Title	Phylotype-level profiling of lactobacilli in highly complex environments by means of an ITS-based metagenomic approach
Authors	Milani, Christian;Duranti, Sabrina;Mangifesta, Marta;Lugli, Gabriele A.;Turroni, Francesca;Mancabelli, Leonardo;Viappiani, Alice;Anzalone, Rosaria;Alessandri, Giulia;Ossiprandi, Maria Cristina;van Sinderen, Douwe;Ventura, Marco
Publication date	2018-05-04
Original Citation	Milani, C., Duranti, S., Mangifesta, M., Lugli, G. A., Turroni, F., Mancabelli, L., Viappiani, A., Anzalone, R., Alessandri, G., Ossiprandi, M. C., van Sinderen, D. and Ventura, M. (2018) 'Phylotype-level profiling of lactobacilli in highly complex environments by means of an ITS-based metagenomic approach', Applied and Environmental Microbiology, In Press. doi: 10.1128/aem.00706-18
Type of publication	Article (peer-reviewed)
Link to publisher's version	http://aem.asm.org/content/early/2018/04/30/AEM.00706-18.abstract - 10.1128/aem.00706-18
Rights	Open access - http://creativecommons.org/licenses/by/4.0/© 2018 American Society for Microbiology.
Download date	2025-07-06 08:32:58
Item downloaded from	https://hdl.handle.net/10468/6181

1   **Phylotype-level profiling of lactobacilli in highly complex environments by means of an ITS-**

2   **based metagenomic approach.**

3   Key words: *Lactobacillus*, microbiota, ITS, Next Generation Sequencing, Illumina.

4

5   Christian Milani[1]*, Sabrina Duranti[1]*, Marta Mangifesta[1]*, Gabriele Andrea Lugli[1], Francesca

6   Turroni[1,2], Leonardo Mancabelli[1], Alice Viappiani[3], Rosaria Anzalone[1], Giulia Alessandri[1], Maria

7   Cristina Ossiprandi[4], Douwe van Sinderen[5] and Marco Ventura[1,2]

8

9   *These authors contributed equally.

10

11   Laboratory of Probiogenomics, Department of Life Sciences, University of Parma, Italy[1];

12   Microbiome Research Hub, University of Parma, Parma, Italy[2] GenProbio srl, Parma, Italy[3];

13   Department of Veterinary Medical Science, University of Parma, Parma, Italy[4]; APC Microbiome

14   Ireland and School of Microbiology, Biosciences building, University College Cork, Ireland[5]

15

16

17

18

19   Corresponding author. Mailing address for Marco Ventura Department of Life Sciences, University

20   of Parma, Parco Area delle Scienze 11a, 43124 Parma, Italy. Phone: +39-521-905666. Fax: +39-

21   521-905604. E-mail: marco.ventura@unipr.it

Abstract

The genus *Lactobacillus* is a widespread taxon, members of which are highly relevant to functional and fermented foods, while they are also commonly present in host-associated gut and vaginal microbiota. Substantial efforts have been undertaken to disclose the genetic repertoire of all members of the genus *Lactobacillus*, yet their species-level profiling in complex matrices is still undeveloped due to the poor phylotype resolution of profiling approaches based on the 16S rRNA gene. To overcome this limitation, an ITS-based profiling method was developed to accurately profile lactobacilli at species-level. This approach encompasses a genus-specific primer pair combined with a database of ITS sequences retrieved from all available *Lactobacillus* genomes and a script for the Qiime software suite that performs all required steps to reconstruct a species-level profile. This methodology was applied to several environments, i.e., human gut and vagina, cecum of free range chickens, as well as whey and fresh cheese. Interestingly, data collected confirmed a relevant role of lactobacilli present in functional and fermented foods in defining the population harbored by the human gut, while, unsurprisingly perhaps, the cecum of free range chickens was observed to be dominated by lactobacilli characterized in birds living in natural environments. Moreover, vaginal swabs confirmed the existence of previously-hypothesized community state types, while analysis of whey and fresh cheese revealed a dominant presence of single *Lactobacillus* species used as starters for cheese production. Furthermore, application of this ITS profiling method to a mock *Lactobacillus* community allowed a minimal resolution level of <0.006 ng/µl.

43 **Importance**

44 The genus *Lactobacillus* is a large and ubiquitous taxon of high scientific and commercial

45 relevance. Despite the fact that the genetic repertoire of lactobacilli species has been extensively

46 characterized, the ecology of this genus has been explored by metataxonomic techniques that are

47 accurate down to the genus or phylogenetic group level only. Thus, the distribution of lactobacilli in

48 environmental or processed food samples is relatively unexplored. The profiling protocol described

49 here relies on the use of the Internally Transcribed Spacer to perform an accurate classification in a

50 target population of lactobacilli with <0.006 ng/µl sensitivity. This approach was used to analyze

51 five sample types collected from both human and animal host-associated microbiota as well as from

52 the cheese production chain. Availability of a tool for species-level profiling of lactobacilli may be

53 highly useful for both academic research and a wide range of industrial applications.

54   **Introduction**

55   The genus *Lactobacillus* is a widespread and diverse taxon encompassing more than 170

56   species and 17 subspecies, which are classified as Gram-positive, non-spore-forming and catalase-

57   negative facultative anaerobes (1, 2). Moreover, based on their metabolic capability to produce

58   lactic acid as the main metabolic end product of carbohydrate fermentation, lactobacilli are

59   classified as members of the Lactic Acid Bacteria (LAB). Notably, 16S rRNA gene-based

60   phylogenetic analyses revealed the existence of 22 distinct phylogenetic groups of *Lactobacillus*

61   species (24 including pediococci) (2-4).

62   Regarding their ecological distribution, lactobacilli are found in a wide range of

63   environments, including plants, water, soil, silage and different body sites of humans and other

64   animals as members of host-associated microbiomes, such as those colonizing the oral cavity, the

65   vagina and the gastrointestinal tract (GIT) (4, 5). Moreover, 37 species of this genus have been

66   granted the Qualified Presumption of Safety (QPS) status by the European Food Safety Authority

67   (EFSA) (6). Thus, they are extensively used in the food industry, in particular in fermented foods

68   due to their high performance in lactic acid fermentation coupled with high tolerance for low pH,

69   preservative and organoleptic properties, and production of exopolysaccharides that contribute to

70   the texture of foods (2, 7). In this context, members of the genus *Lactobacillus* have in recent years

71   gained significant scientific and commercial interest as health-promoting microorganisms, as

72   evidenced by the fact that 22 species encompass strains patented as probiotics in Europe (8).

73   The high commercial and scientific relevance of lactobacilli coupled to the recent

74   introduction of next-generation sequencing technologies has recently led to genome decoding of all

75   (then) known *Lactobacillus* species (3, 7). The retrieved genomic data has been exploited for

76   comparative genomic analyses, and has allowed identification of many shared or distinct genetic

77   features of this genus. Furthermore, this genomic information has permitted the reconstruction of

78   their metabolic potential, has shed light on host-microbe interactions, such as adhesion to the mucus

79 layer and modulation of the immune system of the host, and has revealed particular microbe-

80 microbe interactions with other commensals or (opportunistic) pathogens (1, 7, 8).

81   Despite the large body of data concerning the physiology and genetics of lactobacilli,

82 knowledge about the ecology and distribution in environmental or host-associated niches of

83 individual species relies mainly on culture-dependent studies. This is partly due to the resolution

84 limit of currently used metagenomic approaches. Although microbial profiling based on partial 16S

85 rRNA gene is able to discriminate between phylogenetic groups of lactobacilli due to the high

86 phylogenetic diversity of this genus, it cannot provide an accurate species-level resolution.

