<table>
<thead>
<tr>
<th>Title</th>
<th>Phylotype-level profiling of lactobacilli in highly complex environments by means of an ITS-based metagenomic approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Milani, Christian; Duranti, Sabrina; Mangifesta, Marta; Lugli, Gabriele Andrea; Turroni, Francesca; Mancabelli, Leonardo; Viappiani, Alice; Anzalone, Rosaria; Alessandro, Giulia; Ossiprandi, Maria Cristina; van Sinderen, Douwe; Ventura, Marco</td>
</tr>
<tr>
<td>Publication date</td>
<td>2018-05-04</td>
</tr>
<tr>
<td>Type of publication</td>
<td>Article (peer-reviewed)</td>
</tr>
<tr>
<td>Link to publisher's version</td>
<td><a href="http://aem.asm.org/content/early/2018/04/30/AEM.00706-18.abstract">http://aem.asm.org/content/early/2018/04/30/AEM.00706-18.abstract</a> <a href="http://dx.doi.org/10.1128/aem.00706-18">http://dx.doi.org/10.1128/aem.00706-18</a></td>
</tr>
<tr>
<td>Rights</td>
<td>© 2018 American Society for Microbiology.</td>
</tr>
<tr>
<td>Embargo information</td>
<td>Access to this article is restricted until 6 months after publication by request of the publisher.</td>
</tr>
<tr>
<td>Embargo lift date</td>
<td>2018-11-04</td>
</tr>
<tr>
<td>Item downloaded from</td>
<td><a href="http://hdl.handle.net/10468/6181">http://hdl.handle.net/10468/6181</a></td>
</tr>
</tbody>
</table>

Downloaded on 2018-12-26T20:06:43Z
Phylotype-level profiling of lactobacilli in highly complex environments by means of an ITS-based metagenomic approach.

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

Christian Milani\(^1\)*, Sabrina Duranti\(^1\)*, Marta Mangifesta\(^1\)*, Gabriele Andrea Lugli\(^1\), Francesca Turroni\(^{1,2}\), Leonardo Mancabelli\(^1\), Alice Viappiani\(^3\), Rosaria Anzalone\(^1\), Giulia Alessandri\(^1\), Maria Cristina Ossiprandi\(^4\), Douwe van Sinderen\(^5\) and Marco Ventura\(^{1,2}\)

*These authors contributed equally.

Laboratory of Probiogenomics, Department of Life Sciences, University of Parma, Italy\(^1\);
Microbiome Research Hub, University of Parma, Parma, Italy\(^2\) GenProbio srl, Parma, Italy\(^3\);
Department of Veterinary Medical Science, University of Parma, Parma, Italy\(^4\); APC Microbiome Ireland and School of Microbiology, Biosciences building, University College Cork, Ireland\(^5\)

Corresponding author. Mailing address for Marco Ventura Department of Life Sciences, University of Parma, Parco Area delle Scienze 11a, 43124 Parma, Italy. Phone: +39-521-905666. Fax: +39-521-905604. 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.
Importance

The genus *Lactobacillus* is a large and ubiquitous taxon of high scientific and commercial relevance. Despite the fact that the genetic repertoire of lactobacilli species has been extensively characterized, the ecology of this genus has been explored by metataxonomic techniques that are accurate down to the genus or phylogenetic group level only. Thus, the distribution of lactobacilli in environmental or processed food samples is relatively unexplored. The profiling protocol described here relies on the use of the Internally Transcribed Spacer to perform an accurate classification in a target population of lactobacilli with <0.006 ng/µl sensitivity. This approach was used to analyze five sample types collected from both human and animal host-associated microbiota as well as from the cheese production chain. Availability of a tool for species-level profiling of lactobacilli may be highly useful for both academic research and a wide range of industrial applications.
Introduction

The genus *Lactobacillus* is a widespread and diverse taxon encompassing more than 170 species and 17 subspecies, which are classified as Gram-positive, non-spore-forming and catalase-negative facultative anaerobes (1, 2). Moreover, based on their metabolic capability to produce lactic acid as the main metabolic end product of carbohydrate fermentation, lactobacilli are classified as members of the Lactic Acid Bacteria (LAB). Notably, 16S rRNA gene-based phylogenetic analyses revealed the existence of 22 distinct phylogenetic groups of *Lactobacillus* species (24 including pediococci) (2-4).

Regarding their ecological distribution, lactobacilli are found in a wide range of environments, including plants, water, soil, silage and different body sites of humans and other animals as members of host-associated microbiomes, such as those colonizing the oral cavity, the vagina and the gastrointestinal tract (GIT) (4, 5). Moreover, 37 species of this genus have been granted the Qualified Presumption of Safety (QPS) status by the European Food Safety Authority (EFSA) (6). Thus, they are extensively used in the food industry, in particular in fermented foods due to their high performance in lactic acid fermentation coupled with high tolerance for low pH, preservative and organoleptic properties, and production of exopolysaccharides that contribute to the texture of foods (2, 7). In this context, members of the genus *Lactobacillus* have in recent years gained significant scientific and commercial interest as health-promoting microorganisms, as evidenced by the fact that 22 species encompass strains patented as probiotics in Europe (8).

The high commercial and scientific relevance of lactobacilli coupled to the recent introduction of next-generation sequencing technologies has recently led to genome decoding of all (then) known *Lactobacillus* species (3, 7). The retrieved genomic data has been exploited for comparative genomic analyses, and has allowed identification of many shared or distinct genetic features of this genus. Furthermore, this genomic information has permitted the reconstruction of their metabolic potential, has shed light on host-microbe interactions, such as adhesion to the mucus...
layer and modulation of the immune system of the host, and has revealed particular microbe-
microbe interactions with other commensals or (opportunistic) pathogens (1, 7, 8).

