MICROBIAL GENOMICS

RESEARCH ARTICLE
Wuyts *et al.*, *Microbial Genomics* 2019;5
DOI 10.1099/mgen.0.000286

MICROBIOLOGY SOCIETY

OPEN DATA OPEN MICROBIOLOGY

Comparative genome analysis of *Lactobacillus mudanjiangensis*, an understudied member of the *Lactobacillus plantarum* group

Sander Wuyts[1,2], Camille Nina Allonsius[1], Stijn Wittouck[1], Sofie Thys[3], Bart Lievens[4], Stefan Weckx[2], Luc De Vuyst[2] and Lebeer Sarah[1,*]

Abstract

The genus *Lactobacillus* is known to be extremely diverse and consists of different phylogenetic groups that show a diversity that is roughly equal to the expected diversity of a typical bacterial genus. One of the most prominent phylogenetic groups within this genus is the *Lactobacillus plantarum* group, which contains the understudied *Lactobacillus mudanjiangensis* species. Before this study, only one *L. mudanjiangensis* strain, DSM 28402^T, had been described, but without whole-genome analysis. In this study, three strains classified as *L. mudanjiangensis* were isolated from three different carrot juice fermentations and their whole-genome sequence was determined, together with the genome sequence of the type strain. The genomes of all four strains were compared with publicly available *L. plantarum* group genome sequences. This analysis showed that *L. mudanjiangensis* harboured the second largest genome size and gene count of the whole *L. plantarum* group. In addition, all members of this species showed the presence of a gene coding for a cellulose-degrading enzyme. Finally, three of the four *L. mudanjiangensis* strains studied showed the presence of pili on scanning electron microscopy (SEM) images, which were linked to conjugative gene regions, coded on a plasmid in at least two of the strains studied.

DATA SUMMARY

(1) The sequencing data and genome assemblies are available at the European Nucleotide Archive under accession number PRJEB29655 (https://www.ebi.ac.uk/ena/data/view/PRJEB29655).

(2) The complete data analysis pipeline can be found on GitHub (https://github.com/swuyts/mudAnalysis).

INTRODUCTION

The genus *Lactobacillus* is known to be extremely diverse [1]. Furthermore, it has been shown that different phylogenetic groups within this genus display a diversity that is roughly equal to the expected diversity of a typical bacterial genus [2–6]. Each of these phylogenetic groups can be recognized as an entity with unique properties and a distinct natural history, ecology, function and physiology [5]. Therefore, the study of these phylogenetic groups separately, as if they were one genus, can be an interesting approach that might reveal new, previously overlooked, phylogenetic relationships and functional properties.

One of the more abundantly studied species within the genus *Lactobacillus* is *Lactobacillus plantarum*. Previous genome-based phylogenetic studies have defined *L. plantarum* as a member of the *L. plantarum* group, together with *Lactobacillus fabifermentans*, *Lactobacillus paraplantarum*, *Lactobacillus pentosus* and *Lactobacillus xiangfangensis* [1, 7]. In addition, the species *Lactobacillus herbarum* [8], *Lactobacillus*

Received 21 March 2019; Accepted 16 July 2019; Published 01 August 2019

Author affiliations: [1]Research Group Environmental Ecology and Applied Microbiology (ENdEMIC), Department of Bioscience Engineering, University of Antwerp, Antwerp, Belgium; [2]Research Group of Industrial Microbiology and Food Biotechnology (IMDO), Faculty of Sciences and Bioengineering Sciences, Vrije Universiteit Brussel, Brussels, Belgium; [3]Laboratory of Cell Biology and Histology, Antwerp Centre for Advanced Microscopy (ACAM), University of Antwerp, Antwerp, Belgium; [4]Laboratory for Process Microbial Ecology and Bioinspirational Management (PME&BIM), Department of Microbial and Molecular Systems (M2S), KU Leuven, Campus De Nayer, Sint-Katelijne-Waver, Belgium.

***Correspondence:** Lebeer Sarah, sarah.lebeer@uantwerpen.be

Keywords: *Lactobacillus*; comparative genomics; conjugation; scanning electron microscopy; cellulase.

Abbreviations: c.f.u., colony forming units; CMC, carboxymethyl cellulose; GH, glycosyl hydrolase; GT, glycosyltransferase; ICEs, integrative and conjugative elements; LAB, lactic acid bacteria; LGC, Lactobacillus genus complex; MRS, de Man-Rogosa-Sharpe; NCBI, National Center for Biotechnology Information; PBS, phosphate-buffered-saline; SEM, scanning electron microscopy; T4CP, type 4 coupling protein; T4SS, type 4 secretion system.