87 Moreover, the majority of the currebt exiting studies of lactobacilli populations based on 16S rRNA

88 gene profiling do not even perform phylogenetic group-level analyses. To offer a more refined

89 taxonomic view of lactobacilli in a given environment or sample, we developed a profiling

90 approach based on amplification of the internally transcribed spacer (ITS) sequence. Notably, due

91 to their high variability, ITS sequences have previously been exploited in a wide range of studies

92 encompassing the identification of unique species-specific restriction patterns of lactobacilli, as well

93 as the identification and characterization of *Leuconostoc* strains and for the genotyping of

94 *Streptococcus pneumoniae* strains (9-11). The developed methodology in the current work is able to

95 determine the composition of lactobacilli-containing communities down to the species level. The

96 method was validated through the analysis of a sample artificially constituted by DNA of 14

97 lactobacilli taxa at known concentration. Furthermore, we applied this methodology for the precise

98 investigation of bacterial communities harbored by human-, animal- and food-associated matrices

99 that were previously explored down to the genus-level only.

100 **Results and discussion**

101 **Analysis of ITS variability within the *Lactobacillus* genus.**

102 Genomes of 1523 strains assigned to the genus *Lactobacillus* and corresponding to 176 species

103 were retrieved from the NCBI genome database, and then processed using the MEGAnnotator

104 software (12) for prediction of rRNA genes in order to ensure the same high-quality standard for all

105 sequences of ribosomal loci included in this study (Table S1). Notably, the genomic sequences of

106 892 *Lactobacillus* strains, representing the 58.6 % of the total strain pool analyzed, did not harbor

107 complete rRNA loci, i.e. encompassing complete 5S, 16S and 23S rRNA genes. In contrast, at least

108 one complete ribosomal rRNA genes locus was identified for 631 of the 1523 analyzed strains,

109 corresponding to 70 species and a custom script was then used to extract a total of 1788 Internally

110 Transcribed Spacer (ITS) sequences. Assembly of draft genomes generally generates the collapse of

111 reads that correspond to rRNA genes into a single rRNA locus. However, availability of multiple

112 draft sequences of a given *Lactobacillus* taxon, complemented with analysis of 217 complete

113 genomes of *Lactobacillus* species, allowed us to retrieve an average of 25.5 ITS sequences per

114 species. Interestingly, 137 of the 1788 retrieved ITS sequences include stretches of >3 undefined

115 (N) nucleotides, thus highlighting a high rate of assembly-related issues and/or low-quality regions

116 in genomes deposited at the NCBI genome database. Comparative analysis of the 1651 complete

117 ITS sequences without multiple contiguous nucleotide ambiguities revealed that 92.5 % of the ITS

118 sequences range between 200 and 500 bp.

119 As previously observed for bifidobacteria (13, 14), alignment of ITS sequences from *Lactobacillus*

120 genomes shows a high level of diversity, probably due to a high mutation frequency, and

121 corresponding to a high evolutionary rate, as reflected by multiple substitutions at a given

122 nucleotide position and indicative of mutational saturation of such ITS sequences. While this

123 particularly high mutation frequency prevents phylogeny inference (15), it is suitable for

124 metagenomic amplicon-based profiling below the genus level, as previously validated for members

125 of the genus *Bifidobacterium* (13).

126

Design of a PCR primer pair for ITS-profiling of the *Lactobacillus* genus.

128 Many profiling approaches have been developed to accurately reconstruct the taxonomic

129 composition of complex bacterial communities. These include methods based on low-coverage

130 sequencing of full-length 16S rRNA genes and the use of technologies providing long reads, i.e.

131 Sanger and PacBio. Nevertheless, despite the fact that full-length sequencing of the 16S rRNA gene

132 allows high accuracy in taxonomic assignment, the low sequencing coverage permits the detection

133 only of dominant taxa and prevents profiling of bacteria present at low relative abundance in a

134 given population (10). Furthermore, the use of alternative marker genes has also been proposed,

135 though their use remains limited due to difficulties in the definition of universal primers as well as

136 in the lack of a complete reference database. The advent of next-generation sequencing,

137 characterized by high coverage and short reads, facilitated the amplification and sequencing of

138 partial 16S rRNA genes, i.e. 16S rRNA gene profiling. This metagenomic method has in recent

139 years been used as the gold standard for taxonomic characterization of environmental and host-

140 associated microbiomes. While this methodology covers all bacterial biodiversity, it is generally

141 only accurate for the reconstruction of taxonomic profiles down to genus level (16) or down to

142 phylogenetic groups in case of genera with a high level of phylogenetic diversity, e.g. the genus

143 *Lactobacillus* (3, 4) since it relies on sequencing of a small region of the whole 16S rRNA gene

144 through next-generation sequencing. To overcome this limitation and to obtain species-level

145 resolution, the use of the ITS sequence as an alternative molecular marker has been proposed (13).

146 In order to develop a universal primer pair suitable for profiling of all members of the *Lactobacillus*

147 genus, we aligned the 16S and 23S rRNA genes flanking the 1651 complete ITS sequences without

148 stretches of undefined nucleotides that were retrieved from lactobacilli genomes deposited at the

149 NCBI database. Manual inspection of the alignments allowed the identification of 'universal'

150 primers located at the 5'-end of the 16S rRNA gene and at the 3'-end of the 23S rRNA gene, i.e.,

151 Probio-lac_Uni (CGTAACAAGGTAGCCGTAGG) and Probio-lac_Rev

152 (GTYVCGTCCTTCWTCGSC), respectively (Figure 1). Sequence conservation amongst the

153 aligned 16S and 23S rRNA genes is reported in Figure 1 through a WebLogo representation. These

154 primers generate an amplicon of an average length of 380 bp covering the complete ITS region and

155 suitable for 2 X 250 bp paired-end Illumina sequencing followed by single-end bioinformatic

156 analysis of both paired reads (see below). Analysis of single-end reads provided reliable assignment

157 to species level even in cases where a tRNA gene was located within the ITS region (see below).

158 Notably, the final sequence of the primers was defined after multiple iterative alignments to the

159 Silva SSU and LSU databases (17) using the Silva TestProbe v. 3.0 tool (https://www.arb-

160 silva.de/search/testprobe/). The latter approach led to introduction of specific IUPAC bases in order

161 to maximize alignment of the primers to all currently available 16S and 23S rRNA gene sequences

162 of lactobacilli corresponding to all known species of this genus, while minimizing alignment to

163 non-lactobacilli ribosomal RNA genes. The usefulness of the Probio-lac_Uni/Probio-lac_Rev

164 primer pair was *in vitro* validated through successful amplicon generation in the case of 31

165 lactobacilli species belonging to the 23 phylogenetic groups identified previously in the genus

166 *Lactobacillus* (3, 4) (Figure S1). In contrast, no amplification was observed when the Probio-

167 lac_Uni/Probio-lac_Rev primer pair was used to amplify DNA extracted from nine non-

168 *Lactobacillus* taxa (Figure S1). Interestingly, for all tested lactobacilli we observed two PCR

169 fragments, each with a molecular size ranging from 300 to 350 bp, and 500 to 550 bp,

170 corresponding to the ITS region with and without a tRNA gene (see below for details), respectively

171 (Figure S1) (Figure 1). Such ITS patterns confirmed those displayed in previous studies targeting

172 the amplification of the ITS region of lactobacilli (18). Notably, for few taxa we observed a faint

173 amplification fragment of 500 to 550 bp, which might suggest a lower copy number of ITS regions

174 encompassing tRNA genes in the same genome.

175 The Probio-lac_Uni/Probio-lac_Rev primer pair was employed for *in silico* PCR amplification of

176 the 631 genomes of the genus *Lactobacillus* encoding at least one rRNA genes locus. This approach

177 facilitated the development of a database encompassing 1651 complete ITS sequences without

8

178   multiple ambiguous nucleotides, and flanked by partial 16S and 23S rRNA sequences, together

179   constituting the <u>L</u>actobacillus <u>ITS</u> <u>A</u>mplicon database (LITSA database).