Despite the large body of data concerning the physiology and genetics of lactobacilli,
knowledge about the ecology and distribution in environmental or host-associated niches of
individual species relies mainly on culture-dependent studies. This is partly due to the resolution
limit of currently used metagenomic approaches. Although microbial profiling based on partial 16S
rRNA gene is able to discriminate between phylogenetic groups of lactobacilli due to the high
phylogenetic diversity of this genus, it cannot provide an accurate species-level resolution.
Moreover, the majority of the current exiting studies of lactobacilli populations based on 16S rRNA
gene profiling do not even perform phylogenetic group-level analyses. To offer a more refined
taxonomic view of lactobacilli in a given environment or sample, we developed a profiling
approach based on amplification of the internally transcribed spacer (ITS) sequence. Notably, due
to their high variability, ITS sequences have previously been exploited in a wide range of studies
encompassing the identification of unique species-specific restriction patterns of lactobacilli, as well
as the identification and characterization of Leuconostoc strains and for the genotyping of
Streptococcus pneumoniae strains (9-11). The developed methodology in the current work is able to
determine the composition of lactobacilli-containing communities down to the species level. The
method was validated through the analysis of a sample artificially constituted by DNA of 14
lactobacilli taxa at known concentration. Furthermore, we applied this methodology for the precise
investigation of bacterial communities harbored by human-, animal- and food-associated matrices
that were previously explored down to the genus-level only.
Results and discussion

Analysis of ITS variability within the *Lactobacillus* genus.

Genomes of 1523 strains assigned to the genus *Lactobacillus* and corresponding to 176 species were retrieved from the NCBI genome database, and then processed using the MEGAnnotator software (12) for prediction of rRNA genes in order to ensure the same high-quality standard for all sequences of ribosomal loci included in this study (Table S1). Notably, the genomic sequences of 892 *Lactobacillus* strains, representing the 58.6 % of the total strain pool analyzed, did not harbor complete rRNA loci, i.e. encompassing complete 5S, 16S and 23S rRNA genes. In contrast, at least one complete ribosomal rRNA genes locus was identified for 631 of the 1523 analyzed strains, corresponding to 70 species and a custom script was then used to extract a total of 1788 Internally Transcribed Spacer (ITS) sequences. Assembly of draft genomes generally generates the collapse of reads that correspond to rRNA genes into a single rRNA locus. However, availability of multiple draft sequences of a given *Lactobacillus* taxon, complemented with analysis of 217 complete genomes of *Lactobacillus* species, allowed us to retrieve an average of 25.5 ITS sequences per species. Interestingly, 137 of the 1788 retrieved ITS sequences include stretches of >3 undefined (N) nucleotides, thus highlighting a high rate of assembly-related issues and/or low-quality regions in genomes deposited at the NCBI genome database. Comparative analysis of the 1651 complete ITS sequences without multiple contiguous nucleotide ambiguities revealed that 92.5 % of the ITS sequences range between 200 and 500 bp.

As previously observed for bifidobacteria (13, 14), alignment of ITS sequences from *Lactobacillus* genomes shows a high level of diversity, probably due to a high mutation frequency, and corresponding to a high evolutionary rate, as reflected by multiple substitutions at a given nucleotide position and indicative of mutational saturation of such ITS sequences. While this particularly high mutation frequency prevents phylogeny inference (15), it is suitable for metagenomic amplicon-based profiling below the genus level, as previously validated for members of the genus *Bifidobacterium* (13).
Design of a PCR primer pair for ITS-profiling of the *Lactobacillus* genus.

Many profiling approaches have been developed to accurately reconstruct the taxonomic composition of complex bacterial communities. These include methods based on low-coverage sequencing of full-length 16S rRNA genes and the use of technologies providing long reads, i.e. Sanger and PacBio. Nevertheless, despite the fact that full-length sequencing of the 16S rRNA gene allows high accuracy in taxonomic assignment, the low sequencing coverage permits the detection only of dominant taxa and prevents profiling of bacteria present at low relative abundance in a given population (10). Furthermore, the use of alternative marker genes has also been proposed, though their use remains limited due to difficulties in the definition of universal primers as well as in the lack of a complete reference database. The advent of next-generation sequencing, characterized by high coverage and short reads, facilitated the amplification and sequencing of partial 16S rRNA genes, i.e. 16S rRNA gene profiling. This metagenomic method has in recent years been used as the gold standard for taxonomic characterization of environmental and host-associated microbiomes. While this methodology covers all bacterial biodiversity, it is generally only accurate for the reconstruction of taxonomic profiles down to genus level (16) or down to phylogenetic groups in case of genera with a high level of phylogenetic diversity, e.g. the genus *Lactobacillus* (3, 4) since it relies on sequencing of a small region of the whole 16S rRNA gene through next-generation sequencing. To overcome this limitation and to obtain species-level resolution, the use of the ITS sequence as an alternative molecular marker has been proposed (13).