Sequencing data and genome assemblies are available at the European Nucleotide Archive under the accession number ERP111972.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Three supplementary figures and four supplementary tables are available with the online version of this article.

plajomi [9], *Lactobacillus modestisalitolerans* [9] and *Lactobacillus mudanjiangensis* [10] are closely related to *L. plantarum* and thus should be regarded as members of the *L. plantarum* group. *L. mudanjiangensis* is a species that was described for the first time in 2013 and that was isolated from a traditional pickle fermentation in the Heilongjiang province in China [10]. Since its first description, no other study has provided additional characterization or reported the isolation of other strains of the *L. mudanjiangensis* species. Therefore, before this study, not a single genomic assembly of this species was publicly available. However, in this study, four strains isolated from three different spontaneous carrot juice fermentations [11] were putatively classified as members of this *Lactobacillus* species.

Since the discovery of the mucus-binding pili, fimbriae or adhesins, in *Lactobacillus rhamnosus* GG [12, 13], several comparative genomic studies have focused on exploring similar gene clusters in other lactobacilli, including the members of the *L. plantarum* group [1, 14–18]. Whereas these specific pili play an important role in cell surface adhesion, pili can be of importance for an array of other functions as well, ranging from biofilm formation to the uptake of extracellular DNA via natural competence (type IV pili) or facilitation of DNA transfer via conjugation [19–21]. The latter is a process that uses conjugative pili to bring bacterial cells together to provide an interface to exchange macromolecules, such as DNA or DNA–protein complexes [20]. Historically, these conjugation systems and their pili have only been associated with conjugative plasmids [22], one of the main drivers of horizontal gene transfer [23, 24]. However, recently, integrative and conjugative elements (ICEs), which harbour conjugation systems as well, have also been found to be another important driver of horizontal gene transfer [24–26].

This study aimed to provide more insights into the genomic features of the understudied *L. mudanjiangensis* species, in relation to the other members of the *L. plantarum* group, using a comparative genomics approach. Therefore, the genome of the type strain of *L. mudanjiangensis* was sequenced together with three strains isolated from fermented carrot juice. These and other publicly available genome sequences were used to screen for *L. mudanjiangensis* species-specific properties, which included an analysis for the presence of genes related to pili formation and conjugation. In total, 304 genomes were subjected to an in-depth analysis focusing on the phylogenetic relationships as well as the predicted functional capacity of these strains.

METHODS

Sequencing of the *L. mudanjiangensis* type strain and downloading of publicly available assemblies

The type strain of *L. mudanjiangensis* [*L. mudanjiangensis* DSM 28402^T (=LMG 27194^T=CCUG 62991^T)] was purchased from a public micro-organism collection (BCCM-LMG, Ghent, Belgium). This strain and *L. mudanjiangensis* AMBF197, AMBF209 and AMBF249 were grown overnight in de Man–Rogosa–Sharpe (MRS) medium (Carl Roth, Karlsruhe, Germany) and DNA was extracted using the NucleoSpin 96 tissue kit (Macherey-Nagel, Düren, Germany), with an extra cell lysis step using 20 mg ml^{-1} of lysozyme (Sigma-Aldrich, St Louis, MO, USA) and 100 U ml^{-1} of mutanolysin (Sigma-Aldrich). Whole-genome sequencing was performed using the Nextera XT DNA Sample Preparation kit (Illumina, San Diego, CA, USA) and the Illumina MiSeq platform, using 2×250 cycles, at the Laboratory of Medical Microbiology (University of Antwerp, Antwerp, Belgium) in the case of the strains AMBF197, AMBF249 and DSM28402^T or 2×300 cycles at the Center of Medical Genetics Antwerp (University of Antwerp) for strain AMBF209. *De novo* assembly of the genome sequence was performed using SPAdes v 3.12.0 [27]. In addition, all genome sequences annotated as *L. fabifermentans*, *L. herbarum*, *L. paraplantarum*, *L. pentosus*, *L. plantarum* and *L. xiangfangensis* were downloaded from the National Center for Biotechnology Information (NCBI) Assembly database on 24 July 2018 using in-house scripts. In total, 310 genomes were used as an input for quality control.

Impact Statement

Members of the bacterial genus *Lactobacillus* are well known because of their use in foods and probiotics. Currently more than 200 species have been described as members of this genus and every year several new species are discovered. One of them is *Lactobacillus mudanjiangensis*, first described in 2013. Since its first description, no other study has provided additional characterization or reported the isolation of other strains of the *L. mudanjiangensis* species. In this study, three new strains, isolated from fermented carrot juice, are reported and the first comparative genome analysis of this species is presented. This resulted in the discovery of a cellulose-degrading enzyme, which has not been found in any other *Lactobacillus* species, and which could be useful for several industrial applications wherein breakdown of this important skeletal plant component is necessary. Furthermore, scanning electron microscopy detected the presence of pili, which were linked to bacterial conjugation, a process in which DNA is transferred from one bacterial cell to another. In general, this genome-based study of *L. mudanjiangensis* thus provides the first insights into the biology of this species, which could lead to novel applications.