180   Cross-alignment of all retrieved LITSA sequences using MatGAT software (19) was performed in

181   order to evaluate the level of identity between predicted amplicons (Table S2) and to evaluate

182   possible limits imposed by actual lactobacilli taxonomy to the proposed ITS profiling methodology.

183   Notably, this analysis highlighted cases in which comparison of multiple LITSA sequences from

184   the same strain showed low identity. In-depth investigation revealed that 46 of the 62 lactobacilli

185   species included in the LITSA database contain at least one ITS sequence that harbors two tRNA

186   genes (for Alanine and Isoleucine) (Figure 1). Notably, despite the fact that this prediction is limited

187   due to the small number of complete genomes available, the presence of tRNA genes in one or

188   multiple rRNA loci appears to be a common feature of genomes from members of the *Lactobacillus*

189   genus.

190   Furthermore, cross-alignment analysis also revealed that the majority of the 62 *Lactobacillus*

191   species, for which a complete LITSA sequence was available, can be discriminated (Table S2), with

192   the exception of putatively misclassified strains and/or species (see below). In this context, despite

193   the fact that lactobacilli are known to possess a very high level of phylogenetic diversity (3, 4),

194   strains corresponding to 18 species showed an average LITSA sequence identity of >97 % with at

195   least one other *Lactobacillus* species, thus showing a very close phylogenetic relationship between

196   such taxa (Table 1). Amongst lactobacilli, *Lactobacillus casei* and *Lactobacillus paracasei* strains

197   possess an average LITSA sequence identity of 99 %, while the amplicon sequences of

198   *Lactobacillus pentosus*, *Lactobacillus plantarum* and *Lactobacillus paraplantarum* strains show up

199   to 100 % identity (Table S2). An in-depth analysis of each strain revealed that 23 of the 25 strains

200   classified as *L. casei* share an average LITSA sequence identity ≤96.1 % with the type strain *L.*

201   *casei* ATCC 393, while the average identity with the type strain *L. paracasei* ATCC 394 is ≥99.7 %

202   (Table S2). In contrast, the putative lactobacilli species *Lactobacillus* sp. FMNP02 shares 99.7 %

9

203 identity with *L. casei* ATCC 393 (Table S2), thus representing a possible misclassification of the

204 latter strain.

205 In our attempts to obtain insights into the phylogeny of *L. pentosus*, *L. plantarum* and *L.*

206 *paraplantarum*, we observed an average LITSA sequence identity of 98.9 % between *L. pentosus*

207 and *L. plantarum* strains (Table S2). Moreover, the two strains of *L. paraplantarum*, for which we

208 were able to predict an rRNA gene locus, show an average LITSA identity of 99.5 % with *L.*

209 *plantarum* strains (Table S2), thus indicating that such taxa may belong to the same species and

210 therefore cannot be discriminated using metataxonomic techniques. Nevertheless, evaluation of the

211 average nucleotide identity is needed to confirm this hypothesis. Furthermore, we could not retrieve

212 an *in silico* Probio-lac_Uni/Probio-lac_Rev-corresponding amplicon for the type strains of *L.*

213 *pentosus* and *L. paraplantarum* due to absence of a complete ITS region in the deposited genomes,

214 and we were therefore unable to evaluate their amplicon identity with the LITSA sequences of *L.*

215 *plantarum* strains.

216 Notably, these observations suggest that major issues in the classification of the genus

217 *Lactobacillus* still exist, resulting in the unfeasibility of distinguish a number of species through ITS

218 profiling. Thus, as has been proposed previously, it is desirable that a re-evaluation of the taxonomy

219 of lactobacilli is undertaken based on a phylogenomic approach (20, 21), as was also corroborated

220 by recent studies (3, 4, 7).

221

222 **Development of a bioinformatic tool for ITS-profiling of the *Lactobacillus* genus.**

223 The length of the amplicon produced by the Probio-lac_Uni/Probio-lac_Rev primer pair may

224 exceed 600 bp, particularly when tRNA-encoding sequences are present in the ITS sequence. Thus,

225 sequencing produced non-overlapping paired-end reads even with the maximum length obtainable

226 using Next-Generation Illumina sequencing, i.e. 2 X 250 bp paired-end reads, using the MiSeq

227 Reagents Kit v3 600 cycles chemistry. Nevertheless, each forward and reverse read covers 42 and

228 60 nucleotides corresponding to the 16S rRNA gene 3'-end and the 23S rRNA gene 5'-end,

229   respectively, which are followed by 190-208 bp of hyper-variable ITS sequence suitable for

230   profiling at species-level (Figure 1). Thus, we developed a package for QIIME software suite v1.9.1

231   (22) that encompasses the LITSA database and a bash script for analysis of both forward and

232   reverse reads of the *Lactobacillus* ITS profiling data (probiogenomics.unipr.it/pbi). Notably, the

233   LITSA database will be updated regularly to include additional ITS sequences as new lactobacilli

234   genome sequences become available, thus increasing the number of lactobacilli species that can be

235   profiled. The script performs quality-filtering, *de novo* OTU clustering at 100 % identity and

236   taxonomic classification of OTU reference sequences through RDP classifier with a confidence

237   level of 0.80. Notably, these cut-off values permit discrimination of closely related taxa. Due to the

238   average size of the amplicon, the paired-end reads are not joined prior classification. Instead, the

239   script analyzes both the forward and the reverse reads altogether and provides an average profile.

240   Notably, the different number of rRNA loci predicted in the genomes of *Lactobacillus* species may

241   generate biases in the retrieved profiles. Thus, we evaluated the average number of ITS regions

242   present in the 217 available complete *Lactobacillus* genomes. This analysis provided data for

243   normalization of 45 of the 62 species of lactobacilli for which a LITSA sequence could be retrieved.

244   Moreover, the average number of rRNA genes loci of the remaining 17 species with only draft

245   genomes was set at 5.6, i.e., the average obtained for all the species with at least a complete

246   genome. Notably, the *Lactobacillus* ITS profiling analysis script includes a normalization step

247   based on the number of rRNA genes loci predicted for all the 62 *Lactobacillus* species for which a

248   LITSA sequence could be retrieved. The output produced by the script is summarized in the

249   "output" folder, which contains the predicted taxonomic profile based on the LITSA database (both

250   non-normalized and normalized for the number of rRNA loci) and the OTU table in tabular text

251   format that reports the reference sequence and associated taxonomy. All *Lactobacillus* ITS profiles

252   reported in this manuscript correspond to the average between forward and reverse read profiles

253   after normalization for the number of predicted rRNA genes loci.

254

11

Assessing detection sensitivity and accuracy using the *Lactobacillus* ITS profiling protocol

255   In order to provide an evaluation of the sensitivity and accuracy of the Probio-lac_Uni/Probio-

256   lac_Rev primer pair, 14 *Lactobacillus* type strains were employed to artificially compose a mock

257   community (Table S3). The DNA extracted from each taxon grown in pure culture was added to the

258   mix at known amount, ranging from 0.006 ng to 50 ng of DNA, corresponding to 0.006 % to 50 %

259   of the total DNA pool (Figure 2). Sequencing of the mock sample was performed using an Illumina

260   MiSeq with 2X250 bp chemistry, producing 45,146 quality-filtered paired-end reads. Interestingly,

261   *Lactobacillus* ITS profiling of this dataset successfully profiled all *Lactobacillus* species included in

262   this sample, except *Lactobacillus vaginalis* and *Lactobacillus pontis*, for which we could not

263   retrieve a LITSA sequence from analysis of available genome sequences (Figure 2). Thus, even

264   though the Probio-lac_Uni/Probio-lac_Rev primer pair produces an amplicon for these species, the

265   latter cannot be taxonomically classified due to absence of *L. vaginalis* and *L. pontis* in the present

266   version of the LITSA database. This is a temporary limitation and the LITSA database will be

267   updated regularly (probiogenomics.unipr.it/pbi) to include LITSA sequences of newly sequenced

268   genomes in order to cover all the lactobacilli species that currently cannot be profiled. Moreover,

269   comparison of the retrieved profile with the expected composition revealed a strong correlation for

270   each taxon with few discrepancies (Figure 2). The causes of such differences between expected and

271   observed relative abundance may be imputed to the lack of sufficient information in the LITSA

272   database, at this time, regarding the average number of rRNA loci per genome used for

273   normalization of the ITS profiling data.

