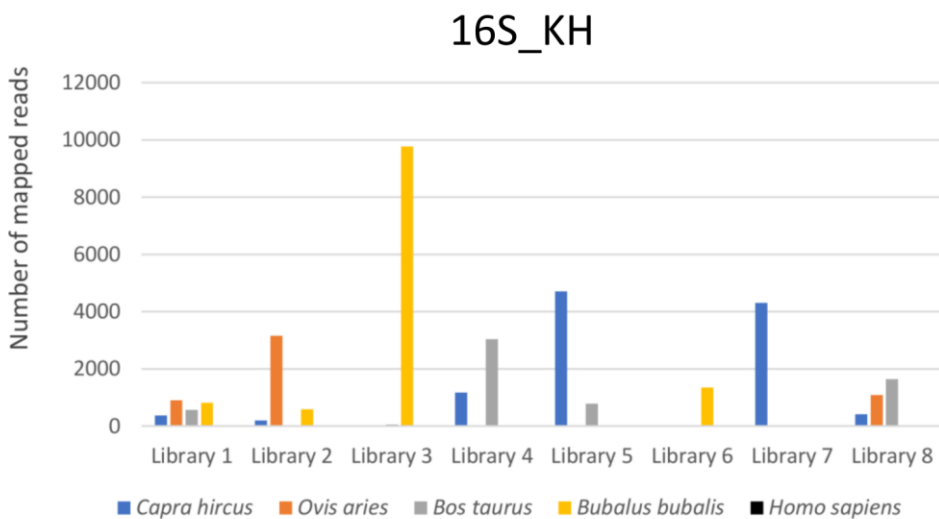
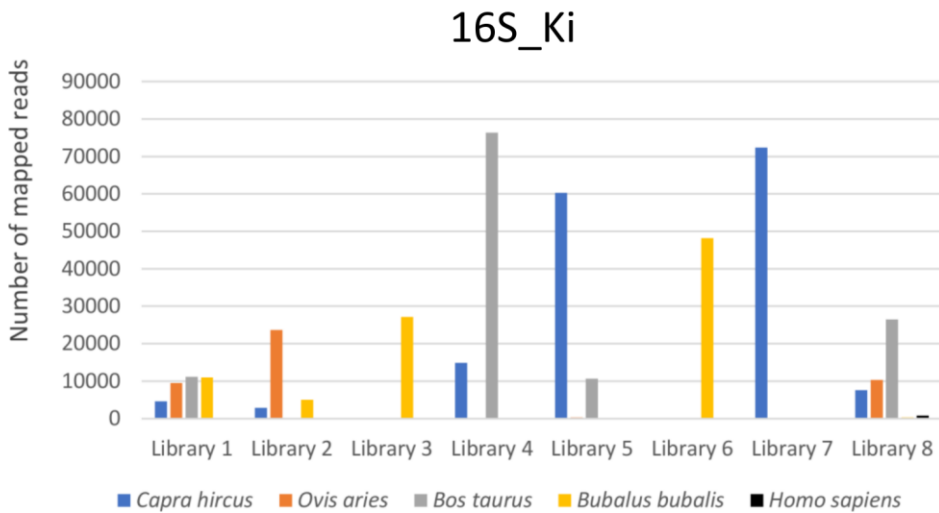
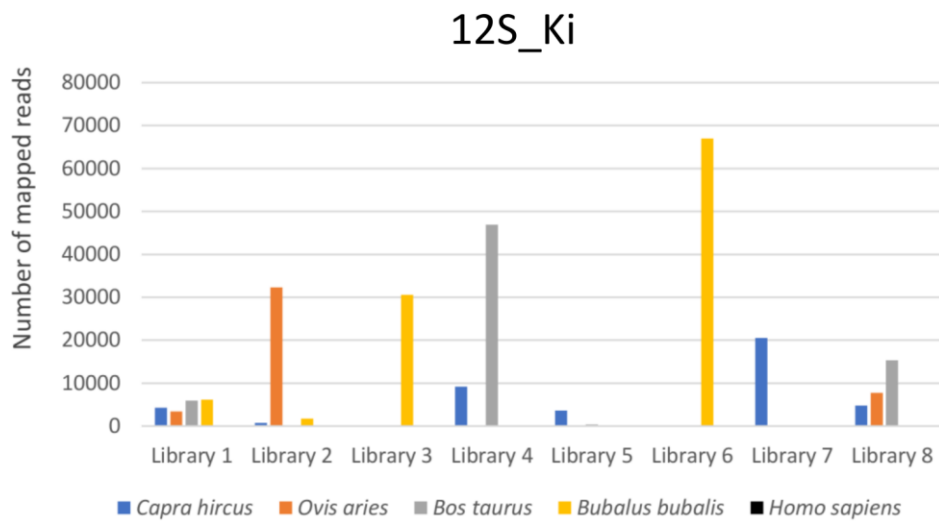
759