Quality control and annotation

Basic genome characteristics, including genome size, GC content, completeness and N50 value, were estimated using Quast 4.6.3 [28] and CheckM v1.0.12 [29]. The quality of the genome assemblies was evaluated using the Quast and CheckM output. After visualization of several quality control parameters using ggplot2 [30], genomes with an N50 value <25000 bp, a number of undefined nucleotides (N) per 100000 bases >500 and a completeness <94% were discarded. Furthermore, one *L. plantarum* genome with an extremely low total genome length (GCA_001660655) was also discarded. An overview of all the genome sequences and strains that passed this quality control (304 assemblies) can be found in Table S1 (available in the online version of this article). Finally, Prokka 1.12 [31] was used to predict and annotate genes for all genome sequences, including an estimation of the number of tRNA and rRNA sequences. In addition to its internal databases, a customized genus-specific BLAST database was used for higher quality annotation with Prokka's *–usegenus* option. This database was created using BLAST [32, 33] and all complete *Lactobacillus* genomes found in the NCBI Assembly database.

Defining the pangenomes of all *L. plantarum* group species

To define the pangenome, all genes were clustered into orthogroups using OrthoFinder 2.2.6 [34] and further analysed in R [35]. Here, a core orthogroup is defined as an orthogroup present in more than 95% of a set of genomes. All other orthogroups are defined as accessory orthogroups. An upset plot was created using the R package UpSetR [36]. Unique orthogroups belonging to *L. mudanjiangensis* were further annotated using EggNOG-mapper [37] and visualized using ggplot2 [30].

Phylogenetic tree construction

Single-copy core orthogroups found by Orthofinder were used as input for the construction of a phylogenetic tree. *Lactobacillus algidus* DSM 15638 (NCBI Assembly accession number GCA_001434695) served as an outgroup, as it is the species most closely related to the *L. plantarum* group [1]. The first protein sequence of each fasta file of the single-copy core orthogroups was compared with a BLAST database of all genome proteins of the outgroup's genome sequence. All hits with a coverage >75% and a percentage similarity >50% were added to the alignment of each orthogroup. These alignments, on the amino acid level, were concatenated into a supermatrix that was used in RaxML 8.2.9 [38] to build a maximum-likelihood phylogenetic tree with the *–f a* option, which combines a rapid bootstrap algorithm with an extensive search of the tree space, starting from multiple different starting trees. The tree and subtrees were plotted with the R package ggtree [39].

Average nucleotide identity

All pairwise ANI values were calculated with the Python pyani package [40] using a BLASTN approach [32, 33] based on the methodology described by Goris *et al.* [41].

Characterization of a cellulase-encoding gene

The predicted cellulase-encoding genes were further characterized by performing a mafft [42] alignment on the nucleotide and protein level to assess their similarity. Furthermore a BLASTP search against the NCBI protein (nr) database was performed [32, 33]. Only hits with a query coverage >80% and a percentage identity >60% were retained. A conserved domain analysis was performed using the NCBI Conserved Domain web interface [43]. Genes encoding glycosyl hydrolases (GHs) were detected by scanning all *L. mudanjiangensis* genomes and all the *Lactobacillus* genus complex (LGC) as described by Wittouck *et al.* [44] (genome sequences downloaded 18 December 2018) against hidden Markov model (HMM) profiles of the CAZyme families [45]. The profiles were downloaded from the dbCAN webserver [46] and queried using HMMSCAN [47]. An E-value of 1×10^{-15} and a coverage of 0.35 were used as a cut-off, similar to what has been described before [48].

Experimental validation of the cellulase activity was performed by growing the bacterial strains *L. mudanjiangensis* AMBF197, AMBF209, AMBF249 and DSM28402^{T} and *L. plantarum* CMPG5300 overnight in MRS medium with the addition of 0.5% carboxymethyl cellulose (CMC) at 37 °C. The overnight cultures were centrifuged for 10 min at 4000 ***g***, the supernatant was removed and the cells were resuspended in 8 ml phosphate-buffered saline (PBS) medium. Next, the cells were centrifuged again, followed by the removal of supernatant and finally resuspended in 80 µl PBS [56 g of NaCl, 1.4 g of KCl, 10.48 g of Na_2HPO_4 and 1.68 g of KH_2PO_4 (pH 7.4) l^{-1}]. Then 60 µl of each culture was spotted in the middle of a CMC agar medium plate (5 g of CMC, 1 g of $NaNO_3$, 1 g of K_2HPO_4, 1 g of KCl, 0.5 g of $MgSO_4$, 0.5 g of yeast extract and 15 g of agar l^{-1}) followed by a 3-day incubation at 37 °C. Finally, to visualize halo formation, the agar plates were stained for 30 min with 3 ml of a 0.1% Congo red solution and destained with 3 ml of 1M NaCl twice for 15 min.

