# Phylogenomic Analysis of Lactobacillus curvatus Reveals Two Lineages Distinguished by Genes for Fermenting Plan-Derived Carbohydrates

Lucrecia C. Terán, Gwendoline Coeuret, Raul Raya, Monique Zagorec, Marie-Christine Champomier-Verges, Stephane Chaillou

# Phylogenomic Analysis of *Lactobacillus curvatus* Reveals Two Lineages Distinguished by Genes for Fermenting Plant-Derived Carbohydrates

Lucrecia C. Terán[1], Gwendoline Coeuret[2], Raúl Raya[1], Monique Zagorec[3], Marie-Christine Champomier-Vergès[2], and Stéphane Chaillou[2,*]

[1]CERELA-CONICET, Centro de Referencia para Lactobacilos, San Miguel de Tucumán, Argentina

[2]MICALIS Institute, INRA, AgroParisTech, Université Paris-Saclay, Domaine de Vilvert, Jouy-en-Josas, France

[3]SECALIM, INRA, Oniris, Université Bretagne Loire, Nantes, France

*Corresponding author: E-mail: stephane.chaillou@inra.fr.

Accepted: May 24, 2018

## Abstract

*Lactobacillus curvatus* is a lactic acid bacterium encountered in many different types of fermented food (meat, seafood, vegetables, and cereals). Although this species plays an important role in the preservation of these foods, few attempts have been made to assess its genomic diversity. This study uses comparative analyses of 13 published genomes (complete or draft) to better understand the evolutionary processes acting on the genome of this species. Phylogenomic analysis, based on a coalescent model of evolution, revealed that the 6,742 sites of single nucleotide polymorphism within the *L. curvatus* core genome delineate two major groups, with lineage 1 represented by the newly sequenced strain FLEC03, and lineage 2 represented by the type-strain DSM20019. The two lineages could also be distinguished by the content of their accessory genome, which sheds light on a long-term evolutionary process of lineage-dependent genetic acquisition and the possibility of population structure. Interestingly, one clade from lineage 2 shared more accessory genes with strains of lineage 1 than with other strains of lineage 2, indicating recent convergence in carbohydrate catabolism. Both lineages had a wide repertoire of accessory genes involved in the fermentation of plant-derived carbohydrates that are released from polymers of $\alpha/\beta$-glucans, $\alpha/\beta$-fructans, and N-acetylglucosan. Other gene clusters were distributed among strains according to the type of food from which the strains were isolated. These results give new insight into the ecological niches in which *L. curvatus* may naturally thrive (such as silage or compost heaps) in addition to fermented food.

**Key words:** pangenome, *Lactobacillus curvatus*, plant fermentation, food, lactic acid bacteria, phylogenomic.

## Introduction

*Lactobacillus curvatus* is a facultative heterofermentative lactic acid bacterium, commonly associated with fermented and vacuum-packaged refrigerated meat and fish products (Hammes et al. 1990; Hammes and Knauf 1994; Hammes and Hertel 1998; Lyhs et al. 2002; Lyhs and Björkroth 2008; Lucquin et al. 2012; Chaillou et al. 2015). In addition, this species has also been isolated from dairy products such as milk and cheese (Kask et al. 2003). More recently, many studies have identified *L. curvatus* in fermented plant products like sauerkraut (Vogel et al. 1993), sourdough (Michel et al. 2016), radish pickles (Nakano et al. 2016), and kimchi (Jung et al. 2011); in other plant-derived materials like honey (Bulgasem et al. 2016); or from the environmental fermentation process of corn or grass silage (Tohno et al. 2012; Zhou et al. 2016). These observations suggest that *L. curvatus* is ubiquitous in lactic acid fermentation and that foods of vegetable origins are a common environment for this species. Based on this, it is perhaps not surprising that *L. curvatus* has also been identified in the feces or gut of many animal species that feed on plants or cereals, including snails (Koleva et al. 2014), chickens (Zommiti et al. 2017), and humans (Dal Bello et al. 2003).

*Lactobacillus curvatus* belongs to the *Lactobacillus sakei* clade of psychrotrophic *Lactobacillus*, which comprises four species: *Lactobacillus sakei, Lactobacillus fuchuensis, Lactobacillus graminis*, and *L. curvatus* (Sun et al. 2015;

Zheng et al. 2015). To date, thirteen *L. curvatus* genomes are available, of which five are complete and eight are draft (Hebert et al. 2012; Cousin 2015; Nakano et al. 2016; Inglin et al. 2017; Jans et al. 2017; Lee et al. 2017; Terán et al. 2017). Some of these strains were sequenced to highlight their ability to produce multiple bacteriocins, like strain CRL705, which was isolated from an Argentinean artisanal dry sausage (Hebert et al. 2012), or because of surprising features of flagellum-mediated motility, like the Japanese strain NRIC0822, which was isolated from sushi (Cousin 2015). Besides these few examples, very little is known about the intraspecies genomic repertoire of *L. curvatus* strains. In particular, improved knowledge of this species' genome could help to distinguish it from the closely related species *L. sakei*, which is known to be separated into three phylogenic lineages (Chaillou et al. 2013).

Several recent publications have reported the genome sequencing of strains of *L. curvatus* isolated from various food products. We took advantage of these resources to perform a detailed phylogenomic and pangenomic analysis of *L. curvatus* as a species. In order to improve our understanding of the evolution and population structure of the 13 strains studied, we performed multiple, complementary analyses. These included an evolutionary analysis of the core genome, which revealed the existence of two lineages, and an in-depth comparison of the biological functions encoded in the accessory genome, which highlighted the strong relationship of this species with different plant-based environments.

## Materials and Methods

### Genome Data and Curation of Annotations

Our data set consisted of 13 genomes of *L. curvatus* strains, of which five were complete and eight were draft versions, all available from the Genbank/EMBL databases (table 1). All genomes were downloaded to the MAGE annotation platform (Vallenet et al. 2013) and strain-specific genes were all manually curated in order to standardize the comparative analysis. Metabolic pathways were reconstructed using the METACYC database (Caspi et al. 2016) or from the literature when indicated.

### Pangenome Analysis and Clusters of Orthologous Genes

The composition of the core and variable genomes was calculated using a pairwise estimation of orthologous proteins in CDhit (Li and Godzik 2006) at a threshold of 80% identity on 80% of the protein's total length. We then modeled the progression of pangenome size with respect to the number of genomes included by randomly picking genomes and iterating the process 13 times, as described on the MAGE annotation platform (Vallenet et al. 2013). R statistical software (R Development Core Team 2010) and the HEATMAP.2 function of the GPLOT package were used to construct a heatmap based on the variable components of the pangenome. The Euclidean distance based on presence/absence was used to calculate the distance matrix between the variable genomes, and clustering was performed by unsupervised complete linkage.