274   Furthermore, since PCR amplicon size has been identified as a source of bias in ITS-based profiling

275   studies of fungi (23), we evaluated the presence of possible biases introduced by amplification of

276   lactobacilli ITS sequences of different length due to the presence or absence of tRNA genes (see

277   above). The 14 *Lactobacillus* species that constitute the mock community (Table S3) were

278   subjected to manual characterization of corresponding rRNA loci. Notably, the ten species for

279   which a complete genome was available, confirmed what had been observed for the *in vitro* PCR,

255   **Assessing detection sensitivity and accuracy using the *Lactobacillus* ITS profiling protocol**

256   In order to provide an evaluation of the sensitivity and accuracy of the Probio-lac_Uni/Probio-

257   lac_Rev primer pair, 14 *Lactobacillus* type strains were employed to artificially compose a mock

258   community (Table S3). The DNA extracted from each taxon grown in pure culture was added to the

259   mix at known amount, ranging from 0.006 ng to 50 ng of DNA, corresponding to 0.006 % to 50 %

260   of the total DNA pool (Figure 2). Sequencing of the mock sample was performed using an Illumina

261   MiSeq with 2X250 bp chemistry, producing 45,146 quality-filtered paired-end reads. Interestingly,

262   *Lactobacillus* ITS profiling of this dataset successfully profiled all *Lactobacillus* species included in

263   this sample, except *Lactobacillus vaginalis* and *Lactobacillus pontis*, for which we could not

264   retrieve a LITSA sequence from analysis of available genome sequences (Figure 2). Thus, even

265   though the Probio-lac_Uni/Probio-lac_Rev primer pair produces an amplicon for these species, the

266   latter cannot be taxonomically classified due to absence of *L. vaginalis* and *L. pontis* in the present

267   version of the LITSA database. This is a temporary limitation and the LITSA database will be

268   updated regularly (probiogenomics.unipr.it/pbi) to include LITSA sequences of newly sequenced

269   genomes in order to cover all the lactobacilli species that currently cannot be profiled. Moreover,

270   comparison of the retrieved profile with the expected composition revealed a strong correlation for

271   each taxon with few discrepancies (Figure 2). The causes of such differences between expected and

272   observed relative abundance may be imputed to the lack of sufficient information in the LITSA

273   database, at this time, regarding the average number of rRNA loci per genome used for

274   normalization of the ITS profiling data.

275   Furthermore, since PCR amplicon size has been identified as a source of bias in ITS-based profiling

276   studies of fungi (23), we evaluated the presence of possible biases introduced by amplification of

277   lactobacilli ITS sequences of different length due to the presence or absence of tRNA genes (see

278   above). The 14 *Lactobacillus* species that constitute the mock community (Table S3) were

279   subjected to manual characterization of corresponding rRNA loci. Notably, the ten species for

280   which a complete genome was available, confirmed what had been observed for the *in vitro* PCR,

281   i.e. presence of longer ITS sequences that encompass two tRNA genes (Figure 1; Figure S1).

282   Interestingly, the different intensities observed in the PCR fragments, i.e. 300-350 bp and 500-550

283   bp (Figure S1), did not influence

284   expected relative abundance of the mock community (Figure 2). Notably, detection of *Lactobacillus*

285   *rhamnosus* whose concentration in the mock community is 0.006 ng/µl indicates that the limit of

286   detection of the lactobacilli ITS profiling is <0.006 ng/µl, corresponding to $1.85*10^3$ cells/ µl.

287

288   **Validation of the *Lactobacillus* ITS profiling protocol through analysis of samples from**

289   **multiple environments.**

290   *Lactobacillus* is a highly diverse microbial genus, members of which are found in a wide range of

291   environments (5). To perform a comprehensive testing of the performances of the *Lactobacillus* ITS

292   profiling protocol, we analyzed a total of 25 samples encompassing five human faecal samples, five

293   human vaginal swab samples , five free range chicken cecal samples, five whey samples and five

294   parmesan cheese samples   (Table S4). Sequencing was performed with an Illumina MiSeq

295   instrument using 2x250 bp chemistry, producing an average of 15,529 forward and 15,293 reverse

296   quality-filtered reads per sample (Table S4).

297   Interestingly, analysis of the human faecal samples revealed the presence of human gut colonizers,

298   such as *Lactobacillus rhamnosus*, along with a range of lactobacilli used in functional or fermented

299   foods that are typically part of the human diet, such as *L. plantarum*, *Lactobacillus helveticus,*

300   *Lactobacillus delbrueckii* and *Lactobacillus sakei* (Figure 3).

301   Moreover, the obtained profiles of the five human vaginal swab samples confirmed the proposed

302   existence of community state types (CSTs) of the vaginal microbiota dominated by specific

303   *Lactobacillus* taxa (24, 25). In fact, HV1 is dominated by *Lactobacillus gasseri*, while

304   *Lactobacillus iners* and *Lactobacillus crispatus* are the most abundant lactobacilli taxa in HV2/HV5

305   and HV3/HV4, respectively (Figure 3). Furthermore, in all five reconstructed human vaginal

306   profiles, *L. helveticus* is the second most abundant *Lactobacillus* species, as observed in the

13

307   aforementioned CSTs (24, 25) (Figure 3). Thus, based on the classification proposed by DiGiulio et

308   al. (24), HV1 can be classified as a CST 2, while HV2/HV5 falls within the CST 3, whereas

309   HV3/HV4 can be attributed to CST 1.

310   To demonstrate the relevance of an efficient methodology for precise cataloguing of the

311   *Lactobacillus* species for which a complete LITSA sequence is available in different environments,

312   we analyzed five free range chickens cecal samples. The retrieved profiles revealed a high relative

313   abundance (ranging from a total of 53.1 % to 96.8 %) of *Lactobacillus* species previously

314   characterized in poultry or other birds, such as *Lactobacillus salivarius, Lactobacillus reuteri,*

315   *Lactobacillus ingluviei, Lactobacillus amylovorus, Lactobacillus agilis, Lactobacillus aviarius* and

316   *Lactobacillus johnsonii* (26-33) (Figure 3). Notably, samples FRC1, FRC2 and FRC3 showed a

317   similar profile with high abundance of *L. salivarius*, *L. ingluviei* and *L. amylovorus*, reflecting the

318   fact that they were kept in the same hen house (Figure 3). Accordingly, samples FRC4 and FRC5,

319   collected in two additional hen houses, showed different profiles characterized by high abundance

320   of *L. aviarius* and *L. johnsonii*, respectively (Figure 3).

321   For milk and milk-related products, profiling of five whey and five fresh parmesan cheeses (at 1

322   day of ripening) samples revealed, as expected, similar profiles dominated by *L. helveticus* and *L.*

323   *delbrueckii* (Figure 3), which represent two lactobacilli species typically used as starter cultures for

324   the production of cheese (34). These data indicate that the *Lactobacillus* ITS profiling approach also

325   represents a valuable tool for monitoring the population of lactobacilli across the cheese production

326   chain.