Scanning electron microscopy (SEM)

To assess the presence or absence of pili or fimbriae on the cell surface of *L. mudanjiangensis* strains AMBF197, AMBF209, AMBF249 and DSM 28402^{T}, SEM was performed. To this end, the bacterial strains were grown overnight (MRS medium, 37 °C), gently washed with PBS and spotted on a gold-coated membrane [approximately 5×10^{7} colony-forming units (c.f.u.) per membrane]. Bacterial spots were fixed with 2.5% (m/v) glutaraldehyde in 0.1 M sodium cacodylate buffer (2.5% glutaraldehyde, 0.1 M sodium cacodylate, 0.05% $CaCl_2.2H_2O$; pH 7.4) by gently shaking the membrane for 1 h at room temperature, followed by a further overnight fixation at 4 °C. After fixation, the membranes were washed three times for 20 min with cacodylate buffer [containing 7.5% (m/v) saccharose]. Subsequently, the bacteria were dehydrated in an ascending series of ethanol (50, 70, 90 and 95%, each for 30 min at room temperature, and 100% for 2×1 h and 1×30 min) and dried in a Leica EM CPD030 (Leica Microsystems Belgium, Diegem, Belgium). The membranes were mounted on a stub and coated with 5 nm of carbon (Leica Microsystems Belgium) in a Leica EM Ace 600 coater (Leica Microsystems Belgium). SEM imaging was performed using a Quanta FEG250 SEM system (Thermo Fisher, Asse, Belgium) at the Antwerp Centre for Advanced Microscopy (ACAM, University of Antwerp) and the Electron Microscopy for Material Science group (EMAT, University of Antwerp).

Detection of genomic clusters encoding pili or fimbriae

To screen for the presence of the *spaCBA* gene cluster, the gene cluster that is responsible for expression of the fimbriae in *L. rhamnosus* GG [12, 13], a BLAST search [32, 33] on the protein level was performed against a BLAST database constructed for each genome separately. The gene sequences of *spaA* (NCBI GenBank accession number BAI40953.1), *spaB* (BAI40954.1) and *spaC* (BAI40955.1) were used as queries. Furthermore, the genomes were screened for genes encoding pili-related protein secretion systems using the predicted amino acid sequences as a query and the TXSScan definitions and profile models [49] as references in MacSyFinder v1.0.5 [50]. As only genes related to conjugation systems were found, all protein sequences of all genomes were scanned again, this time using the CONJScan definitions and profile models [22, 24] using MacSyFinder. In brief, a conjugation region was only considered if the conjugation genes were separated by fewer than 31 genes, except for genes encoding relaxases that can be separated by maximal 60 genes. The region was considered to be conjugative when it contained genes coding for (i) a VirB4/TraU homologue, (ii) a relaxase, (iii) a type 4 coupling protein (T4CP) and (iv) a minimum number of type 4 secretion system (T4SS) type-specific genes [24]. For both scans, hits with alignments covering >50% of the protein profile and with an independent E-value $<10^{-3}$ were kept for further analysis (default parameters) in R [35]. Conserved domain analysis of genes of interest was performed using the NCBI Conserved Domain web interface [43]. The gene regions were visualized using the R package gggenes (available at https://github.com/wilkox/gggenes).

Plasmid identification

Detection and reconstruction of plasmids in the different *L. mudanjiangensis* strains was performed using Recycler v0.7 [51], with the original fastq files and SPAdes assembly graphs as input. In addition, plasmidSPAdes (SPAdes v3.12.0) [52] was used and only circular sequences detected with both assembly strategies were retained for further analysis. The assembled plasmids were annotated with Prokka and further characterized by scanning against the EggNOG database, as described above. The presence of a conjugation system was confirmed with CONJScan, as described above. The percentage identity between the different plasmids found was assessed using BLAST [32, 33]. The similarity with any previously described plasmid was checked by performing a Mash (v 2.1.1) [53] distance search against the PLSDB database with the PLSDB webserver (v 0.4.1–2) [54]. Only plasmids with a maximum distance of 0.05 were kept for further analysis. The presence of shared genes was assessed by performing a BLAST search [24] for all *L. mudanjiangensis* predicted plasmid gene sequences against the reference plasmid sequences. Only hits with a nucleotide identity >80% and a query coverage >70% were assessed as true hits. A plasmid map was created using Geneious v8 [32, 33].