### Evolutionary and Phylogenomic Analysis

The alignment of the nucleotide genome sequences was performed using PROGRESSIVEMAUVE (Darling et al. 2010); from this, we extracted the core alignment by keeping only the regions where all genomes aligned over at least 500 bp. This core alignment was submitted to five independent runs of CLONALFRAME software (Didelot and Falush 2006; Vos and Didelot 2009), which consisted of a burn-in cycle of the MCMC (Markov Chain Monte Carlo) algorithm fixed to 50,000 iterations and a posterior sampling of 50,000 iterations. The prior iterations were discarded and model parameters were sampled in the second period of the run every 50 iterations, resulting in 1,000 samples from the posterior. Satisfactory convergence of the MCMC algorithm in the different runs was estimated based on the Gelman-Rubin statistic calculated in CLONALFRAME. The genealogy of the population was summarized and the robustness of the tree topology was evaluated by concatenating the posterior samples of the five runs to build a 50% majority rule consensus tree using the CLONALFRAME GUI and MEGA6 software (Tamura et al. 2013). From these runs, several measurements were also taken, such as $\rho/\theta$ (relative frequencies of occurrence of recombination and mutation) and r/m (relative impact of recombination and mutation in the diversification of the lineages). A Bayesian approach, implemented in STRUCTURE software version 2.3 (Pritchard et al. 2000; Falush et al. 2003), was used to infer the lineage ancestry of the core genome by assuming that each strain derived all of its SNPs from one of the $K$ ancestral subpopulations. The number of populations, $K$, was determined under the linkage model. Five individual runs per value of $K$ (chosen to be 2 or 3) were performed using 50,000 burn-in iterations and 50,000 sampling iterations.

## Results and Discussion

### Evolutionary Analysis Reveals the Existence of Two Lineages Within *L. curvatus*

The phylogeny of *L. curvatus* was inferred using two different approaches. The first strategy was based on Bayesian inference with the coalescent model implemented in CLONALFRAME software (Didelot and Falush 2006) whereas the second strategy was to statistically estimate the probable number of ancestral subpopulations ($K$) within the genetic population of strains; this was performed using STRUCTURE with the linkage model (Pritchard et al. 2000; Falush et al. 2003). The initial step of these two analysis consisted of the selection of the high-quality core genome using PROGRESSIVEMAUVE software (Darling et al. 2010), which

**Table 1**
List of *Lactobacillus curvatus* Genome Sequences Used in This Study

| Strain | Origin | Chromosome (Mb) | Sequencing Status (Sequencing Technology) | Number of Contigs | Number of CDS | Accession Number | Reference |
|---|---|---|---|---|---|---|---|
| FBA2 | Radish/Carrot pickles | 1.849 | Complete (PacBio RS II platform) | 1 | 1,718 | CP016028.1 | Nakano et al. (2016) |
| Wikim38 | Baechu (Chinese Kimchi) | 1.940 | Complete (PacBio RS II platform) | 1 | 1,810 | CP017124.1 | Lee et al. (2017) |
| Wikim52 | Kimchi | 1.987 | Complete (PacBio RS II platform) | 1 | 1,875 | CP016602.1 | NP |
| MRS6 | Fermented meat | 2.114 | Complete (PacBio RS II platform) | 1 | 1,935 | CP022474.1 | Jans et al. (2017) |
| KG6 | Fermented meat | 2.002 | Complete (PacBio RS II platform) | 2 | 1,884 | CP022475.1-76.1 | Jans et al. (2017) |
| CRL705 | Argentinean dry-sausage | 1.838 | Draft (454 GS Titanium pyrosequencing) | 145 | 1,708 | GCA_000235705.2 | Hebert et al. (2012) |
| FLEC03 | Vacuum-packed beef carpaccio | 1.902 | Draft (Illumina MiSeq pair-end) | 47 | 1,944 | GCA_900178545.1 | Terán et al. (2017) |
| DSM20019 | Milk | 1.815 | Draft (Ion Torrent PGM) | 72 | 1,828 | GCA_001311645.1 | NP |
| NRCI0822 | Kabura zushi | 1.945 | Draft (Illumina HiSeq pair-end) | 144 | 1,831 | GCA_000805355.1 | Cousin (2015) |
| RI-406 | Meat | 2.001 | Draft (Illumina MiSeq pair-end) | 52 | 1,873 | GCA_001981905.1 | Inglin et al. (2017) |
| RI-198 | Meat | 1.804 | Draft (Illumina MiSeq pair-end) | 77 | 1,727 | GCA_001981925.1 | Inglin et al. (2017) |
| RI-193 | Meat | 1.805 | Draft (Illumina MiSeq pair-end) | 82 | 1,727 | GCA_001982045.1 | Inglin et al. (2017) |
| RI-124 | Meat | 1.810 | Draft (Illumina MiSeq pair-end) | 77 | 1,722 | GCA_001982025.1 | Inglin et al. (2017) |

NP, no publication available.

took into account only the coding sequences (CDS) and the aligned regions with no frameshift between the 13 chromosomes. Therefore, all CDS of potential low-sequencing quality (i.e., due to the draft genome sequencing status) were discarded from this analysis. A total of 131 alignment blocks longer than 500 bp were selected. The core aligned genome consisted of 199,762 bp, which contained 6,742 Single Nucleotide Polymorphic loci (SNPs) among the 13 chromosomes. Results are shown in figure 1. From the coalescent tree, we were able to define two separate lineages: lineage 1 comprised the strains FLEC03, MRS6, and RI-406, and lineage 2 comprised the ten other strains. The results of the STRUCTURE analysis confirmed the presence of two populations. Furthermore, this method enabled the characterization of the allele frequencies at each locus, then it probabilistically assigned individuals to $K$ (unknown) ancestral populations. For both lineages (when $K$ was set to two), >75% of the genetic contribution to each strain came from its own group. However, we observed less genetic homogeneity in one clade of lineage 2 (named clade 2B), which contained strains RI-124, RI-193, RI-198, and KG6. We therefore investigated if inference to $K = 3$ ancestral populations would cause this clade to be assigned to a third lineage, a population structure which would be similar to that observed in the closely related species *L. sakei* (Chaillou et al. 2013). However, there was no statistical support for this hypothesis, suggesting that clade 2B does not originate from a third ancestral population. Therefore, at this stage of the analysis it could only be concluded that strains RI-124, RI-193, RI-198, and KG6 are most likely affiliated to the broader lineage 2 but have some degree of admixture (from 35% to 45%) with lineage 1. The CLONALFRAME analysis estimated statistically the $\rho/\theta$ ratio, a measure of how often recombination events occur relative to neutral genetic drift (mutation). This value was 0.137 (0.127–0.147 at 95% credibility interval), which indicated that the recombination rate is significantly lower than the mutation rate and therefore, recombination has played only a minor role in the evolution of the two lineages. Nevertheless, the admixture status of strains RI-124, RI-193, RI-198, and KG6 clearly indicated that recombination events between the two lineages may occur, perhaps when strains from both lineages are in physical proximity such as in solid food.