327   Results obtained from ITS-profiling were also compared to profiles reconstructed through analysis

328   of OTUs generated at 99 % identity from 16S rRNA profiling data (Figure 3) (Table S5). Notably,

329   only OTUs classified as lactobacilli have been included in the representation, thus the relative

330   abundance of unclassified lactobacilli reported in the bar plot do not include additional OTUs that

331   could not be attributed to this genus. Moreover, lactobacilli species whose relative abundance is

332   below 5 % in each sample were collapsed under "Others <5 %" in the bar plot representation.

14

333   Interestingly, the ITS profiling approach provided a more accurate species-level reconstruction of

334   the lactobacilli populations when used to analyze human faecal and vaginal samples as well as free

335   range chicken faecal samples. Moreover, it confirmed and partially improved the simple lactobacilli

336   community of whey and fresh Parmesan cheese samples observed through 16S rRNA gene

337   profiling. In fact, differences in the profiles obtained through 16S rRNA gene and ITS profiling can

338   be observed in all cases (Figure 3) (Table S5). Such differences are caused by the limited number of

339   *Lactobacillus* species that could be discriminated based on partial16S rRNA gene sequence respect

340   to ITS sequence (Figure 3) (Table S5).

341   Altogether, these results confirm the performance of the *Lactobacillus* ITS profiling protocol

342   observed from analysis of the artificial sample and validate their use, complemental to 16S rRNA

343   gene profiling, for analysis of a wide range of complex environmental and host-associated matrices.

344

345   **Conclusions**

346   We developed a newly designed method for characterization of the *Lactobacillus* population in

347   complex environments based on the use of the internally transcribed spacer (ITS), which represents

348   a hypervariable region located between the 16S and the 23S rRNA genes that allows high-accuracy

349   species-level profiling. The accuracy and sensitivity of this method allowed profiling of complex

350   communities of lactobacilli with a successful identification of taxa with abundance of $1.85*10^3$

351   cells/ μl, which is even lower to what was previously identified for a similar approach developed for

352   the profiling of bifidobacterial communities (13). Notably, despite the fact that the current LITSA

353   database allows the precise profiling of just 62 species, the ITS-profiling approach represents a new

354   metagenomic tool for species-level profiling of complex lactobacilli communities that complements

355   phylogenetic group assignments that can be obtained from 16S rRNA gene profiling data.

356   Moreover, the database will be regularly updated to represent additional lactobacilli species as

357   genomes encompassing complete LITSA sequences are becoming available. When the ITS

358   lactobacilli profiling method was applied to different biological samples, encompassing the stool of

15

359   human as well as birds, vaginal swabs and cheese, it allowed the reconstruction of the cataloguing

360   of lactobacilli communities residing in these environments. Altogether, these results highlight that

361   ITS-mediated profiling of populations of lactobacilli could be useful not only for academic

362   purposes, but also for industrial applications such as tracing the microbial composition of probiotic

363   products based on lactobacilli as well as of starter cultures in food manufacture.

364

365   **Material and methods**

366   **Sample collection**

367   In the frame work of a more extensive bacterial cataloguing project, this study enrolled stool,

368   vaginal swab, fresh parmesan cheese (one day of ripening), whey and cecal (from free range

369   chickens) samples.

370   Five fresh stool samples obtained from human healthy volunteers and five cecal samples retrieved

371   from free range chickens were immediately frozen upon collection at -80°C until processing for

372   DNA extraction. The DNA extraction was performed using the QIAamp DNA Stool Mini Kit

373   following the manufacturer's instructions (Qiagen, Manchester, UK). Additionally, five vaginal

374   swab samples were collected in sterile tubes containing 1 ml of DNA-RNA shield from ZYMO

375   Research until bacterial DNA extraction using ZymoBIOMICS™ DNA Miniprep Kit (ZYMO

376   Research). Furthermore, 10 ml samples of whey and 2-4 gr of fresh parmesan cheese were collected

377   in sterile tubes and the DNA was extracted using the DNeasy Mastitis Mini Kit (Qiagen Ltd,

378   Strasse, Germany) following the manufacturer's instructions (Qiagen Ltd). Notably, whey samples

379   and cheese samples at one day of ripening were collected from the same Parmesan cheese producer

380   in Parma, Italy.

381   **Ethical statement.** This study was carried out in accordance with the recommendations of the

382   ethical committee of the University of Parma and was approved by the "Comitato di Etica

383   Università degli Studi di Parma", Italy. All animal procedures were performed according to national

384   guidelines (Decreto legislativo 26/2014).

16

385 **Bacterial growth conditions and DNA extraction.** Type strains of several lactobacilli taxa (Table

386 S3) were growth in Man-Rogosa-Sharpe (MRS) medium (Scharlau Chemie) supplemented with

387 0.05 % (w/v) L-cysteine hydrochloride and incubated in an anaerobic atmosphere (2.99 % H2,

388 17.01 % CO2 and 80 % N2) in a chamber (Concept 400; Ruskin) at 37°C for 24 h. In addition, nine

389 non-lactobacilli microorganisms were used in this study. These included *Bifidobacterium bifidum*

390 LMG11041, which was cultivated in MRS broth as *Lactobacillus* strains; *Collinsella intestinalis*

391 DSM 13280, *Escherichia coli* LMG 2092, and *Klebsiella pneumoniae* CECT 143, which were

392 grown in de MRS broth (Difco, Detroit, MI) supplemented with 0.05% (w/v) l-cysteine (MRSC;

393 Sigma, St. Louis, MO). *Prevotella copri* DSM 18205 and *Blautia coccoides* DSM 935 were

394 cultivated in a combination of Reinforced Clostridial Broth (Merck, Darmstadt, Germany) and

395 Brain-Heart Infusion (Difco), supplemented with 5% (v/v) heat-inactivated fetal bovine serum

396 (LabClinics, Barcelona, Spain) respectively. For *Bacteroides thetaiotaomicron* DSMZ 2079, the

397 latter medium was supplemented with 0.005 % hemin (Sigma) and 0.005 % Vitamin K1 (Sigma).

398 *Faecalibacterium prausnitzii* DSM 17677 was grown in Wilkins-Chalgren Anaerobe broth (Merck),

399 following the recommendations included in the DSMZ medium 339. Finally, an active culture of

400 *Methanobrevibacter smithii* DSM 861 grown in *Methanobacterium* medium (DSMZ 119) was

401 directly supplied by DSMZ.

402 Bacterial DNA was extracted using GenEluteTM Bacterial Genomic DNA kits (SIGMA-

403 ALDRICH) following the manufacturer's instructions. Taxonomic identity of the microorganisms

404 was validated by sequencing the V3 variable region of the 16S rRNA gene using primer pair

405 Probio_Uni and/Probio_Rev (14).

406 *Lactobacillus* mock community

407 The cultures of fourteen different *Lactobacillus* strains were grown separately on Man-Rogosa-

408 Sharpe (MRS) medium (Scharlau Chemie) supplemented with 0.05 % (w/v) L-cysteine

409 hydrochloride and incubated in an anaerobic atmosphere (2.99 % H2, 17.01 % CO2 and 80 % N2)

410 in a chamber (Concept 400; Ruskin) at 37°C until they reached late log phase. The bacteria were

17

411 enumerated by counting colonies on solid medium and the optical density at 600 nm was

412 determined. The final bacterial cell concentration was approximately 10^7 cfu/ml. Chromosomal

413 DNA of each strains was extracted as previously described and subsequently mixed.

414 Specifically, the mock community consists of a pool of known concentration of fourteen different

415 *Lactobacillus* strains to obtain the final quantity of DNA indicated in Table S3. Furthermore, the

416 mix was prepared by combining equal volumes (20 μL) of DNA.