In order to develop a universal primer pair suitable for profiling of all members of the *Lactobacillus* genus, we aligned the 16S and 23S rRNA genes flanking the 1651 complete ITS sequences without stretches of undefined nucleotides that were retrieved from lactobacilli genomes deposited at the NCBI database. Manual inspection of the alignments allowed the identification of ‘universal’ primers located at the 5’-end of the 16S rRNA gene and at the 3’-end of the 23S rRNA gene, i.e., Probio-lac_Uni (CGTAACAAGGTAGCCGTAGG) and Probio-lac_Rev...
(GTYVCGTCCTTCWTCGSC), respectively (Figure 1). Sequence conservation amongst the aligned 16S and 23S rRNA genes is reported in Figure 1 through a WebLogo representation. These primers generate an amplicon of an average length of 380 bp covering the complete ITS region and suitable for 2 X 250 bp paired-end Illumina sequencing followed by single-end bioinformatic analysis of both paired reads (see below). Analysis of single-end reads provided reliable assignment to species level even in cases where a tRNA gene was located within the ITS region (see below). Notably, the final sequence of the primers was defined after multiple iterative alignments to the Silva SSU and LSU databases (17) using the Silva TestProbe v. 3.0 tool (https://www.arb-silva.de/search/testprobe/). The latter approach led to introduction of specific IUPAC bases in order to maximize alignment of the primers to all currently available 16S and 23S rRNA gene sequences of lactobacilli corresponding to all known species of this genus, while minimizing alignment to non-lactobacilli ribosomal RNA genes. The usefulness of the Probio-lac_Uni/Probio-lac_Rev primer pair was in vitro validated through successful amplicon generation in the case of 31 lactobacilli species belonging to the 23 phylogenetic groups identified previously in the genus Lactobacillus (3, 4) (Figure S1). In contrast, no amplification was observed when the Probio-lac_Uni/Probio-lac_Rev primer pair was used to amplify DNA extracted from nine non-Lactobacillus taxa (Figure S1). Interestingly, for all tested lactobacilli we observed two PCR fragments, each with a molecular size ranging from 300 to 350 bp, and 500 to 550 bp, corresponding to the ITS region with and without a tRNA gene (see below for details), respectively (Figure S1) (Figure 1). Such ITS patterns confirmed those displayed in previous studies targeting the amplification of the ITS region of lactobacilli (18). Notably, for few taxa we observed a faint amplification fragment of 500 to 550 bp, which might suggest a lower copy number of ITS regions encompassing tRNA genes in the same genome.

The Probio-lac_Uni/Probio-lac_Rev primer pair was employed for in silico PCR amplification of the 631 genomes of the genus Lactobacillus encoding at least one rRNA genes locus. This approach facilitated the development of a database encompassing 1651 complete ITS sequences without...
multiple ambiguous nucleotides, and flanked by partial 16S and 23S rRNA sequences, together constituting the **Lactobacillus ITS** Amplicon database (LITSA database).

Cross-alignment of all retrieved LITSA sequences using MatGAT software (19) was performed in order to evaluate the level of identity between predicted amplicons (Table S2) and to evaluate possible limits imposed by actual lactobacilli taxonomy to the proposed ITS profiling methodology. Notably, this analysis highlighted cases in which comparison of multiple LITSA sequences from the same strain showed low identity. In-depth investigation revealed that 46 of the 62 lactobacilli species included in the LITSA database contain at least one ITS sequence that harbors two tRNA genes (for Alanine and Isoleucine) (Figure 1). Notably, despite the fact that this prediction is limited due to the small number of complete genomes available, the presence of tRNA genes in one or multiple rRNA loci appears to be a common feature of genomes from members of the **Lactobacillus** genus.

Furthermore, cross-alignment analysis also revealed that the majority of the 62 *Lactobacillus* species, for which a complete LITSA sequence was available, can be discriminated (Table S2), with the exception of putatively misclassified strains and/or species (see below). In this context, despite the fact that lactobacilli are known to possess a very high level of phylogenetic diversity (3, 4), strains corresponding to 18 species showed an average LITSA sequence identity of >97% with at least one other *Lactobacillus* species, thus showing a very close phylogenetic relationship between such taxa (Table 1). Amongst lactobacilli, *Lactobacillus casei* and *Lactobacillus paracasei* strains possess an average LITSA sequence identity of 99%, while the amplicon sequences of *Lactobacillus pentosus*, *Lactobacillus plantarum* and *Lactobacillus paraplantarum* strains show up to 100% identity (Table S2). An in-depth analysis of each strain revealed that 23 of the 25 strains classified as *L. casei* share an average LITSA sequence identity ≤96.1% with the type strain *L. casei* ATCC 393, while the average identity with the type strain *L. paracasei* ATCC 394 is ≥99.7% (Table S2). In contrast, the putative lactobacilli species *Lactobacillus* sp. FMNP02 shares 99.7%
identity with *L. casei* ATCC 393 (Table S2), thus representing a possible misclassification of the latter strain.

In our attempts to obtain insights into the phylogeny of *L. pentosus*, *L. plantarum* and *L. paraplantarum*, we observed an average LITSA sequence identity of 98.9 % between *L. pentosus* and *L. plantarum* strains (Table S2). Moreover, the two strains of *L. paraplantarum*, for which we were able to predict an rRNA gene locus, show an average LITSA identity of 99.5 % with *L. plantarum* strains (Table S2), thus indicating that such taxa may belong to the same species and therefore cannot be discriminated using metataxonomic techniques. Nevertheless, evaluation of the average nucleotide identity is needed to confirm this hypothesis. Furthermore, we could not retrieve an *in silico* Probio-lac_Uni/Probio-lac_Rev-corresponding amplicon for the type strains of *L. pentosus* and *L. paraplantarum* due to absence of a complete ITS region in the deposited genomes, and we were therefore unable to evaluate their amplicon identity with the LITSA sequences of *L. plantarum* strains.

Notably, these observations suggest that major issues in the classification of the genus *Lactobacillus* still exist, resulting in the unfeasibility of distinguish a number of species through ITS profiling. Thus, as has been proposed previously, it is desirable that a re-evaluation of the taxonomy of lactobacilli is undertaken based on a phylogenomic approach (20, 21), as was also corroborated by recent studies (3, 4, 7).