## The Accessory Genome Corroborates the Existence of Two Lineages

Another way to assess population structure is to perform a comparative analysis of the variable or accessory genome. Indeed, the speciation of a species into several lineages may arise from positive selection for a given ecological niche and thus for the acquisition of lineage-specific metabolic traits (Doolittle and Papke 2006). The general genome features of the *L. curvatus* strains, shown in table 1, clearly highlighted some genetic variability (from 1708 CDS in strain CRL705 to 1944 CDS in strain FLEC03). We thus compared the 13 strains in terms of gene content by estimating the core shared genome and then evaluating how the accessory genome could contribute to the differentiation of the two lineages. The results, shown in figure 2, first indicated that the core genome is composed of 1,215 clusters of orthologous genes (orthologs) whereas the pangenome contains 3,435 orthologs. It

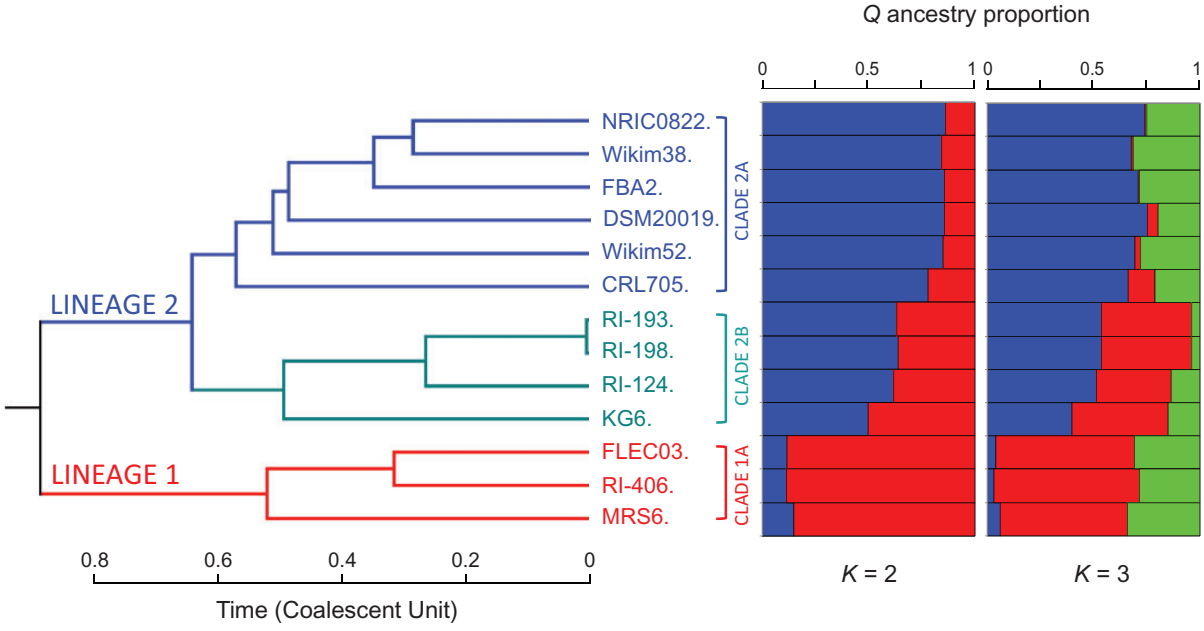


**Fig. 1.**—Phylogenomic clonal genealogy and population structure of *Lactobacillus curvatus* strains. On the left: Fifty-percent majority rule consensus tree inferred with CLONALFRAME software using the coalescent model (branch lengths are given in coalescent units). All branches are supported by a posterior probability of >95%. Strains are colored according to their lineage affiliation. On the right: Proportion of genetic material derived from each of *K* subpopulations as inferred by STRUCTURE (linkage model) and assuming *K*=2 or 3 populations. Ancestral subpopulations are colored in red (lineage1), blue (lineage 2), and green (an unlikely lineage 3), respectively. Clade 2B of lineage 2 is colored in dark cyan to highlight its divergence from Clade 2A and a strong degree of admixture between the two lineages.

should be noted that although it has been shown that draft assemblies provide highly relevant insights into pangenomic studies (Sun et al. 2015), the use of these types of assemblies may raise the possibility to underestimate the core genome, the mobile genome and the strain-specific genes. For this reason, we then relaxed the threshold used to estimate the core genome by including the possibility that a given gene might be absent in one of the thirteen strains, a possibility that could occur due to the draft sequencing of eight out of the 13 strains studied. With these settings, the core genome was estimated to contain 1,407 orthologs. Of these, 414 orthologs formed the mobile genome (IS elements and prophages), with 55 different putative transposase families represented overall. The cloud genome (genes present in only one strain and not from the mobile genome) contained 648 orthologs which were mainly distributed into three major groups: proteins of unknown functions (50.1%), cell-surface or exported proteins (15%), and proteins involved in the production of surface or exported polysaccharides or teichoic acid (20%). The remaining 901 orthologs (when the relaxed threshold was used to estimate the core genome) form part of the accessory genome, in which genes are shared by at least two strains. This group of 901 orthologs from the accessory genome was used for cluster analysis of the strains (fig. 3). We observed that strains grouped according to their lineage, with the exception of strains from clade 2B; these shared more accessory genes with strains of lineage 1 than with the other strains of their own putative lineage. It is important to remember that the core genome analysis is based on the mutation and recombination rates among SNPs in housekeeping genes and thus reflects a rather long-term evolutionary process. Instead, the accessory genome analysis is affected to a large degree by horizontal gene transfer, which might be influenced by the lifestyle of the strains and represents a recent and ongoing process of fitness acquisition by the strains. Therefore, the striking finding of a discrepancy between core and accessory genome clustering suggests that strains from clade 2B have recently evolved from lineage 2 through the acquisition of functional traits from lineage 1. It should be noted that strains FLEC03, RI-406, RI-124, RI-193, RI-198, and MRS6 share a common source of isolation (meat), whereas the other strains were isolated from fermented nonmeat products (except for CRL705). Furthermore, strains isolated from Asian-type food products (sushi and kimchi) formed a closely related subgroup of strains in lineage 2. Therefore, patterns in the accessory genome of *L. curvatus* suggest that certain traits that affect environmental fitness have been recently acquired. However, it should be acknowledge that a bias on the origin of strains might still exist since eight out thirteen of the strains being sequenced and publicly available are from fresh or fermented meat. It would therefore be valuable to sequence more strains from other sources (vegetables and silage) to validate this conclusion. Based on figure 3, several groups of accessory genes were defined (A to E)

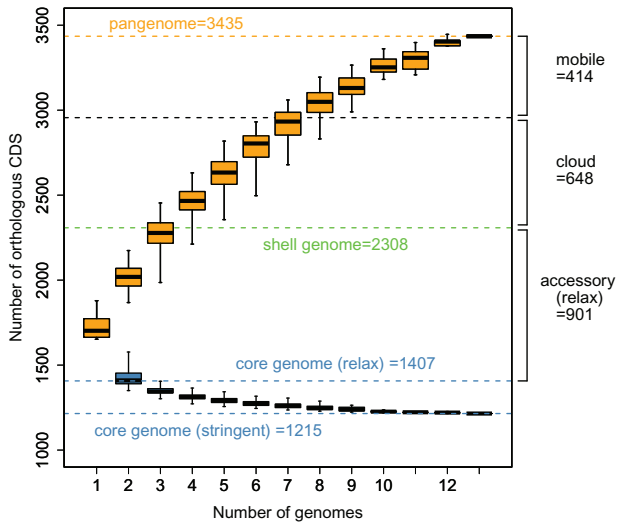


**Fig. 2.**—Progression of the core genome and pangenome of *Lactobacillus curvatus*. Each boxplot represents the pairwise evolution of the core genome (blue) and pangenome (yellow) of clusters of orthologous proteins calculated iteratively as genomes were added to the analysis, for a total of 2–13 genomes. Dashed lines represent the values obtained for the progression of the core genome (using a stringent or relaxed estimation; see text), the pangenome, and for another important step in the estimation of the accessory genome, the shell genome. The shell genome is a more realistic functional estimation of the pangenome that excludes mobile selfish DNA (mobile genome) and unique gene clusters found in only one strain (cloud genome) from the sum of accessory genes.

according to their frequency in the different strains and these will be addressed later in the discussion.