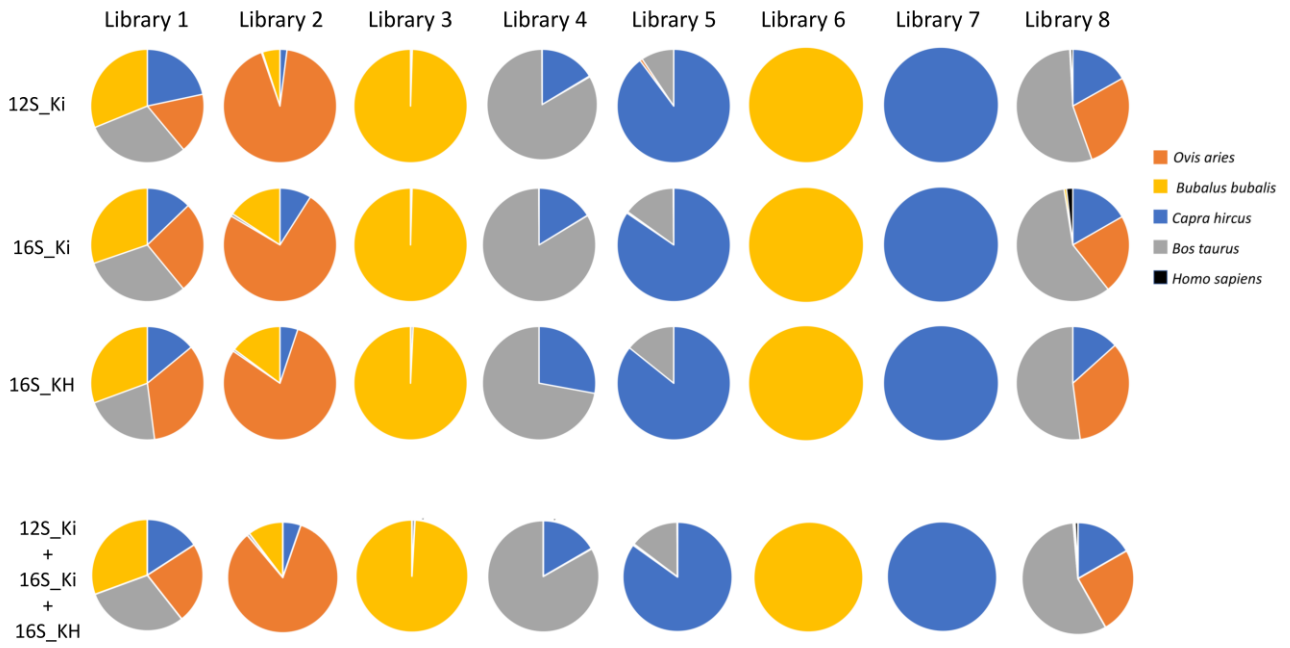
760 **Fig. 2**

761 Proportion of reads of the three mtDNA fragments (12S\_Ki, 16S\_Ki and 16S\_KH) presented  
762 separately and combined, obtained from the eight libraries, assigned to different species with the *aln*  
763 algorithm. Precise information is presented in Table S3.

764

ACCEPTED MANUSCRIPT





767

768

ACCEPTED MANUSCRIPT

**Table 1**

Primer pairs used for DNA amplification.

Primer pair name	Primers (5' to 3'): forward and reverse	Annealing T (°C)	Amplified fragment (bp) <sup>1</sup>	References
12S_Ki	CCCAAACCTGGGATTAGATACCC <i>GTTTGCTGAAGATGGCGGTA</i>	59	215-219	Kitano et al. (2007)
16S_Ki	GCCTGTTTACCAAAAACATCAC <i>CTCCATAGGGTCTTCTCGTCTT</i>	62	244-245	Kitano et al. (2007)
16S_KH	GACGAGAAGACCCTATGGAGC <i>TCCGAGGTCGCCCAACC</i>	59	112-114	Karlsson & Holmlund (2007)

<sup>1</sup> The size of the amplified regions is different in the considered species. The range is reported.

770  
771  
772  
773  
774**Table 2**

Libraries sequenced with the Ion Torrent Personal Genome Machine prepared including fragments obtained with all primer pairs. The total number of mapped reads obtained with the algorithm *aln* and considering *q30*, that performed better in this study, is reported for each library.

Library ID	Original source/product of sequenced DNA	Total no. of mapped reads ( <i>aln_q30</i> )
1	Equimolar pool of PCR products obtained separately from cattle, sheep, goat and buffalo DNA	58670
2	DNA pool containing equimolar DNA content from cattle, sheep, goat and buffalo	70589
3	DNA pool containing 90% buffalo DNA + 10% cattle DNA	67876
4	DNA extracted from a mixture containing 10% goat milk + 90% cow milk in volume	151874
5	DNA extracted from a mixture containing 50% goat milk + 50% cow milk in volume	80815
6	DNA isolated from buffalo mozzarella cheese	116842
7	DNA isolated from goat crescenza cheese	97590
8	DNA extracted from artisanal cured ricotta cheese (derived from cow milk, goat milk and sheep milk; volume ratio: □2:1:1)	76784

775  
776

777 **Table 3**  
 778 Error rate of the Ion Torrent sequencing calculated for the three amplified regions of the four dairy  
 779 species obtained from reads generated from Library 1 and aligned with *aln* algorithm.  
 780

Species	mtDNA region	Total no. of aligned nucleotides	Error rate
<i>Capra hircus</i>	12S_Ki	7513135	0.0206
<i>Capra hircus</i>	16S_KH	497965	0.0577
<i>Capra hircus</i>	16S_Ki	10742381	0.0236
<i>Ovis aries</i>	12S_Ki	4498431	0.0243
<i>Ovis aries</i>	16S_KH	105773	0.0461
<i>Ovis aries</i>	16S_Ki	5769484	0.0239
<i>Bos taurus</i>	12S_Ki	6222102	0.0220
<i>Bos taurus</i>	16S_KH	112704	0.0461
<i>Bos taurus</i>	16S_Ki	9279828	0.0210
<i>Bubalus bubalis</i>	12S_Ki	4871864	0.0203
<i>Bubalus bubalis</i>	16S_KH	416016	0.0486
<i>Bubalus bubalis</i>	16S_Ki	6876920	0.0224

781

782