**Development of a bioinformatic tool for ITS-profiling of the *Lactobacillus* genus.** The length of the amplicon produced by the Probio-lac_Uni/Probio-lac_Rev primer pair may exceed 600 bp, particularly when tRNA-encoding sequences are present in the ITS sequence. Thus, sequencing produced non-overlapping paired-end reads even with the maximum length obtainable using Next-Generation Illumina sequencing, i.e. 2 X 250 bp paired-end reads, using the MiSeq Reagents Kit v3 600 cycles chemistry. Nevertheless, each forward and reverse read covers 42 and 60 nucleotides corresponding to the 16S rRNA gene 3’-end and the 23S rRNA gene 5’-end,
respectively, which are followed by 190-208 bp of hyper-variable ITS sequence suitable for profiling at species-level (Figure 1). Thus, we developed a package for QIIME software suite v1.9.1 (22) that encompasses the LITSA database and a bash script for analysis of both forward and reverse reads of the Lactobacillus ITS profiling data (probiogenomics.unipr.it/pbi). Notably, the LITSA database will be updated regularly to include additional ITS sequences as new lactobacilli genome sequences become available, thus increasing the number of lactobacilli species that can be profiled. The script performs quality-filtering, de novo OTU clustering at 100% identity and taxonomic classification of OTU reference sequences through RDP classifier with a confidence level of 0.80. Notably, these cut-off values permit discrimination of closely related taxa. Due to the average size of the amplicon, the paired-end reads are not joined prior classification. Instead, the script analyzes both the forward and the reverse reads altogether and provides an average profile.

Notably, the different number of rRNA loci predicted in the genomes of Lactobacillus species may generate biases in the retrieved profiles. Thus, we evaluated the average number of ITS regions present in the 217 available complete Lactobacillus genomes. This analysis provided data for normalization of 45 of the 62 species of lactobacilli for which a LITSA sequence could be retrieved. Moreover, the average number of rRNA genes loci of the remaining 17 species with only draft genomes was set at 5.6, i.e., the average obtained for all the species with at least a complete genome. Notably, the Lactobacillus ITS profiling analysis script includes a normalization step based on the number of rRNA genes loci predicted for all the 62 Lactobacillus species for which a LITSA sequence could be retrieved. The output produced by the script is summarized in the “output” folder, which contains the predicted taxonomic profile based on the LITSA database (both non-normalized and normalized for the number of rRNA loci) and the OTU table in tabular text format that reports the reference sequence and associated taxonomy. All Lactobacillus ITS profiles reported in this manuscript correspond to the average between forward and reverse read profiles after normalization for the number of predicted rRNA genes loci.
Assessing detection sensitivity and accuracy using the *Lactobacillus* ITS profiling protocol

In order to provide an evaluation of the sensitivity and accuracy of the Probio-lac_Uni/Probio-
lac_Rev primer pair, 14 *Lactobacillus* type strains were employed to artificially compose a mock
community (Table S3). The DNA extracted from each taxon grown in pure culture was added to the
mix at known amount, ranging from 0.006 ng to 50 ng of DNA, corresponding to 0.006 % to 50 %
of the total DNA pool (Figure 2). Sequencing of the mock sample was performed using an Illumina
MiSeq with 2X250 bp chemistry, producing 45,146 quality-filtered paired-end reads. Interestingly,
*Lactobacillus* ITS profiling of this dataset successfully profiled all *Lactobacillus* species included in
this sample, except *Lactobacillus vaginalis* and *Lactobacillus pontis*, for which we could not
retrieve a LITSA sequence from analysis of available genome sequences (Figure 2). Thus, even
though the Probio-lac_Uni/Probio-lac_Rev primer pair produces an amplicon for these species, the
latter cannot be taxonomically classified due to absence of *L. vaginalis* and *L. pontis* in the present
version of the LITSA database. This is a temporary limitation and the LITSA database will be
updated regularly (probiogenomics.unipr.it/pbi) to include LITSA sequences of newly sequenced
genomes in order to cover all the lactobacilli species that currently cannot be profiled. Moreover,
comparison of the retrieved profile with the expected composition revealed a strong correlation for
each taxon with few discrepancies (Figure 2). The causes of such differences between expected and
observed relative abundance may be imputed to the lack of sufficient information in the LITSA
database, at this time, regarding the average number of rRNA loci per genome used for
normalization of the ITS profiling data.

Furthermore, since PCR amplicon size has been identified as a source of bias in ITS-based profiling
studies of fungi (23), we evaluated the presence of possible biases introduced by amplification of
lactobacilli ITS sequences of different length due to the presence or absence of tRNA genes (see
above). The 14 *Lactobacillus* species that constitute the mock community (Table S3) were
subjected to manual characterization of corresponding rRNA loci. Notably, the ten species for
which a complete genome was available, confirmed what had been observed for the *in vitro* PCR,
i.e. presence of longer ITS sequences that encompass two tRNA genes (Figure 1; Figure S1).

Interestingly, the different intensities observed in the PCR fragments, i.e. 300-350 bp and 500-550 bp (Figure S1), did not influence expected relative abundance of the mock community (Figure 2). Notably, detection of *Lactobacillus rhamnosus* whose concentration in the mock community is 0.006 ng/µl indicates that the limit of detection of the lactobacilli ITS profiling is <0.006 ng/µl, corresponding to 1.85*10^3 cells/µl.

**Validation of the *Lactobacillus* ITS profiling protocol through analysis of samples from multiple environments.**

*Lactobacillus* is a highly diverse microbial genus, members of which are found in a wide range of environments (5). To perform a comprehensive testing of the performances of the *Lactobacillus* ITS profiling protocol, we analyzed a total of 25 samples encompassing five human faecal samples, five human vaginal swab samples, five free range chicken cecal samples, five whey samples and five parmesan cheese samples (Table S4). Sequencing was performed with an Illumina MiSeq instrument using 2x250 bp chemistry, producing an average of 15,529 forward and 15,293 reverse quality-filtered reads per sample (Table S4).