## Cell-Surface Complexes as a Major Difference between the Two Lineages

Cell-surface complexes (Cscs) are conserved among Firmicutes and, in particular, in species belonging to the order *Lactobacillales*. Cscs are multicomponent complexes composed of four types of protein families which differ according to their domains: CscA has a DUF916 domain of unknown function, CscB and CscC contain a WxL1 and WxL2 domain which binds noncovalently to the murein polymer of the cell wall, and CscD contains an LPxTG motif for covalent anchoring to the cell wall (Siezen et al. 2006). Csc components can vary in number and position among clusters and not all of them are necessarily present in one complex. In particular, CscCs are large secreted proteins with putative carbohydrate polymer binding domains that are involved in adhesion and/or carbohydrate scavenging. They are often highly variable between gene clusters and were shown to cross-react with CscA and CscB proteins of noncognate gene clusters (Brinster et al. 2007). Of the accessory orthologous genes investigated here, we found Csc-encoding regions in both groups A and D (see fig. 3). Strains FLEC03, RI-406, and MRS6 of lineage 1 share eight putative Csc clusters, two of which are also shared with strains RI-124, RI-193, and RI-198 from the admixed clade 2B. These gene clusters are very similar to those previously identified in *L. sakei* strain 23 K (Chaillou et al. 2005), which were hypothesized to be important for growth fitness in meat. It is therefore interesting to observe that such cell surface complexes are almost absent from strains in lineage 2, which were not isolated from fresh meat.

## A Wide Repertoire of Phosphotransferase Transport Systems for Sugar and Polyol Uptake

The accessory genome of the thirteen *L. curvatus* strains contains >30 phosphotransferase systems (PTSs) and 10 other systems dedicated to carbohydrate uptake. Altogether, these account for 240 genes (30.8% of the accessory genome shown in fig. 3). Interestingly, it seems that the repertoire of PTS gene clusters enables the utilization of a wide range of carbohydrates and polyols from plants (vegetables, fruits, cereals) and insect/microbe-derived sugar polymers. In particular, a significant proportion of these gene clusters encode uptake systems for the utilization of $\alpha/\beta$-glucan, $\alpha/\beta$-fructan, and N-acetylglucosan. An overview of these systems with respect to the plant carbohydrates that they transport is shown in figure 4 and details about the gene content of these clusters can be found in supplementary table S1, Supplementary Material online.

### *Systems for α– and β-Glucans*

We found at least three different systems for maltose utilization. Two of these use starch and maltodextrins via the intracellular α-amylase pathway and the maltose phosphorylase pathway, both of which are linked to an ABC transporter (*map* gene cluster N°1 an N°2). Interestingly, these two gene clusters have a slightly different gene synteny (gene *mapG* encoding a hypothetical protein is respectively absent from cluster N°1 and gene *mapL1* encoding an oligo alpha-1,6-glusosidase is absent from cluster N°2). Furthermore, the encoded proteins were not considered to be orthologs at a threshold of 80% similarity, which indicates that they have different phylogenetic origins. One gene cluster was present in all strains of lineage 1 (strains FLEC03, RI-406, and MRS6), but was also found in strains KG6 and FBA2 of lineage 2; it shares a high level of identity (~75%) with homologs in *Lactobacillus alimentarius*. Conversely, the second gene cluster was found only in strains of lineage 2, with a high level of identity (~70%) to homologs in *Pediococcus pentosaceus*. The third system for maltose utilization (cluster N°3), present only in strains NRIC0822 and MRS6, was the *mal* PTS, which was coupled with the *malA* gene that encodes 6-phospho α-glucosidase.

Ten of the thirteen strains also contained genes coding for the *tre* PTS (cluster N°4), which would enable them to use the α-glucan-derived disaccharide trehalose; the three strains that