Table 1. An overview of the studied species and strains

Public data			
Species	**No. of genomes**	**Type strain**	**Reference**
L. fabifermentans	2	DSM 21115^T	[80]
L. herbarum	1	DSM 100358^T	[8]
L. mudanjiangensis	0	DSM 28402^T	[10]
L. paraplantarum	5	DSM 10667^T	[81]
L. pentosus	13	DSM 20314^T	[82]
L. plantarum	278	DSM 20174^T	[83]
L. xiangfangensis	1	DSM 27103^T	[84]
Type strains sequenced in this study			
Species	**No. of genomes**	**Type strain**	**Reference**
L. mudanjiangensis	1	DSM 28402^T	[85]
In-house isolates			
Species	**No. of genomes**	**Isolation source**	**Reference**
L. mudanjiangensis	3	Spontaneously fermented carrot juice	[11]

Delimitation of integrative and conjugative elements

The presence of ICEs was explored by using a similar approach to the pipeline described previously [10]. Briefly, all strict core genes, i.e. genes present in all strains of *L. mudanjiangensis*, were found using the Orthofinder output (see above). Next, all flanking core genes of each conjugative region were identified. Since within one species an ICE is expected to be found between the same core orthogroups, the flanking core genes of each conjugative region found were evaluated to determine whether or not they could be defined as an ICE.

RESULTS

The assembled genome of the type strain *L. mudanjiangensis* DSM 28402^T was analysed together with the genome sequences of three putative *L. mudanjiangensis* strains isolated from carrot juice fermentations, namely AMBF197, AMBF209 and AMBF249, to confirm their putative classification as *L. mudanjiangensis* members. Furthermore, to allow comparison with other closely related *Lactobacillus* species and the detection of *L. mudanjiangensis* species-specific properties, all publicly available genome sequences (NCBI Assembly database, 24 July 2018) of *L. plantarum* group members were included in this comparative genomics study, and 304 genomes in total were analysed (Table 1).

Phylogeny of the *Lactobacillus plantarum* group

To obtain a detailed view on the phylogeny of *L. mudanjiangensis* in relation to the whole *L. plantarum* group, a maximum-likelihood phylogenetic tree was constructed, based on 612 single-copy core orthogroups, found with Orthofinder (Figs 1 and S1). The resulting topology of this tree showed seven major clades, mostly following the species annotation as described in the NCBI Assembly database. However, these results exposed a few wrongly annotated genomic assemblies. For example, both *L. plantarum* MPL16 and *L. plantarum* AY01 had previously been annotated as *L. plantarum*, whereas here they were found within a clade that contained the *L. paraplantarum* type strain. Similarly, *L. plantarum* EGD-AQ4 was found within the clade of the *L. pentosus* type strain, whereas it was annotated as *L. plantarum* previously. These strains were reclassified for subsequent analysis. Furthermore, the type strain of *L. mudanjiangensis* formed a separate clade together with the strains AMBF197, AMBF209 and AMBF249 (Fig. 1). Based on its single-copy core orthogroups, this species was phylogenetically the least similar to *L. plantarum*, whereas its most similar relative was *L. fabifermentans*, followed by *L. xiangfangensis*.

To confirm that each major phylogenetic clade represented at least one different species, the pairwise average nucleotide identity (ANI) values of all genome assemblies were calculated (Fig. 2). The intraclade ANI values all exceeded the commonly used 95% species-level threshold [55] for *L. mudanjiangensis* (99.0–99.4%), *L. fabifermentans* (99.7–99.9%) and *L. paraplantarum* (99.7–99.9%), whereas their interclade ANI values were far below this threshold, showing that these clades all represented a single species. However, this was not the case for *L. pentosus* and *L. plantarum*, for which multiple pairwise comparisons led to intraclade ANI values below this threshold, suggesting that these phylogenetic clades contained at least two species (Figs 2 and S2). Conversely, further analysis described elsewhere [44], showed that this was not the case and therefore none of the major clades were split for subsequent analysis. Finally, for *L. xiangfangensis* and *L. herbarum*, no intraclade comparisons could be performed, as only one genome assembly was available for these species.

Genomic features of *L. mudanjiangensis*

The above results confirmed that strains AMBF197, AMBF209 and AMBF249 were members of the *L. mudanjiangensis* species. Therefore, the first four genomes of this species are presented here. Their estimated genome size varied between 3.4 Mb (strain DSM 28402^T) and 3.6 Mb (strain AMBF209), whereas their GC content varied between 42.73% (strain AMBF209) and 43.06% (strain DSM 28402^T) (Table 2). Finally, a high number of transfer RNA (tRNA) genes were found in all four strains.