Interestingly, analysis of the human faecal samples revealed the presence of human gut colonizers, such as *Lactobacillus rhamnosus*, along with a range of lactobacilli used in functional or fermented foods that are typically part of the human diet, such as *L. plantarum*, *Lactobacillus helveticus*, *Lactobacillus delbrueckii* and *Lactobacillus sakei* (Figure 3).

Moreover, the obtained profiles of the five human vaginal swab samples confirmed the proposed existence of community state types (CSTs) of the vaginal microbiota dominated by specific *Lactobacillus* taxa (24, 25). In fact, HV1 is dominated by *Lactobacillus gasseri*, while *Lactobacillus iners* and *Lactobacillus crispatus* are the most abundant lactobacilli taxa in HV2/HV5 and HV3/HV4, respectively (Figure 3). Furthermore, in all five reconstructed human vaginal profiles, *L. helveticus* is the second most abundant *Lactobacillus* species, as observed in the
aforementioned CSTs (24, 25) (Figure 3). Thus, based on the classification proposed by DiGiulio et
al. (24), HV1 can be classified as a CST 2, while HV2/HV5 falls within the CST 3, whereas
HV3/HV4 can be attributed to CST 1.

To demonstrate the relevance of an efficient methodology for precise cataloguing of the
*Lactobacillus* species for which a complete LITSA sequence is available in different environments,
we analyzed five free range chickens cecal samples. The retrieved profiles revealed a high relative
abundance (ranging from a total of 53.1 % to 96.8 %) of *Lactobacillus* species previously
characterized in poultry or other birds, such as *Lactobacillus salivarius*, *Lactobacillus reuteri*,
*Lactobacillus ingluviei*, *Lactobacillus amylovorus*, *Lactobacillus agilis*, *Lactobacillus aviaris* and
*Lactobacillus johnsonii* (26-33) (Figure 3). Notably, samples FRC1, FRC2 and FRC3 showed a
similar profile with high abundance of *L. salivarius*, *L. ingluviei* and *L. amylovorus*, reflecting the
fact that they were kept in the same hen house (Figure 3). Accordingly, samples FRC4 and FRC5,
collected in two additional hen houses, showed different profiles characterized by high abundance
of *L. aviaris* and *L. johnsonii*, respectively (Figure 3).

For milk and milk-related products, profiling of five whey and five fresh parmesan cheeses (at 1
day of ripening) samples revealed, as expected, similar profiles dominated by *L. helveticus* and *L.
delbrueckii* (Figure 3), which represent two lactobacilli species typically used as starter cultures for
the production of cheese (34). These data indicate that the *Lactobacillus* ITS profiling approach also
represents a valuable tool for monitoring the population of lactobacilli across the cheese production
chain.

Results obtained from ITS-profiling were also compared to profiles reconstructed through analysis
of OTUs generated at 99 % identity from 16S rRNA profiling data (Figure 3) (Table S5). Notably,
only OTUs classified as lactobacilli have been included in the representation, thus the relative
abundance of unclassified lactobacilli reported in the bar plot do not include additional OTUs that
could not be attributed to this genus. Moreover, lactobacilli species whose relative abundance is
below 5 % in each sample were collapsed under “Others <5 %” in the bar plot representation.
Interestingly, the ITS profiling approach provided a more accurate species-level reconstruction of the lactobacilli populations when used to analyze human faecal and vaginal samples as well as free range chicken faecal samples. Moreover, it confirmed and partially improved the simple lactobacilli community of whey and fresh Parmesan cheese samples observed through 16S rRNA gene profiling. In fact, differences in the profiles obtained through 16S rRNA gene and ITS profiling can be observed in all cases (Figure 3) (Table S5). Such differences are caused by the limited number of Lactobacillus species that could be discriminated based on partial 16S rRNA gene sequence respect to ITS sequence (Figure 3) (Table S5).

Altogether, these results confirm the performance of the Lactobacillus ITS profiling protocol observed from analysis of the artificial sample and validate their use, complementary to 16S rRNA gene profiling, for analysis of a wide range of complex environmental and host-associated matrices.

**Conclusions**

We developed a newly designed method for characterization of the Lactobacillus population in complex environments based on the use of the internally transcribed spacer (ITS), which represents a hypervariable region located between the 16S and the 23S rRNA genes that allows high-accuracy species-level profiling. The accuracy and sensitivity of this method allowed profiling of complex communities of lactobacilli with a successful identification of taxa with abundance of $1.85 \times 10^3$ cells/µl, which is even lower to what was previously identified for a similar approach developed for the profiling of bifidobacterial communities (13). Notably, despite the fact that the current LITSA database allows the precise profiling of just 62 species, the ITS-profiling approach represents a new metagenomic tool for species-level profiling of complex lactobacilli communities that complements phylogenetic group assignments that can be obtained from 16S rRNA gene profiling data. Moreover, the database will be regularly updated to represent additional lactobacilli species as genomes encompassing complete LITSA sequences are becoming available. When the ITS lactobacilli profiling method was applied to different biological samples, encompassing the stool of...
human as well as birds, vaginal swabs and cheese, it allowed the reconstruction of the cataloguing of lactobacilli communities residing in these environments. Altogether, these results highlight that ITS-mediated profiling of populations of lactobacilli could be useful not only for academic purposes, but also for industrial applications such as tracing the microbial composition of probiotic products based on lactobacilli as well as of starter cultures in food manufacture.