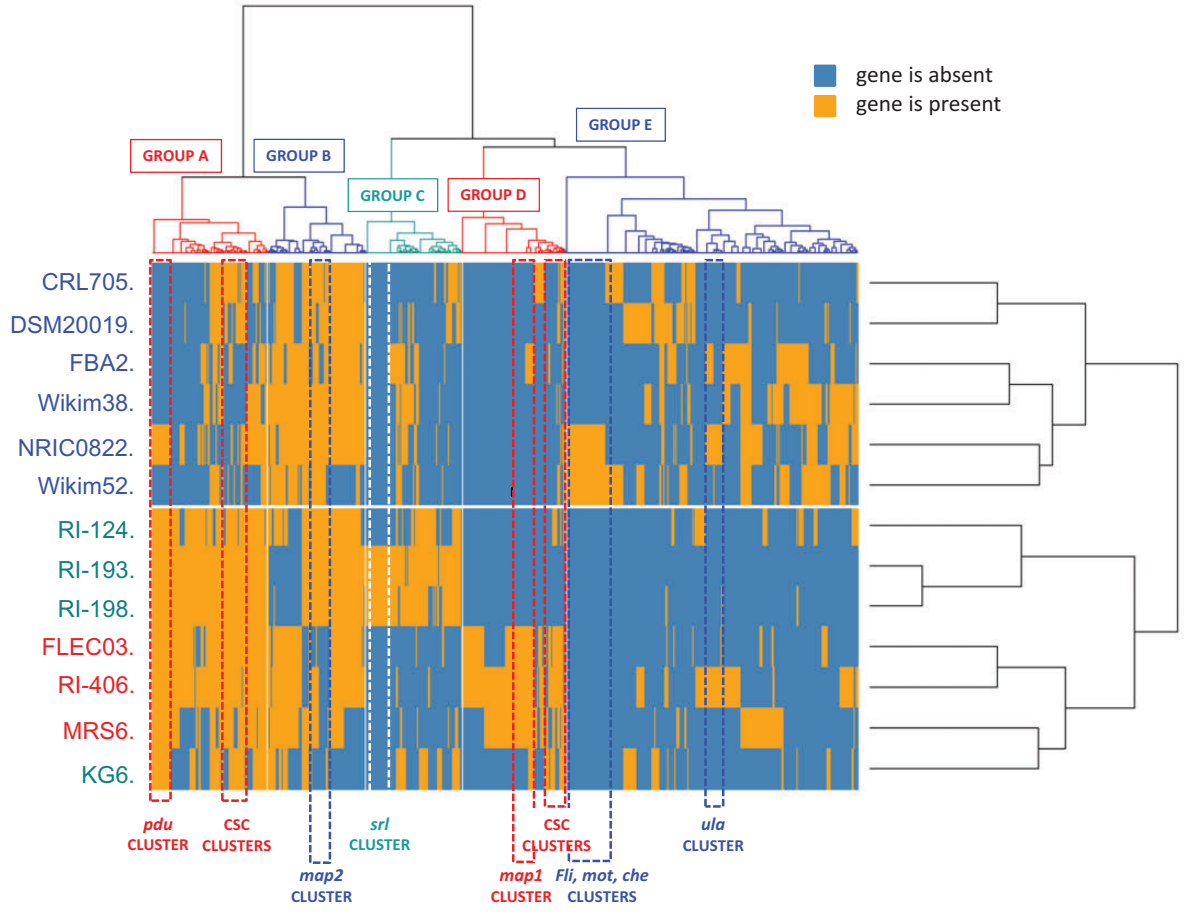


**Fig. 3.**—Heatmap showing the clustering analysis of *Lactobacillus curvatus* strains based on the content of their accessory genomes. Unsupervised complete linkage clustering of *L. curvatus* strains based on the presence (orange) or absence (blue) of 901 orthologs that constitute the *L. curvatus* accessory genome (without mobile and cloud genomes). Names of strains are colored according to their phylogenomic clade and lineage as shown in figure 1. The five main groups of orthologs prevalent among the strains are indicated above the heatmap and inside the clustering tree (from groups A to E). Similarly, gene prevalence groups are colored based on their specificity to each of the phylogenomic clades. Eight gene clusters representative of each groups are boxed with dashed lines: CSC clusters (Cell Surface Complexes); *pdu* cluster (propanediol catabolic pathway); *map1* and *map2* clusters (lineage-specific maltose phosphorylase pathways); *srl* cluster (sorbitol phosphotransferase system); *fli*, *mot*, and *che* clusters (motility operons); *ula* cluster (ascorbate catabolic pathway).

lacked this system were RI-124, RI-193, and RI-198 from the admixed clade 2B. There was much more redundancy in PTSs for the use of $\beta$-glucans, in which there was also extensive variation among strains. These systems included a *lic* PTS (cluster N°5) for catabolizing lichenan (barley glucan) and several *bgl*-like PTSs (up to three systems within strains of lineage 1 from cluster N°06a to N°06c, although some of these may not be complete). However, one peculiar *bgl* system (cluster N°6c) in strain RI-406 also encoded an $\alpha$-xylosidase (xylQ), indicating that this cluster might be involved in the degradation of xyloglucan (plant hemicellulose) (Chaillou et al. 1998). We found evidence that all strains are able to take up glucose and fructose with the *manXYZ* and *fruKRI* PTSs, respectively, but strain FLEC03 and RI-406 from lineage 1 also had an additional copy of the *manXYZ* PTS (cluster N°7).

### *Systems for α– and β-Fructans*

Similarly, it appeared that some strains from lineage 2 are able to use sucrose through two different pathways, one involving the sucrose *src* PTS and sucrose-6-phosphate hydrolase pathway (clusterN° 08a and 08b), and the other involving a symport system coupled with the catabolism of bacterial levan ($\beta$-fructan) by the *lev* PTS (identified in strains FAB2 and Wikim52) (cluster N°12). Another PTS for fructose utilization was identified in strains FLEC03 and RI-406 from lineage 1 and strains RI-193, RI-198, and NRIC0822 from lineage 2; specifically, we detected the *frl* gene cluster (cluster N°10), which encodes a fructose-lysine deglycation pathway (Wiame et al. 2005). This molecule can be abundant in plant fluids, arising spontaneously via condensation of the sugar and the amino acid when both are present in high concentrations (Bilova et al. 2016). The *frl*

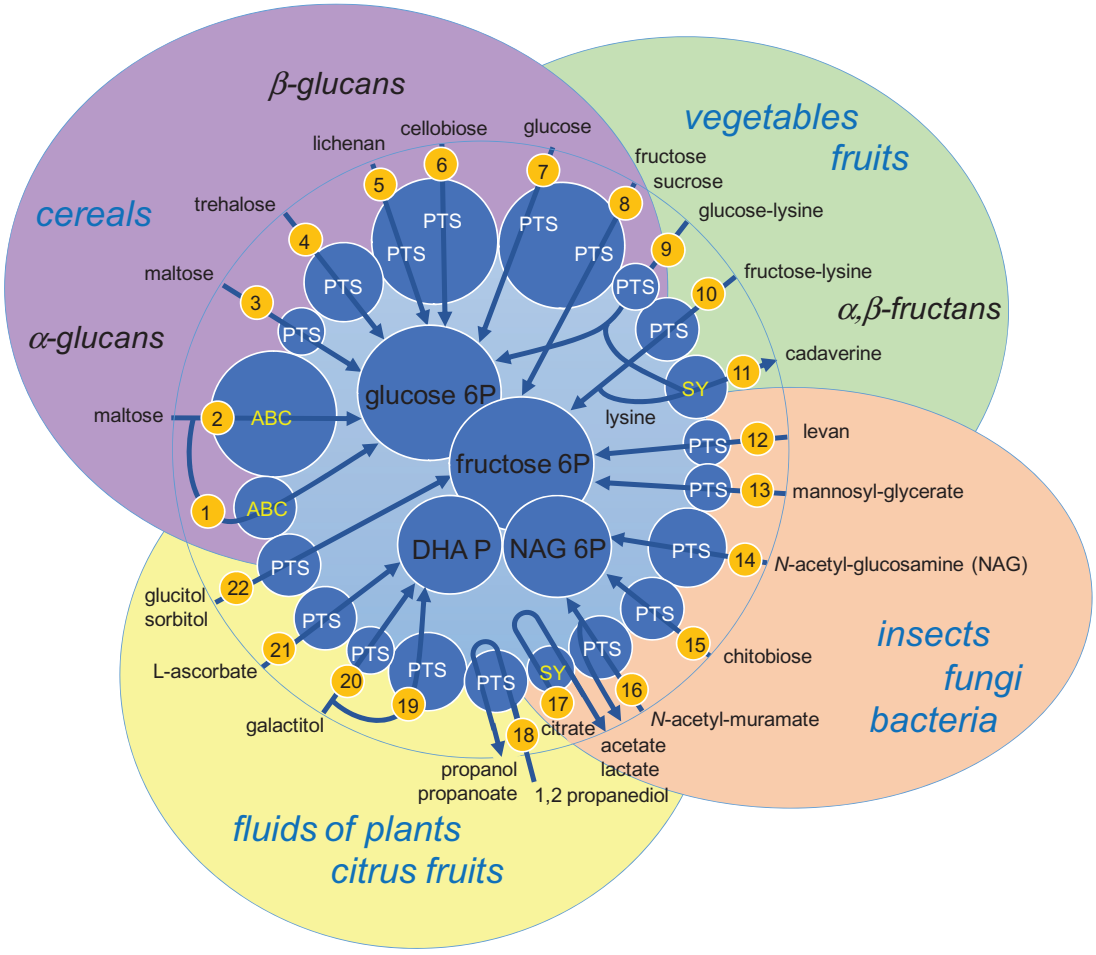


**Fig. 4.**—Overview of *Lactobacillus curvatus* accessory gene repertoire involved in fermentation of plant-derived carbohydrates. Carbohydrates are grouped (external ellipses) according to their type (glucan and fructans) or origin (plants, cereals, and insects). Each uptake and catabolic system is represented by a circle whose size depicts the degree of conservation among the *L. curvatus* strains. The clusters are numbered (small yellow circles) to facilitate their identification using supplementary file S1, Supplementary Material online. The inner circles illustrate the fate of these carbohydrates: into glucose 6P, fructose 6P, N-acetyl glucosamine 6P, or dihydroxy-acetone P (DHAP).