A substantial difference in total genome length between the different species of the *L. plantarum* group was found (Fig. 3a). *L. mudanjiangensis* showed a median estimated genome size of 3.53 Mb, making it the second largest of the whole *L. plantarum* group up to now. *L. pentosus* showed the largest median estimated genome size (3.74 Mb), followed by *L. mudanjiangensis* (3.53 Mb) and *L. fabifermentans* (3.43 Mb), whereas for *L. xiangfangensis* (3.0 Mb) and *L. herbarum* (2.9 Mb) it was much smaller. A high spread in genome length was found within strains belonging to the *L. plantarum* species, as their genome size ranged between 2.9 and 3.8 Mb. Furthermore, *L. mudanjiangensis* showed a GC content of 42.9%, the lowest value within the whole *L. plantarum* group (Fig. 3b). Finally, regarding median gene count, similar trends to those for the genome length were found, with *L. mudanjiangensis* showing the highest count, followed by *L. pentosus* and *L. fabifermentans*, whereas *L. xiangfangensis* and *L. herbarum* harboured the lowest numbers of genes (Fig. 3c).

In total, 947 588 genes were found in the whole *L. plantarum* group, with an average of 3110 genes per genome. These genes were further clustered into 8005 different orthogroups, leading to an average count of 2924 orthogroups per genome. The differences between these numbers were due to the fact that some genes were found in multiple copies within one genome, which clustered together in a single orthogroup. Of all these orthogroups, 2172 were defined as core orthogroups and 5833 as accessory orthogroups. A detailed overview of the number of genes and core and accessory orthogroups can be found in Table S2. Subsequently, the distribution of orthogroups between the different *L. plantarum* group members was explored (Fig. 3d). The species with the highest number of species-specific orthogroups was *L. plantarum*. With 2065 species-specific orthogroups, it greatly outnumbered all other species, although this number was most probably biased, due to the higher number of sequenced genomes available for *L. plantarum* compared with the other *L. plantarum* group species. It was followed by *L. mudanjiangensis* (Fig. 3d, blue), which contained 372 species-specific orthogroups, and *L. pentosus*, harbouring 279 species-specific orthogroups. Furthermore, *L. plantarum* and *L. pentosus* shared the highest number of uniquely shared orthogroups (426), followed by the combination of *L. plantarum* and *L. paraplantarum* (219 uniquely shared orthogroups), which seemed to be in line with the phylogeny described in Fig. 1. In contrast, *L. mudanjiangensis* shared more unique orthogroups with the phylogenetically distant *L. plantarum* (166 orthogroups; Fig. 3d, yellow) than it did with its most similar species, *L. fabifermentans* (59 orthogroups).

To obtain more insights into the unique properties of *L. mudanjiangensis*, all 372 species-specific orthogroups were further classified using the EggNOG database (inset Fig. 3d). However, this resulted in the vast majority of orthogroups (304) being classified under 'category S: function unknown', showing that further experimental work on functional gene validation is necessary. Otherwise, most orthogroups belonged to category G (carbohydrate transport and metabolism; 14 orthogroups), followed by category M (cell wall/membrane/envelope biogenesis; 13 orthogroups).