Material and methods

Sample collection

In the frame work of a more extensive bacterial cataloguing project, this study enrolled stool, vaginal swab, fresh parmesan cheese (one day of ripening), whey and cecal (from free range chickens) samples. Five fresh stool samples obtained from human healthy volunteers and five cecal samples retrieved from free range chickens were immediately frozen upon collection at -80°C until processing for DNA extraction. The DNA extraction was performed using the QIAamp DNA Stool Mini Kit following the manufacturer’s instructions (Qiagen, Manchester, UK). Additionally, five vaginal swab samples were collected in sterile tubes containing 1 ml of DNA-RNA shield from ZYMO Research until bacterial DNA extraction using ZymoBIOMICS™ DNA Miniprep Kit (ZYMO Research). Furthermore, 10 ml samples of whey and 2-4 gr of fresh parmesan cheese were collected in sterile tubes and the DNA was extracted using the DNeasy Mastitis Mini Kit (Qiagen Ltd, Strasse, Germany) following the manufacturer’s instructions (Qiagen Ltd). Notably, whey samples and cheese samples at one day of ripening were collected from the same Parmesan cheese producer in Parma, Italy.

Ethical statement. This study was carried out in accordance with the recommendations of the ethical committee of the University of Parma and was approved by the “Comitato di Etica Università degli Studi di Parma”, Italy. All animal procedures were performed according to national guidelines (Decreto legislativo 26/2014).
Bacterial growth conditions and DNA extraction. Type strains of several lactobacilli taxa (Table S3) were grown in Man-Rogosa-Sharpe (MRS) medium (Scharlau Chemie) supplemented with 0.05 % (w/v) L-cysteine hydrochloride and incubated in an anaerobic atmosphere (2.99 % H2, 17.01 % CO2 and 80 % N2) in a chamber (Concept 400; Ruskin) at 37°C for 24 h. In addition, nine non-lactobacilli microorganisms were used in this study. These included Bifidobacterium bifidum LMG11041, which was cultivated in MRS broth as Lactobacillus strains; Collinsella intestinalis DSM 13280, Escherichia coli LMG 2092, and Klebsiella pneumoniae CECT 143, which were grown in de MRS broth (Difco, Detroit, MI) supplemented with 0.05% (w/v) l-cysteine (MRSC; Sigma, St. Louis, MO). Prevotella copri DSM 18205 and Blautia coccoides DSM 935 were cultivated in a combination of Reinforced Clostridial Broth (Merck, Darmstadt, Germany) and Brain-Heart Infusion (Difco), supplemented with 5% (v/v) heat-inactivated fetal bovine serum (LabClinics, Barcelona, Spain) respectively. For Bacteroides thetaiotaomicron DSMZ 2079, the latter medium was supplemented with 0.005 % hemin (Sigma) and 0.005 % Vitamin K1 (Sigma). Faecalibacterium prausnitzii DSM 17677 was grown in Wilkins-Chalgren Anaerobe broth (Merck), following the recommendations included in the DSMZ medium 339. Finally, an active culture of Methanobrevibacter smithii DSM 861 grown in Methanobacterium medium (DSMZ 119) was directly supplied by DSMZ.

Bacterial DNA was extracted using GenEluteTM Bacterial Genomic DNA kits (SIGMA-ALDRICH) following the manufacturer’s instructions. Taxonomic identity of the microorganisms was validated by sequencing the V3 variable region of the 16S rRNA gene using primer pair Probio_Uni and Probio_Rev (14).

Lactobacillus mock community

The cultures of fourteen different Lactobacillus strains were grown separately on Man-Rogosa-Sharpe (MRS) medium (Scharlau Chemie) supplemented with 0.05 % (w/v) L-cysteine hydrochloride and incubated in an anaerobic atmosphere (2.99 % H2, 17.01 % CO2 and 80 % N2) in a chamber (Concept 400; Ruskin) at 37°C until they reached late log phase. The bacteria were
enumerated by counting colonies on solid medium and the optical density at 600 nm was determined. The final bacterial cell concentration was approximately $10^7$ cfu/ml. Chromosomal DNA of each strain was extracted as previously described and subsequently mixed. Specifically, the mock community consists of a pool of known concentration of fourteen different *Lactobacillus* strains to obtain the final quantity of DNA indicated in Table S3. Furthermore, the mix was prepared by combining equal volumes (20 μL) of DNA.

The DNA from the mix was diluted to produce a final DNA concentration of 2 ng/μL, and 4 μL of these dilutions were used in each PCR reaction. For the PCR reaction, the primer pair Probio-lac_Uni/Probio-lac_Rev was used and the generated amplicons were sequenced using Illumina MiSeq as described below.

### Lactobacillus ITS-specific primer design and gene amplification

The bioinformatics platforms MEGAnnotator (10) and METAnnotatorX (unpublished data) were used to perform 16S and 23S rRNA genes prediction in all 1523 sequenced lactobacilli genomes deposited at the NCBI Genomes database. Primers Probio-lac_Uni (CGTAACAAGGTAGCCGTAGG)/Probio-lac_Rev (GTYVCGTCTCTCWTCGSC) were manually designed based on the alignment of all 16S and 23S rRNA sequences to generate an amplicon encompassing the 3'-end of the 16S rRNA gene, the ITS region and the 5'-end of the 23S rRNA gene. Specificity test was performed using the Silva TestProbe v. 3.0 tool (https://www.arb-silva.de/search/testprobe/) that allows alignment of primers sequences to the Silva SSU and LSU databases (15). A custom bioinformatics script was then used to create a database of all the Probio-lac_Uni/Probio-lac_Rev-generated lactobacilli ITS amplicon sequences (LITSA database), encompassing the Internally Transcribed Spacer (ITS) region and partial 16S and 23S rRNA genes. The PCR conditions used for *Lactobacillus* ITS profiling using the Probio-lac_Uni/Probio-lac_Rev 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, followed by 10 min at 72 °C. Amplification was carried out using a Verity Thermocycler (Applied Biosystems). The integrity of the PCR amplicons was analyzed by gel electrophoresis. An
additional specificity test was performed by PCR using the DNA extracted from all known
*Lactobacillus* species as well as *B. bifidum* ATCC11041, *C. intestinalis* DSM 13280, *E. coli* LMG
*thetaiotaomicron* DSMZ 2079, *F. prausnitzii* DSM 17677 and *M. smithii* DSM 861.
WebLogo representation of primer sequence conservation among the retrieved 16S and 23S rRNA
genes flanking complete ITS sequences was obtained through the WebLogo website
(http://weblogo.berkeley.edu/) (35).