PTS cluster is associated with a fructose 6P-lysine deglycase (*frlF*), which releases fructose 6 P and lysine. The expression of this gene cluster might be controlled by an accessory $\sigma$54-like transcriptional factor, which would indicate that the bacteria might be able to sense this compound in the environment (Francke et al. 2011). The genomes of strains NRIC0822 and Wikim52 contain a second putative glycation PTS (*grl* gene cluster N°09). It should be noted that two pathways for the catabolism of lysine into cadaverine were found (cluster N°11b): a pyridoxal-dependent lysine decarboxylase (*tdcA*) was present in most strains (except strains DSM20019 and Wikim52), while five strains also contained the lysine decarboxylase complex encoded by the *cad* gene cluster (clusterN°11a).

### *Systems for Polyols*

Plant fluids are rich in polyols and vitamin C (ascorbic acid). *Lactobacillus curvatus* strains, in particular those in lineage 2, have acquired multiple PTSs that are specific for these compounds, including a catabolic *ula* PTS pathway for ascorbic acid (Yew and Gerlt 2002) (cluster N°21), a catabolic *srl* PTS pathway for sorbitol (Alcántara 2008) (cluster N°22), and a catabolic *gat* PTS pathway for glucitol/galactitol (clusters N°19 and N°20). Again, strains from lineage 1 and those of the admixed clade 2B had *gat* gene clusters (Nobelmann and Lengeler 1996) that differ in sequence homology and origin from that of strains from lineage 2. Similarly, strains of lineage 1 could also be distinguished from those of lineage 2 by the presence of a pathway that is quite uncommon in lactic acid bacteria (Cluster N°18): a coenzyme B(12)-dependent catabolic pathway (*pdu* gene cluster) for the utilization of 1,2-propanediol (Bobik et al. 1999). This compound is produced by the fermentation of the common plant sugars rhamnose and fucose, and its catabolism creates propionate and propanol as end products.

### *Other Systems*

The differences between the two lineages of *L. curvatus* or between some groups of strains go beyond catabolic pathways for plant-derived carbohydrates. For instance, some strains in lineage 2 (CRL705, DSM20019, and Wikim38) have an *rbsUDKR* gene cluster for ribose catabolism that encodes a ribose transporter *rbsU* and is similar to that of strains of *L. sakei* (Stentz and Zagorec 1999). Instead, all other strains have an *rbsABCDKR* gene cluster which encodes an ABC transporter *rbsABC*. Another example of divergence between the two lineages was also found in the catabolism of N-actetylglucosamine and N-acetylmuramic acid. Whereas most strains (with the exception of FLEC03) harbored the *nagC* PTC and the *nagA* gene (cluster N°14), which encodes the N-acetylglucosamine 6-phosphate deacetylase (Vogler and Lengeler 1989), strains CRL705, DSM20019, and KG6 had an additional *mur* PTS together with the *murQ* gene (cluster N°16), which encodes the D-lactyl ether N-acetylmuramic 6-phosphate acid etherase (Dahl et al. 2004) for the catabolism of N-acetylmurein. Strains from lineage 1 and clade 2B had another unique variation, with what might be a *chi* gene cluster together with *chiK* (cluster N°15); the *chi* cluster is involved in chitobiose uptake and catabolism, and *chiK* encodes an N-acetylglucosamine kinase (Plumbridge and Pellegrini 2004). More surprisingly, strains DSM20019 and MRS6 harbored a *mng* gene cluster (cluster N°13) similar to that of *E. coli*, which has been shown to be involved in 2-*O*-α-Mannosyl-D-glycerate PTS-dependent uptake and catabolism (*mngB* encodes an α-mannosidase; Sampaio et al. 2004). This unusual carbohydrate is known to be abundant in hyperthermophilic prokaryotes, in which it acts as an osmoprotectant. Finally, three strains—FAB2, Wikim38, and RI-406—possessed the *cit* gene cluster (cluster N°17), which is involved in the decarboxylation of citrate to acetate and pyruvate, a common catabolic pathway in *Leuconostoc* (Marty-Teysset et al. 1996). Together, these observations suggest that *L. curvatus* strains may also thrive in natural environments where microbial and insect cell-wall polymeres can be scavenged as well as those derived from plants, environments such as silage or compost heaps.

## Mobile Genome Shows Important Variations between Strains

The mobile genome differs greatly between strains of the two lineages (fig. 5). Strains from clade 2A, and in particular those isolated from Asian types of food, had a broader range of transposase families (up to 20 in strain FAB2, see fig. 5), indicating that gene transfer may occur more frequently in these types of foods than in meat or in environmental silage. However, Schmid et al. (2018) pointed out recently that

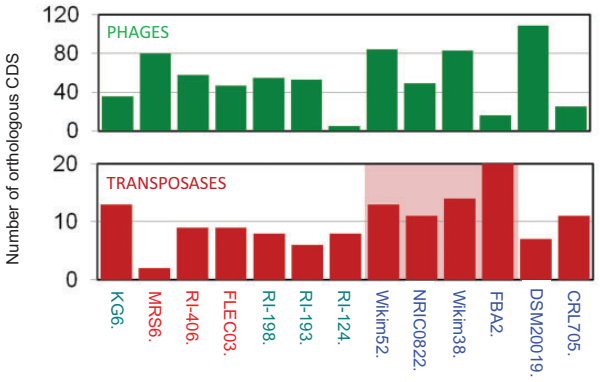


**Fig. 5.**—Barplots showing the number of phage-related genes and transposase families in the *Lactobacillus curvatus* strains. The barplots indicate the numbers of phage genes (green) or transposase families (red) identified in each strain. The red shaded area depicts strains from Asian-type foods, in which a higher number of transposase families was found.