417 The DNA from the mix was diluted to produce a final DNA concentration of 2 ng/μL, and 4 μL of

418 these dilutions were used in each PCR reaction. For the PCR reaction, the primer pair Probio-

419 lac_Uni/Probio-lac_Rev was used and the generated amplicons were sequenced using Illumina

420 MiSeq as described below.

421 ***Lactobacillus* ITS-specific primer design and gene amplification**

422 The bioinformatics platforms MEGAnnotator (10) and METAnnotatorX (unpublished data) were

423 used to perform 16S and 23S rRNA genes prediction in all 1523 sequenced lactobacilli genomes

424 deposited   at   the   NCBI   Genomes   database.   Primers   Probio-lac_Uni

425 (CGTAACAAGGTAGCCGTAGG)/Probio-lac_Rev   (GTYVCGTCCTTCWTCGSC)   were

426 manually designed based on the alignment of all 16S and 23S rRNA sequences to generate an

427 amplicon encompassing the 3'-end of the 16S rRNA gene, the ITS region and the 5'-end of the 23S

428 rRNA gene. Specificity test was performed using the Silva TestProbe v. 3.0 tool (https://www.arb-

429 silva.de/search/testprobe/) that allows alignment of primers sequences to the Silva SSU and LSU

430 databases (15). A custom bioinformatics script was then used to create a database of all the Probio-

431 lac_Uni/Probio-lac_Rev-generated lactobacilli ITS amplicon sequences (LITSA database),

432 encompassing the Internally Transcribed Spacer (ITS) region and partial 16S and 23S rRNA genes.

433 The PCR conditions used for *Lactobacillus* ITS profiling using the Probio-lac_Uni/Probio-lac_Rev

434 primer pair were 5 min at 95 °C, 30 cycles of 30 s at 95 °C, 30 s at 58 °C, and 40 s at 72 °C,

435 followed by 10 min at 72 °C. Amplification was carried out using a Verity Thermocycler (Applied

436 Biosystems). The integrity of the PCR amplicons was analyzed by gel electrophoresis. An

18

437 additional specificity test was performed by PCR using the DNA extracted from all known

438 *Lactobacillus* species as well as *B. bifidum* ATCC11041, *C. intestinalis* DSM 13280, *E. coli* LMG

439 2092*, K. pneumoniae* CECT 143*, P. copri* DSM 18205, *Bl. coccoides DSM 935, Bc.*

440 *thetaiotaomicron* DSMZ 2079, *F. prausnitzii* DSM 17677 and *M. smithii* DSM 861.

441 WebLogo representation of primer sequence conservation among the retrieved 16S and 23S rRNA

442 genes flanking complete ITS sequences was obtained through the WebLogo website

443 (http://weblogo.berkeley.edu/) (35).

444 **Illumina MiSeq sequencing of ITS gene-based amplicons**

445 Illumina adapter overhang nucleotide sequence was added to the PCR amplicons obtained following

446 amplification of the ITS region, as previously described (13). The library of ITS amplicons was

447 prepared following the 16S Metagenomic Sequencing Library Preparation Protocol (Part No.

448 15044223 Rev. B-Illumina). Sequencing was performed using an Illumina MiSeq sequencer with

449 MiSeq Reagent Kit v3 chemicals.

450 **ITS-based microbiota analysis**

451 Fastq files obtained from metagenomic sequencing of each sample were analyzed using a custom

452 script for QIIME software suite (22) and the LITSA database available at

453 (http://probiogenomics.unipr.it/pbi/index.html).

454 Input data were processed in the following steps: filtering of the reads based on length > 100 nt

455 (primers included) to avoid primer dimers, overall quality > 25 and the presence of forward and

456 reverse primers in the forward and reverse reads, respectively, creation of OTUs constituted by

457 identical sequences using prefix_suffix method and removal of OTUs represented by < 10

458 sequences. Taxonomy assignment was performed using RDP method (RDP classifier with a

459 confidence level of 0.80) and the LITSA database constituted by ITS sequences retrieved from the

460 1523 *Lactobacillus* genomes available at the NCBI Genome database. This script is easily

461 modifiable to obtain a profiling based on a different sequence, though will depend on the

462 availability of a corresponding database.

19

463 **Evaluation of the sensitivity of the Probio-lac_Uni/Probio-lac_Rev primer pair**

464 The artificial sample used for the evaluation of the detection sensitivity and accuracy of the Probio-

465 lac_Uni/Probio-lac_Rev primer set was generated using known DNA amounts, ranging from 50 to

466 0.006 ng, of 14 different *Lactobacillus* taxa (Table S3).

467 **Microbiota identification by 16S rRNA gene- amplification, -sequencing and data analysis**.

468 Partial 16S rRNA gene sequences were amplified from extracted DNA using primer pair

469 Probio_Uni / Probio_Rev, which target the V3 region of the 16S rRNA gene sequence (16). 16S

470 rRNA gene amplification and amplicon checks were carried out as previously described (16). 16S

471 rRNA gene sequencing was performed using a MiSeq (Illumina) at the DNA sequencing facility of

472 GenProbio srl (www.genprobio.com) according to a previously reported protocol (16). Following

473 sequencing, the .fastq files were processed using a custom script based on the QIIME software suite

474 (22). Paired-end read pairs were assembled to reconstruct complete Probio_Uni / Probio_Rev

475 amplicons. Quality control retained sequences with a length between 140 and 400 bp and mean

476 sequence quality score >20 while sequences with homopolymers >7 bp and mismatched primers

477 were omitted. 16S rRNA gene Operational Taxonomic Units (OTUs) were defined at $\geq$ 99 %

478 sequence homology using uclust (36) and OTUs with less than 10 sequences were filtered. All reads

479 were classified to the lowest possible taxonomic rank using QIIME (37) and a reference dataset

480 from the SILVA database (Quast et al., 2013).

481

482 **Nucleotide sequence accession numbers**

483 The raw ITS and 16S rRNA gene sequences reported in this article have been deposited in the

484 NCBI Short Read Archive (SRA) under the accession number PRJNA434072.

485

486 **Acknowledgements**

20

References

1. Goldstein EJ, Tyrrell KL, Citron DM. 2015. Lactobacillus species: taxonomic complexity and controversial susceptibilities. Clin Infect Dis 60 Suppl 2:S98-107.

2. Salvetti E, Torriani S, Felis GE. 2012. The Genus Lactobacillus: A Taxonomic Update. Probiotics Antimicrob Proteins 4:217-26.

3. Zheng J, Ruan L, Sun M, Ganzle M. 2015. A Genomic View of Lactobacilli and Pediococci Demonstrates that Phylogeny Matches Ecology and Physiology. Appl Environ Microbiol 81:7233-43.

4. Duar RM, Lin XB, Zheng J, Martino ME, Grenier T, Perez-Munoz ME, Leulier F, Ganzle M, Walter J. 2017. Lifestyles in transition: evolution and natural history of the genus Lactobacillus. FEMS Microbiol Rev 41:S27-S48.

5. Walter J. 2008. Ecological role of lactobacilli in the gastrointestinal tract: implications for fundamental and biomedical research. Appl Environ Microbiol 74:4985-96.

6. Ricci A, Allende A, Bolton D, Chemaly M, Davies R, Girones R, Koutsoumanis K, Herman L, Lindqvist R, Nørrung B, Robertson L, Ru G, Sanaa M, Simmons M, Skandamis P, Snary E, Speybroeck N, Ter Kuile B, Threlfall J, Wahlström H, Cocconcelli PS, Klein G, Peixe L, Maradona MP, Querol A, Suarez JE, Sundh I, Vlak J, Correia S, Fernández Escámez PS. 2017. Update of the list of QPS-recommended biological agents intentionally added to food or feed as notified to EFSA 5: suitability of taxonomic units notified to EFSA until September 2016. EFSA Journal 15:e04663-n/a.