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

Illumina adapter overhang nucleotide sequence was added to the PCR amplicons obtained following
amplification of the ITS region, as previously described (13). The library of ITS amplicons was
prepared following the 16S Metagenomic Sequencing Library Preparation Protocol (Part No.
1504423 Rev. B-Illumina). Sequencing was performed using an Illumina MiSeq sequencer with
MiSeq Reagent Kit v3 chemicals.

**ITS-based microbiota analysis**

Fastq files obtained from metagenomic sequencing of each sample were analyzed using a custom
script for QIIME software suite (22) and the LITSA database available at
(http://probiogenomics.unipr.it/pbi/index.html).
Input data were processed in the following steps: filtering of the reads based on length > 100 nt
(primers included) to avoid primer dimers, overall quality > 25 and the presence of forward and
reverse primers in the forward and reverse reads, respectively, creation of OTUs constituted by
identical sequences using prefix_suffix method and removal of OTUs represented by < 10
sequences. Taxonomy assignment was performed using RDP method (RDP classifier with a
confidence level of 0.80) and the LITSA database constituted by ITS sequences retrieved from the
1523 *Lactobacillus* genomes available at the NCBI Genome database. This script is easily
modifiable to obtain a profiling based on a different sequence, though will depend on the
availability of a corresponding database.
**Evaluation of the sensitivity of the Probio-lac_Uni/Probio-lac_Rev primer pair**

The artificial sample used for the evaluation of the detection sensitivity and accuracy of the Probio-lac_Uni/Probio-lac_Rev primer set was generated using known DNA amounts, ranging from 50 to 0.006 ng, of 14 different *Lactobacillus* taxa (Table S3).

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

Partial 16S rRNA gene sequences were amplified from extracted DNA using primer pair Probio_Uni / Probio_Rev, which target the V3 region of the 16S rRNA gene sequence (16). 16S rRNA gene amplification and amplicon checks were carried out as previously described (16). 16S rRNA gene sequencing was performed using a MiSeq (Illumina) at the DNA sequencing facility of GenProbio srl (www.genprobio.com) according to a previously reported protocol (16). Following sequencing, the .fastq files were processed using a custom script based on the QIIME software suite (22). Paired-end read pairs were assembled to reconstruct complete Probio_Uni / Probio_Rev amplicons. Quality control retained sequences with a length between 140 and 400 bp and mean sequence quality score >20 while sequences with homopolymers >7 bp and mismatched primers were omitted. 16S rRNA gene Operational Taxonomic Units (OTUs) were defined at ≥ 99 % sequence homology using uclust (36) and OTUs with less than 10 sequences were filtered. All reads were classified to the lowest possible taxonomic rank using QIIME (37) and a reference dataset from the SILVA database (Quast et al., 2013).

**Nucleotide sequence accession numbers**

The raw ITS and 16S rRNA gene sequences reported in this article have been deposited in the NCBI Short Read Archive (SRA) under the accession number PRJNA434072.

**Acknowledgements**

This work was funded by the EU Joint Programming Initiative – A Healthy Diet for a Healthy Life (JPI HDHL, http://www.healthydietforhealthylife.eu/) to DvS (in conjunction with Science...
Foundation Ireland [SFI], Grant number 15/JP-HDHL/3280) and to MV (in conjunction with MIUR, Italy). We thank GenProbio srl for financial support of the Laboratory of Probiogenomics. This research benefits from the HPC (High Performance Computing) facility of the University of Parma, Italy. DvS is a member of The APC Microbiome Ireland supported by Science Foundation Ireland (SFI), through the Irish Government’s National Development Plan (Grant number SFI/12/RC/2273). Furthermore, DvS is a visiting professor of the University of Parma supported by Tech In Parma. The authors declare that they have no competing interests.
References


Table 1: List of *Lactobacillus* species with LITSA sequence identity ≥97 % with another *Lactobacillus* species. The percentage reported corresponds to the highest identity observed among all LITSA sequences identified in strains of the two species compared.