mobile elements predominate among the genes that are not captured in fragmented Illumina-based genome assemblies. The set of four Asian L. curvatus genomes (FBA2, Wikim38, Wikim52, NRCI0822; table 1) are enriched in complete genomes (3 out of 4) and this observation might explain the higher number of transposase families in these strains. Between one and three putative prophages were identified in each genome in this group. All prophages were predicted to be noncontratile tail phages of the family Siphoviridae of order Caudoviridae, and their size was between 31 and 42 kb, both characteristics frequently found among prophages of genus *Lactobacillus* (Mahony and van Sinderen 2014). There was little sequence similarities between the different phages genomes, each phage being unique to one strain and this was explained the large contribution of phages to the important size of the mobile genome. It is interesting to note that strain DSM20019 was the richest in prophage content; this strain was isolated from milk, where the concentration of phages is high (from $10^1$ to $10^4$ phages per milliliter) (Marcó et al. 2012). Instead, strain CRL705, which might have been associated with a milk environment in the past because of the presence of a lactose PTS cluster in its genome, only contains remnants of prophages. However, strain CRL705 was unique in possessing two CRISPR/cas systems (Clustered Regulatory Short Palindromic Repeats; Deveau et al. 2010); in addition to the type II system (cas9 gene) present in all strains, CRL705 also had a type I system (cas3). These two clusters might provide strain CRL705 with a stronger immunity against phages. Finally, as it could be expected from the weak assembly performance of repetitive regions using short-read sequencing, the CRISPR spacer regions were largely incomplete in draft genomes and no conclusive information could be extracted on the possible history of the strains versus phages encounter.

### Motility

The mobility operon, which comprises the *fli* (flagellar structral complex), *mot* (flagellar motor complex), and *che* (chemotaxis regulatory complex) gene clusters, was initially characterized in strain NRIC0822 (Cousin 2015). However, we also identified it in the closely related strain Wikim52, suggesting that this feature might not be so unusual among strains of *L. curvatus*.

## Conclusion

Our results showed that, as a species, *L. curvatus* is divided into two ancestral phylogenetic lineages. The traces of this evolutionary path are not only present in the allele frequencies of the core genes but also in the origin and structure of some conserved metabolic gene clusters (i.e., ribose, maltose, galactitol). The degree of variation present in these systems suggests that the two lineages result from different evolution mechanisms ending up to this repertoire. Furthermore, our work demonstrates that the lifestyle and the ecological niche of the strains has a strong influence on the gene content of the accessory genome, which has led to convergence between strains from lineage 1 and those of clade 2B from lineage 2. *Lactobacillus curvatus* pangenome has revealed a wide repertoire of genes for catabolizing plant-derived carbohydrates, and this capacity is representing a major difference with the closely related species *L. sakei*. Finally, an in-depth analysis of the *L. curvatus* accessory genome has led us to conclude that, in addition to living in fermented foods made of vegetables or meat, the species must also thrive in an ecological niche where decaying plants, insects, and bacteria are present in large amounts, for example, in silage or compost heaps.

## Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

## Acknowledgments

L.C.T. was the recipient of a PhD fellowship from the BecAR Program and Campus France Argentine. The LABGeM (CEA/IG/Genoscope and CNRS UMR 8030) and the France Génomique National infrastructure (funded as part of Investissement d'avenir program managed by Agence Nationale pour la Recherche contract no. ANR-10-INBS-09) are acknowledged for support within the MicroScope annotation platform. We are grateful to the INRA MIGALE bioinformatics platform (http://migale.jouy.inra.fr) for providing computational resources and data storage.

## Literature Cited

Alcántara C. 2008. Regulation of *Lactobacillus casei* sorbitol utilization genes requires DNA-binding transcriptional activator GutR and the conserved protein GutM. Appl Environ Microbiol. 74(18):5731–5740.
Bilova T, et al. 2016. A snapshot of the plant glycated proteome: structural, functional, and mechanistic aspects. J Biol Chem. 291(14):7621–7636.
Bobik TA, Havemann GD, Busch RJ, Williams DS, Aldrich HC. 1999. The propanediol utilization (pdu) operon of *Salmonella enterica* serovar Typhimurium LT2 includes genes necessary for formation of polyhedral organelles involved in coenzyme B(12)-dependent 1,2-propanediol degradation. J Bacteriol. 181(19):5967–5975.
Brinster S, Furlan S, Serror P. 2007. C-terminal WxL domain mediates cell wall binding in *Enterococcus faecalis* and other gram-positive bacteria. J Bacteriol. 189(4):1244–1253.
Bulgasem BY, Lani MN, Hassan Z, Wan Yusoff WM, Fnaish SG. 2016. Antifungal activity of lactic acid bacteria strains isolated from natural honey against pathogenic *Candida* species. Mycobiology 44(4):302–309.
Caspi R, et al. 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 44(D1):D471–D480.
Chaillou S, et al. 1998. Cloning, sequence analysis, and characterization of the genes involved in isoprimeverose metabolism in *Lactobacillus pentosus*. J Bacteriol. 180(9):2312–2320.
Chaillou S, et al. 2005. The complete genome sequence of the meat-borne lactic acid bacterium *Lactobacillus sakei* 23K. Nat Biotechnol. 23(12):1527–1533.
Chaillou S, et al. 2015. Origin and ecological selection of core and food-specific bacterial communities associated with meat and seafood spoilage. ISME J. 9(5):1105–1118.
Chaillou S, Lucquin I, Najjari A, Zagorec M, Champomier-Vergès M-C. 2013. Population genetics of *Lactobacillus sakei* reveals three lineages with distinct evolutionary histories. PLoS One 8(9):e73253.
Cousin FJ. 2015. Detection and genomic characterization of motility in *Lactobacillus curvatus*: confirmation of motility in a species outside the *Lactobacillus salivarius* clade. Appl Environ Microbiol. 81(4):1297–1308.
Dahl U, Jaeger T, Nguyen BT, Sattler JM, Mayer C. 2004. Identification of a phosphotransferase system of *Escherichia coli* required for growth on N-acetylmuramic acid. J Bacteriol. 186(8):2385–2392.
Dal Bello F, Walter J, Hammes WP, Hertel C. 2003. Increased complexity of the species composition of lactic acid bacteria in human feces revealed by alternative incubation condition. Microb Ecol. 45(4):455–463.
Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5(6):e11147.
Deveau H, Garneau JE, Moineau S. 2010. CRISPR/Cas system and its role in phage-bacteria interactions. Annu Rev Microbiol. 64(1):475–493.
Didelot X, Falush D. 2006. Inference of bacterial microevolution using multilocus sequence data. Genetics 175(3):1251–1266.
Doolittle WF, Papke RT. 2006. Genomics and the bacterial species problem. Genome Biol. 7(9):116.
Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164(4):1567–1587.
Francke C, et al. 2011. Comparative analyses imply that the enigmatic Sigma factor 54 is a central controller of the bacterial exterior. BMC Genomics 12(1):385.
Hammes WP, Bantleon A, Min S. 1990. Lactic acid bacteria in meat fermentation. FEMS Microbiol Rev. 87(1–2):165–174.
Hammes WP, Hertel C. 1998. New developments in meat starter cultures. Meat Sci. 49:S125–S138.
Hammes WP, Knauf HJ. 1994. Starters in the processing of meat products. Meat Sci. 36(1–2):155–168.