7. Sun Z, Harris HM, McCann A, Guo C, Argimon S, Zhang W, Yang X, Jeffery IB, Cooney JC, Kagawa TF, Liu W, Song Y, Salvetti E, Wrobel A, Rasinkangas P, Parkhill J, Rea MC, O'Sullivan O, Ritari J, Douillard FP, Paul Ross R, Yang R, Briner AE, Felis GE, de Vos WM, Barrangou R, Klaenhammer TR, Caufield PW, Cui Y, Zhang H, O'Toole PW. 2015. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. Nat Commun 6:8322.

8. Salvetti E, O'Toole PW. 2017. The Genomic Basis of Lactobacilli as Health-Promoting Organisms. Microbiol Spectr 5.

9. Sandes SH, Alvin LB, Silva BC, Zanirati DF, Jung LR, Nicoli JR, Neumann E, Nunes AC. 2014. Lactobacillus species identification by amplified ribosomal 16S-23S rRNA restriction fragment length polymorphism analysis. Benef Microbes 5:471-81.

10. Barrangou R, Yoon SS, Breidt F, Jr., Fleming HP, Klaenhammer TR. 2002. Identification and characterization of Leuconostoc fallax strains isolated from an industrial sauerkraut fermentation. Appl Environ Microbiol 68:2877-84.

11. Moore JE, Hirayama J, Hayashi K, Mason C, Coulter W, Matsuda M, Goldsmith CE. 2018. Examination of 16S-23S rRNA intergenic spacer region (ISR) heterogeneity in a population of clinical Streptococcus pneumoniae- a new laboratory epidemiological genotyping tool to aid outbreak analysis. Br J Biomed Sci doi:10.1080/09674845.2017.1382025:1-3.

12. Lugli GA, Milani C, Mancabelli L, van Sinderen D, Ventura M. 2016. MEGAnnotator: a user-friendly pipeline for microbial genomes assembly and annotation. FEMS Microbiol Lett 363.

13. Milani C, Lugli GA, Turroni F, Mancabelli L, Duranti S, Viappiani A, Mangifesta M, Segata N, van Sinderen D, Ventura M. 2014. Evaluation of bifidobacterial community composition in the human gut by means of a targeted amplicon sequencing (ITS) protocol. FEMS Microbiol Ecol 90:493-503.

14. Milani C, Mancabelli L, Lugli GA, Duranti S, Turroni F, Ferrario C, Mangifesta M, Viappiani A, Ferretti P, Gorfer V, Tett A, Segata N, van Sinderen D, Ventura M. 2015. Exploring Vertical Transmission of Bifidobacteria from Mother to Child. Appl Environ Microbiol 81:7078-87.

15. Philippe H, Adoutte A. 1996. What Can Phylogenetic Patterns Tell Us About the Evolutionary Process Generating Biodiversity? Oxford University Press, Oxford, UK.

16. Milani C, Hevia A, Foroni E, Duranti S, Turroni F, Lugli GA, Sanchez B, Martin R, Gueimonde M, van Sinderen D, Margolles A, Ventura M. 2013. Assessing the fecal microbiota: an optimized ion torrent 16S rRNA gene-based analysis protocol. PLoS One 8:e68739.

545   17.   Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. 2013. The SILVA
546   ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic
547   Acids Res 41:D590-6.
548   18.   Chebenova-Turcovska V, Zenisova K, Kuchta T, Pangallo D, Brezna B. 2011. Culture-independent
549   detection of microorganisms in traditional Slovakian bryndza cheese. Int J Food Microbiol 150:73-8.
550   19.   Campanella JJ, Bitincka L, Smalley J. 2003. MatGAT: an application that generates similarity/identity
551   matrices using protein or DNA sequences. BMC Bioinformatics 4:29.
552   20.   Lugli GA, Mangifesta M, Duranti S, Anzalone R, Milani C, Mancabelli L, Alessandri G, Turroni F,
553   Ossiprandi MC, van Sinderen D, Ventura M. 2018. Phylogenetic classification of six novel species
554   belonging to the genus Bifidobacterium comprising Bifidobacterium anseris sp. nov.,
555   Bifidobacterium criceti sp. nov., Bifidobacterium imperatoris sp. nov., Bifidobacterium italicum sp.
556   nov., Bifidobacterium margollesii sp. nov. and Bifidobacterium parmae sp. nov. Syst Appl Microbiol
557   doi:10.1016/j.syapm.2018.01.002.
558   21.   Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA-DNA
559   hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol
560   Microbiol 57:81-91.
561   22.   Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG,
562   Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald
563   D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J,
564   Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community
565   sequencing data. Nat Methods 7:335-6.
566   23.   Bellemain E, Carlsen T, Brochmann C, Coissac E, Taberlet P, Kauserud H. 2010. ITS as an
567   environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. BMC
568   Microbiol 10:189.
569   24.   DiGiulio DB, Callahan BJ, McMurdie PJ, Costello EK, Lyell DJ, Robaczewska A, Sun CL, Goltsman DS,
570   Wong RJ, Shaw G, Stevenson DK, Holmes SP, Relman DA. 2015. Temporal and spatial variation of
571   the human microbiota during pregnancy. Proc Natl Acad Sci U S A 112:11060-5.
572   25.   Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, Karlebach S, Gorle R, Russell J,
573   Tacket CO, Brotman RM, Davis CC, Ault K, Peralta L, Forney LJ. 2011. Vaginal microbiome of
574   reproductive-age women. Proc Natl Acad Sci U S A 108 Suppl 1:4680-7.
575   26.   Svetoch EA, Eruslanov BV, Levchuk VP, Perelygin VV, Mitsevich EV, Mitsevich IP, Stepanshin J,
576   Dyatlov I, Seal BS, Stern NJ. 2011. Isolation of Lactobacillus salivarius 1077 (NRRL B-50053) and
577   characterization of its bacteriocin, including the antimicrobial activity spectrum. Appl Environ
578   Microbiol 77:2749-54.
579   27.   Kobierecka PA, Wyszynska AK, Aleksandrzak-Piekarczyk T, Kuczkowski M, Tuzimek A, Piotrowska W,
580   Gorecki A, Adamska I, Wieliczko A, Bardowski J, Jagusztyn-Krynicka EK. 2017. In vitro characteristics
581   of Lactobacillus spp. strains isolated from the chicken digestive tract and their role in the inhibition
582   of Campylobacter colonization. Microbiologyopen 6.
583   28.   Merhej V, Armougom F, Robert C, Raoult D. 2012. Genome sequence of Lactobacillus ingluviei, a
584   bacterium associated with weight gain in animals. J Bacteriol 194:5697.
585   29.   Dec M, Puchalski A, Urban-Chmiel R, Wernicki A. 2014. Screening of Lactobacillus strains of
586   domestic goose origin against bacterial poultry pathogens for use as probiotics. Poult Sci 93:2464-
587   72.
588   30.   Baele M, Devriese LA, Haesebrouck F. 2001. Lactobacillus agilis is an important component of the
589   pigeon crop flora. J Appl Microbiol 91:488-91.
590   31.   Fujisawa T, Shirasaka S, Watabe J, Mitsuoka T. 1983. Lactobacillus aviarius sp. nov.: A new species
591   isolated from the intestine of chickens. Systematic and Applied Microbiology 5:414-420.
592   32.   La Ragione RM, Narbad A, Gasson MJ, Woodward MJ. 2004. In vivo characterization of Lactobacillus
593   johnsonii FI9785 for use as a defined competitive exclusion agent against bacterial pathogens in
594   poultry. Lett Appl Microbiol 38:197-205.