<table>
<thead>
<tr>
<th>Species</th>
<th>LITSA % identity with the closest species</th>
<th>Closest species</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Lactobacillus acidophilus</em></td>
<td>98</td>
<td><em>Lactobacillus amylovorus</em></td>
</tr>
<tr>
<td></td>
<td>97</td>
<td><em>Lactobacillus crispatus</em></td>
</tr>
<tr>
<td><em>Lactobacillus amylovorus</em></td>
<td>98</td>
<td><em>Lactobacillus acidophilus</em></td>
</tr>
<tr>
<td></td>
<td>97</td>
<td><em>Lactobacillus crispatus</em></td>
</tr>
<tr>
<td><em>Lactobacillus buchneri</em></td>
<td>99</td>
<td><em>Lactobacillus parabuchneri</em></td>
</tr>
<tr>
<td><em>Lactobacillus casei</em></td>
<td>99</td>
<td><em>Lactobacillus paracasei</em></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td><em>Lactobacillus rhamnosus</em></td>
</tr>
<tr>
<td><em>Lactobacillus crispatus</em></td>
<td>97</td>
<td><em>Lactobacillus acidophilus</em></td>
</tr>
<tr>
<td></td>
<td>97</td>
<td><em>Lactobacillus amylovorus</em></td>
</tr>
<tr>
<td><em>Lactobacillus curvatus</em></td>
<td>98</td>
<td><em>Lactobacillus sakei</em></td>
</tr>
<tr>
<td><em>Lactobacillus gallinarum</em></td>
<td>99</td>
<td><em>Lactobacillus helveticus</em></td>
</tr>
<tr>
<td><em>Lactobacillus gasseri</em></td>
<td>99</td>
<td><em>Lactobacillus johnsonii</em></td>
</tr>
<tr>
<td><em>Lactobacillus helveticus</em></td>
<td>99</td>
<td><em>Lactobacillus gallinarum</em></td>
</tr>
<tr>
<td><em>Lactobacillus johnsonii</em></td>
<td>99</td>
<td><em>Lactobacillus gasseri</em></td>
</tr>
<tr>
<td><em>Lactobacillus parabuchneri</em></td>
<td>99</td>
<td><em>Lactobacillus buchneri</em></td>
</tr>
<tr>
<td><em>Lactobacillus paracasei</em></td>
<td>99</td>
<td><em>Lactobacillus casei</em></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td><em>Lactobacillus rhamnosus</em></td>
</tr>
<tr>
<td><em>Lactobacillus paraplantarum</em></td>
<td>100</td>
<td><em>Lactobacillus pentosus</em></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td><em>Lactobacillus paraplantarum</em></td>
</tr>
<tr>
<td><em>Lactobacillus pentosus</em></td>
<td>100</td>
<td><em>Lactobacillus paraplantarum</em></td>
</tr>
<tr>
<td></td>
<td>99</td>
<td><em>Lactobacillus plantarum</em></td>
</tr>
<tr>
<td><em>Lactobacillus plantarum</em></td>
<td>100</td>
<td><em>Lactobacillus pentosus</em></td>
</tr>
<tr>
<td></td>
<td>99</td>
<td><em>Lactobacillus paracasei</em></td>
</tr>
<tr>
<td><em>Lactobacillus rhamnosus</em></td>
<td>100</td>
<td><em>Lactobacillus casei</em></td>
</tr>
<tr>
<td><em>Lactobacillus sakei</em></td>
<td>98</td>
<td><em>Lactobacillus curvatus</em></td>
</tr>
</tbody>
</table>
**Figure legends**

**Figure 1:** Genetic map of the Internally Transcribed Sequence (ITS) region of *Lactobacillus* with and without tRNA genes. Panel a depicts the genetic organization of the five complete ITS regions predicted in the complete genome of *Lactobacillus rhamnosus* ATCC 8530, used here as a test case. Primer sequence conservation is shown through a WebLogo representation where the overall height of the stacks indicates the sequence conservation at that position, while the height of symbols within the stacks indicates the relative frequency of nucleic acids at that position. Panel b illustrates the details of ITS regions identified in the complete genomes of species included in the mock sample for which a complete genome was available. ITS sequences without tRNA genes are highlighted in green, while ITS regions encoding tRNA genes are marked in blue. Black and red text indicates forward and reverse strand orientation, respectively.

**Figure 2:** Evaluation of the sensitivity and accuracy of the *Lactobacillus* ITS profiling protocol. The graph shows the expected and observed relative abundance of 14 *Lactobacillus* taxa constituting an artificial sample. An exponential trendline is reported for the expected and observed data.

**Figure 3:** ITS and 16S rRNA gene profiling of *Lactobacillus* species in five ecological niches. The profile of the *Lactobacillus* population obtained for: a) five human faecal samples (HG); b) five human vaginal swab samples (HV); c) five free range chicken faecal samples (FRC); d) five whey samples (WH), and e) five parmesan cheese samples (PC) is depicted in the corresponding bar plots. Only species with relative abundance >5% in at least a sample are reported. Species below 5% are collapsed in “Others <5 %.”
Figure 1

Lactobacillus rhamnosus ATCC 8530

B) ITS 1  ITS 2  ITS 3  ITS 4  ITS 5  ITS 6  ITS 7  ITS 8  ITS 9

Lactobacillus curvatus YBA201: 211 bp, 426 bp, 211 bp, 211 bp, 211 bp, 211 bp, 426 bp
Lactobacillus acidophilus NCTM15122: 130 bp, 377 bp, 130 bp, 132 bp
Lactobacillus brevis NPS QW 145: 205 bp, 205 bp, 205 bp, 205 bp, 885 bp
Lactobacillus delbrueckii KCTC 13731: 208 bp, 447 bp, 208 bp, 447 bp, 208 bp, 208 bp, 447 bp, 443 bp, 208 bp
Lactobacillus amylovorus DSM20531: 192 bp, 192 bp, 438 bp, 438 bp, 192 bp
Lactobacillus fermentum FTDX8312: 192 bp, 193 bp, 338 bp, 193 bp, 338 bp
Lactobacillus helveticus H110: 192 bp, 442 bp, 192 bp, 460 bp
Lactobacillus plantarum ZS2058: 425 bp, 196 bp, 196 bp, 196 bp, 425 bp
Lactobacillus rhamnosus ATCC 8530: 211 bp, 211 bp, 211 bp, 424 bp, 424 bp
Lactobacillus ruminis ATCC 27782: 427 bp, 205 bp, 305 bp, 405 bp, 205 bp, 205 bp
Figure 3