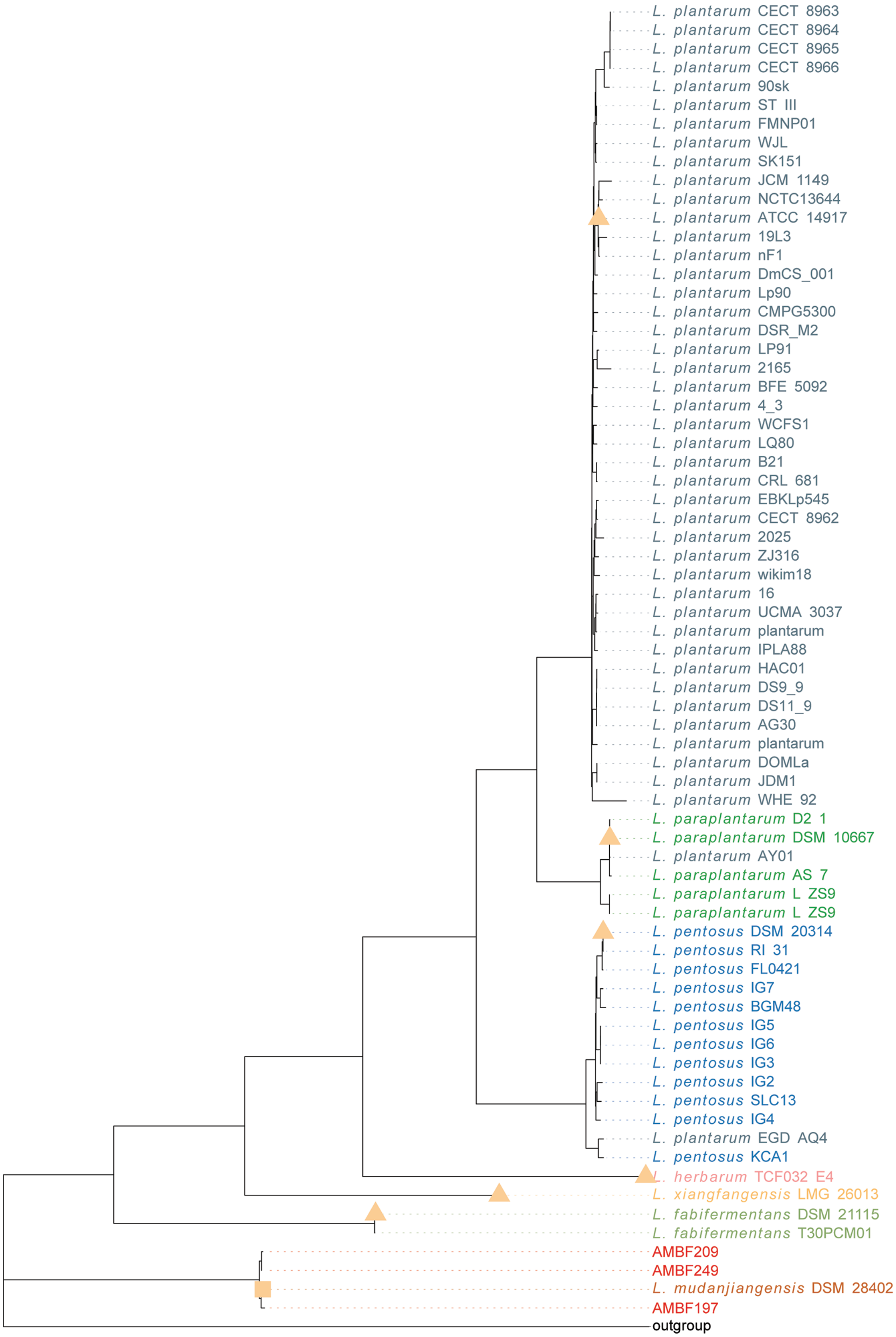

Fig. 1. Maximum-likelihood phylogenetic tree of the *L. plantarum* group. The tree is based on the amino acid sequences of 612 single-copy marker genes. *L. algidus* DSM 15638 was used as an outgroup. The tree was pruned to only keep 70 *L. plantarum* strains to avoid over-plotting. A complete tree can be found in Fig. S1. The branch length of the outgroup was shortened for better visualization. Each tip is colored based on its species name as annotated in the NCBI Assembly database (where applicable), with dark blue for *L. plantarum*, light green for *L. paraplantarum*, light blue for *L. pentosus*, pink for *L. herbarum*, light orange for *L. xiangfangensis*, dark orange for *L. mudanjiangensis* and red for the isolates obtained from the carrot juice fermentations. The type strains of each species are annotated with a triangle (NCBI) or a square (sequenced in this study).