595 33. Ventura M, Zink R. 2002. Specific identification and molecular typing analysis of Lactobacillus
596   johnsonii by using PCR-based methods and pulsed-field gel electrophoresis. FEMS Microbiol Lett
597   217:141-54.
598 34. Gala E, Landi S, Solieri L, Nocetti M, Pulvirenti A, Giudici P. 2008. Diversity of lactic acid bacteria
599   population in ripened Parmigiano Reggiano cheese. Int J Food Microbiol 125:347-51.
600 35. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome
601   Res 14:1188-90.
602 36. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics
603   26:2460-1.
604 37. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG,
605   Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald
606   D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Tumbaugh PJ, Walters WA, Widmann J,
607   Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community
608   sequencing data. Nature Methods 7:335-336.

609

610

611   **Table 1:** List of *Lactobacillus* species with LITSA sequence identity ≥97 % with another

612   *Lactobacillus* species. The percentage reported corresponds to the highest identity observed among

613   all LITSA sequences identified in strains of the two species compared.

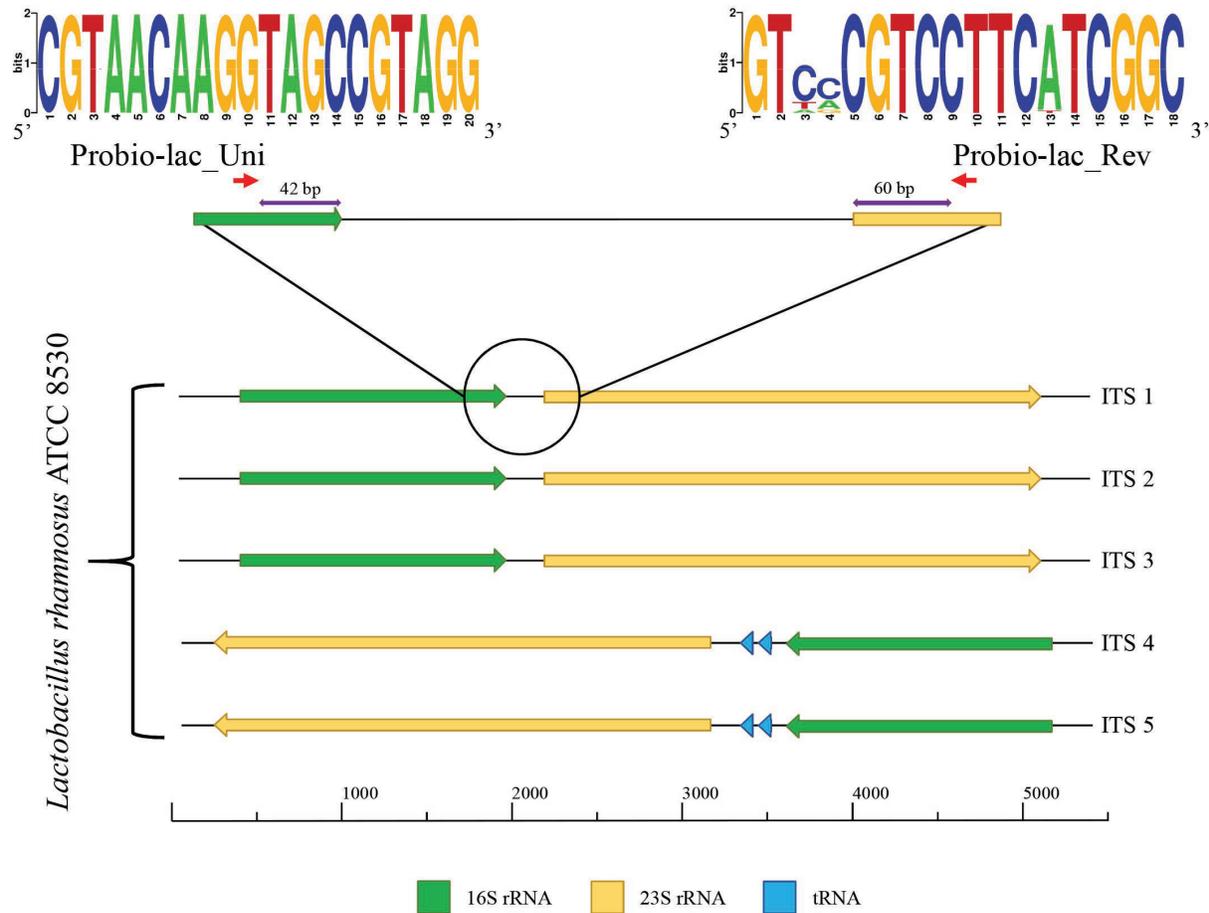| Species	LITSA % identity with the closest species	Closest species
Lactobacillus acidophilus	98	*Lactobacillus amylovorus*
	97	*Lactobacillus crispatus*
Lactobacillus amylovorus	98	*Lactobacillus acidophilus*
	97	*Lactobacillus crispatus*
Lactobacillus buchneri	99	*Lactobacillus parabuchneri*
Lactobacillus casei	99	*Lactobacillus paracasei*
	100	*Lactobacillus rhamnosus*
Lactobacillus crispatus	97	*Lactobacillus acidophilus*
	97	*Lactobacillus amylovorus*
Lactobacillus curvatus	98	*Lactobacillus sakei*
Lactobacillus gallinarum	99	*Lactobacillus helveticus*
Lactobacillus gasseri	99	*Lactobacillus johnsonii*
Lactobacillus helveticus	99	*Lactobacillus gallinarum*
Lactobacillus johnsonii	99	*Lactobacillus gasseri*
Lactobacillus parabuchneri	99	*Lactobacillus buchneri*
Lactobacillus paracasei	99	*Lactobacillus casei*
	100	*Lactobacillus rhamnosus*
Lactobacillus paraplantarum	100	*Lactobacillus pentosus*
	100	*Lactobacillus plantarum*
Lactobacillus pentosus	100	*Lactobacillus paraplantarum*
	99	*Lactobacillus plantarum*
Lactobacillus plantarum	100	*Lactobacillus paraplantarum*
	99	*Lactobacillus pentosus*
Lactobacillus rhamnosus	100	*Lactobacillus paracasei*
	100	*Lactobacillus casei*
Lactobacillus sakei	98	*Lactobacillus curvatus*

614

25

615   **Figure legends**

616   **Figure 1:** Genetic map of the Internally Transcribed Sequence (ITS) region of *Lactobacillus* with

617   and without tRNA genes. Panel a depicts the genetic organization of the five complete ITS regions

618   predicted in the complete genome of *Lactobacillus rhamnosus* ATCC 8530, used here as a test case.

619   Primer sequence conservation is shown through a WebLogo representation where the overall height

620   of the stacks indicates the sequence conservation at that position, while the height of symbols within

621   the stacks indicates the relative frequency of nucleic acids at that position. Panel b illustrates the

622   details of ITS regions identified in the complete genomes of species included in the mock sample

623   for which a complete genome was available. ITS sequences without tRNA genes are highlighted in

624   green, while ITS regions encoding tRNA genes are marked in blue. Black and red text indicates

625   forward and reverse strand orientation, respectively.

626

627   **Figure 2:** Evaluation of the sensitivity and accuracy of the *Lactobacillus* ITS profiling protocol.

628   The graph shows the expected and observed relative abundance of 14 *Lactobacillus* taxa

629   constituting an artificial sample. An exponential trendline is reported for the expected and observed

630   data.

631

632   **Figure 3:** ITS and 16S rRNA gene profiling of *Lactobacillus* species in five ecological niches. The

633   profile of the *Lactobacillus* population obtained for: a) five human faecal samples (HG); b) five

634   human vaginal swab samples (HV); c) five free range chicken faecal samples (FRC); d) five whey

635   samples (WH), and e) five parmesan cheese samples (PC) is depicted in the corresponding bar

636   plots. Only species with relative abundance >5% in at least a sample are reported. Species below

637   5% are collapsed in "Others <5 %.

26

Figure 1

Figure 2

Figure 3