Hebert EM, et al. 2012. Genome sequence of the bacteriocin-producing *Lactobacillus curvatus* strain CRL705. J Bacteriol. 194(2):538–539.
Inglin RC, Meile L, Stevens MJA. 2017. Draft genome sequences of 43 *Lactobacillus* strains from the species *L. curvatus*, *L. fermentum*, *L. paracasei*, *L. plantarum*, *L. rhamnosus*, and *L. sakei*, isolated from food products. Genome Announc. 5(30):e00632-17–e00617.
Jans C, Lagler S, Lacroix C, Meile L, Stevens MJA. 2017. Complete genome sequences of *Lactobacillus curvatus* KG6, L. *curvatus* MRS6, and *Lactobacillus sakei* FAM18311, isolated from fermented meat products. Genome Announc. 5(38):e00915-17.
Jung JY, et al. 2011. Metagenomic analysis of kimchi, a traditional Korean fermented food. Appl Environ Microbiol. 77:2264–2274.
Koleva Z, et al. 2014. Lactic acid microflora of the gut of snail *Cornu aspersum*. Biotechnol Biotechnol Equip. 28(4):627–634.
Kask S, et al. 2003. Physiological properties of Lactobacillus paracasei, L. danicus and L. curvatus strains isolated from Estonian semi-hard cheese. Food Res Int. 36(9–10):1037–1046.
Lee SH, Jung MY, Song J-H, Lee M, Chang JY. 2017. Complete genome sequence of *Lactobacillus curvatus* strain WiKim38 isolated from Kimchi. Genome Announc. 5(18):e00273-17.
Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659.
Lucquin I, Zagorec M, Champomier-Verges M, Chaillou S. 2012. Fingerprint of lactic acid bacteria population in beef carpaccio is influenced by storage process and seasonal changes. Food Microbiol. 29(2):187–196.
Lyhs U, Björkroth JK. 2008. Lactobacillus sakei/curvatus is the prevailing lactic acid bacterium group in spoiled maatjes herring. Food Microbiol. 25(3):529–533.
Lyhs U, Korkeala H, Björkroth J. 2002. Identification of lactic acid bacteria from spoiled, vacuum-packaged 'gravad' rainbow trout using ribotyping. Int J Food Microbiol. 72(1–2):147–153.
Mahony J, van Sinderen D. 2014. Current taxonomy of phages infecting lactic acid bacteria. Front Microbiol. 5:7.
Marcó MB, Moineau S, Quiberoni A. 2012. Bacteriophages and dairy fermentations. Bacteriophage 2(3):149–158.
Marty-Teysset C, et al. 1996. Proton motive force generation by citrolactic fermentation in *Leuconostoc mesenteroides*. J Bacteriol. 178(8):2178–2185.
Michel E, et al. 2016. Characterization of relative abundance of lactic acid bacteria species in French organic sourdough by cultural, qPCR and MiSeq high-throughput sequencing methods. Int J Food Microbiol. 239:35–43.
Nakano K, et al. 2016. First complete genome sequence of the skin-improving *Lactobacillus curvatus* strain FBA2, isolated from fermented vegetables, determined by PacBio single-molecule real-time technology. Genome Announc. 4(5):e00884-16.
Nobelmann B, Lengeler JW. 1996. Molecular analysis of the gat genes from *Escherichia coli* and of their roles in galactitol transport and metabolism. J Bacteriol. 178(23):6790–6795.
Plumbridge J, Pellegrini O. 2004. Expression of the chitobiose operon of *Escherichia coli* is regulated by three transcription factors: nagC, ChbR and CAP. Mol Microbiol. 52(2):437–449.
Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155(2):945–959.
R Development Core Team. 2010. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Sampaio M-M, et al. 2004. Phosphotransferase-mediated transport of the osmolyte 2-O-alpha-mannosyl-D-glycerate in Escherichia coli occurs by the product of the mngA (hrsA) gene and is regulated by the mngR (farR) gene product acting as repressor. J Biol Chem. 279(7):5537–5548.
Schmid M, et al. 2018. Comparative genomics of completely sequenced *Lactobacillus helveticus* genomes provides insights into strain-specific genes and resolves metagenomics data down to the strain level. Front Microbiol. 9:63.
Siezen R, et al. 2006. *Lactobacillus plantarum* gene clusters encoding putative cell-surface protein complexes for carbohydrate utilization are conserved in specific gram-positive bacteria. BMC Genomics 7:126.
Stentz R, Zagorec M. 1999. Ribose utilization in *Lactobacillus sakei*: analysis of the regulation of the rbs operon and putative involvement of a new transporter. J Mol Microbiol Biotechnol. 1:165–173.
Sun Z, et al. 2015. Expanding the biotechnology potential of Lactobacilli through comparative genomics of 213 strains and associated genera. Nat Commun. 6(1):8322.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 30(12):2725–2729.
Terán LC, Coeuret G, Raya R, Champomier-Vergès M-C, Chaillou S. 2017. Draft genome sequence of *Lactobacillus curvatus* FLEC03, a meat-borne isolate from beef carpaccio packaged in a modified atmosphere. Genome Announc. 5(26):e00584-17.
Tohno M, Kobayashi H, Nomura M, Uegaki R, Cai Y. 2012. Identification and characterization of lactic acid bacteria isolated from mixed pasture of timothy and orchardgrass, and its badly preserved silage. Anim Sci J. 83(4):318–330.
Vallenet D, et al. 2013. MicroScope–an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. Nucleic Acids Res. 41(D1):D636–D647.
Vogel RF, Lohmann M, Nguyen M, Weller AN, Hammes WP. 1993. Molecular characterization of Lactobacillus curvatus and Lact. sake isolated from sauerkraut and their application in sausage fermentations. J Appl Bacteriol. 74(3):295–300.
Vogler AP, Lengeler JW. 1989. Analysis of the nag regulon from *Escherichia coli* K12 and *Klebsiella pneumoniae* and of its regulation. Mol Gen Genet. 219(1–2):97–105.
Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 3(2):199–208.
Wiame E, Lamosa P, Santos H, Van Schaftingen E. 2005. Identification of glucoselysine-6-phosphate deglycase, an enzyme involved in the metabolism of the fructation product glucoselysine. Biochem J. 392(2):263–269.
Yew WS, Gerlt JA. 2002. Utilization of L-ascorbate by *Escherichia coli* K-12: assignments of functions to products of the yjf-sga and yia-sgb operons. J Bacteriol. 184(1):302–306.
Zheng J, Ruan L, Sun M, Gänzle M. 2015. A genomic view of Lactobacilli and Pediococci demonstrates that phylogeny matches ecology and physiology. Appl Environ Microbiol. 81(20):7233–7243.
Zhou Y, Drouin P, Lafrenière C. 2016. Effect of temperature (5-25°C) on epiphytic lactic acid bacteria populations and fermentation of whole-plant corn silage. J Appl Microbiol. 121(3):657–671.
Zommiti M, Connil N, Hamida JB, Ferchichi M. 2017. Probiotic characteristics of *Lactobacillus curvatus* DN317, a strain isolated from chicken ceca. Probiotics Antimicrob Proteins. 9(4):415–424.

**Associate editor**: Esther Angert