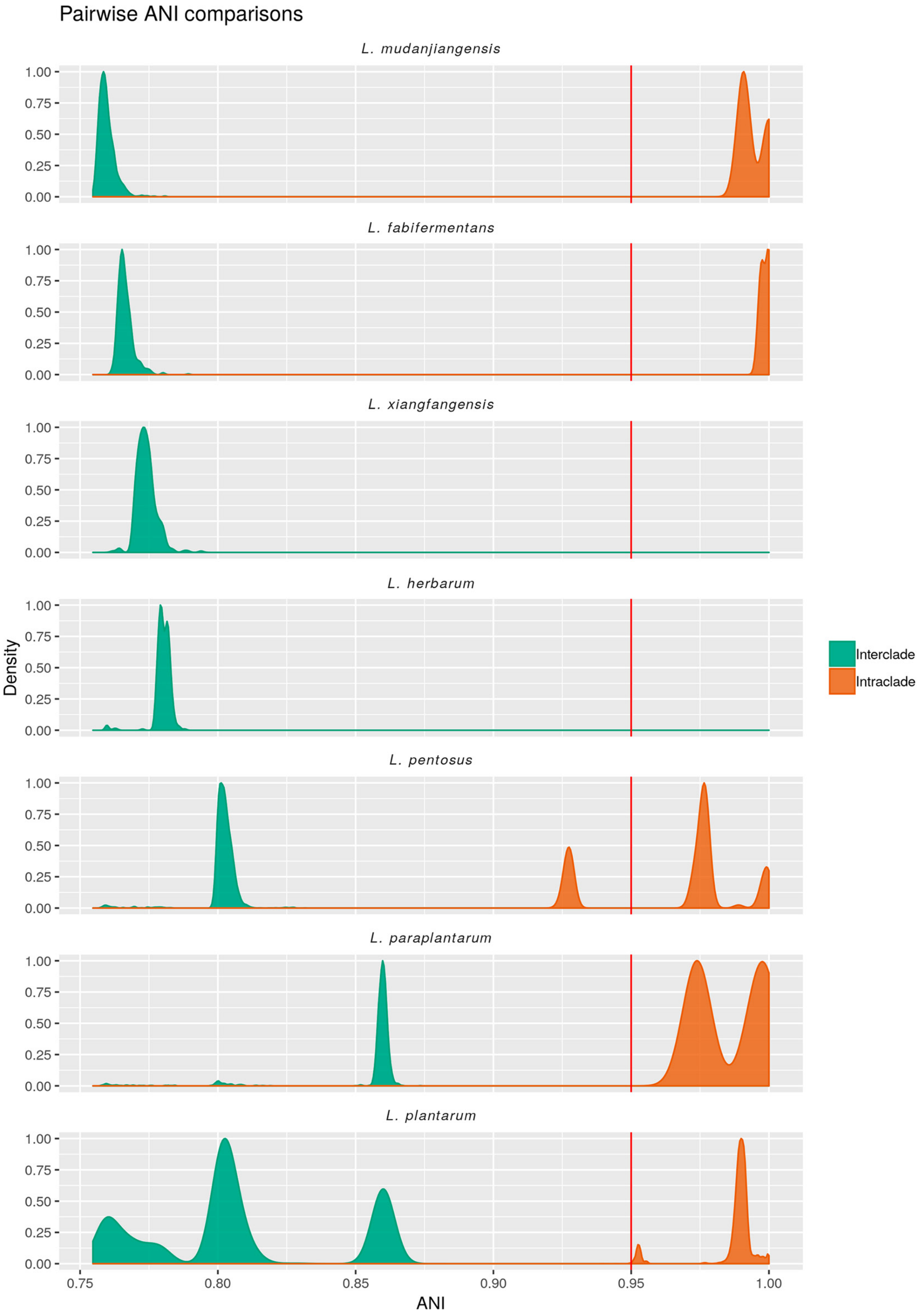

Fig. 2. Density plot of all pairwise average nucleotide identity (ANI) comparisons for each *L. plantarum* group species. All interclade comparisons are shown in green, whereas all intraclade comparisons are shown in orange. The vertical red line shows the commonly used 95% species-level delimitation threshold [55]. No intraclade comparisons could be performed for *L. xiangfangensis* and *L. herbarum*, as only one genome assembly was available for these species.

Table 2. Genome characteristics of *L. mudanjiangensis* strains

Genome	Total length (bp)	# contigs	GC content (%)	Coding sequences	# tRNA genes	# rRNA genes	Accession no.
AMBF197	3501388	52	42.85	3463	65	8	GCA_900617905
AMBF209	3589692	111	42.73	3586	63	8	GCA_900617925
AMBF249	3554025	69	42.83	3503	71	8	GCA_900617935
DSM 28402^T	3389962	42	43.06	3346	66	8	GCA_900617945

L. mudanjiangensis harbours a cellulose-degrading enzyme

Carbohydrate transport and metabolism (category G) was found to be the most abundantly characterized category among the *L. mudanjiangensis* species-specific orthogroups. Further examination of the 14 unique orthogroups that were detected in this category revealed the presence of a gene in all four strains annotated as endoglucanase E1, which is involved in the conversion of cellulose polymers into simple saccharides [56]. Alignment of both the nucleotide and predicted protein sequence (GenBank accession numbers: VDG21000, VDG22783, VDG26647 and VDG31879) showed that the sequences were identical between all four *L. mudanjiangensis* strains studied. Further analysis revealed the presence of a conserved domain in the endoglucanase E1 gene of all *L. mudanjiangensis* strains, annotated as cellulase/glycosyl hydrolase family 5 (NCBI Conserved Domains ID: pfam00150), supporting the view that this gene is related to cellulose degradation. Since cellulases/endoglucanases are thus classified as glycosyl hydrolases (GHs), GHs were predicted for all four *L. mudanjiangensis* strains and all LGC genomes as additional reference. Indeed, for all four *L. mudanjiangensis* strains, this endoglucanase E1 gene was classified as belonging to the GH5_1 family, a GH subfamily that was uniquely found in *L. mudanjiangensis* and not in any other member of the LGC. Although this GH5_1 family shows some degree of polyspecificity, the majority of enzymes (22 out of 24 enzymes characterized) are reported to be endoglucanases/cellulases [57]. Finally, a BLASTP search of the predicted protein sequences of these endoglucanase E1 genes to the NCBI nr database resulted in three additional hits to three predicted proteins of poorly characterized unclassified lactobacilli (GenBank accession numbers: WP_137634547, WP_137628413 and WP_137639555), which were not included in the above-described GH analysis, indicating that not only *L. mudanjiangensis* harboured putative cellulase genes within the LGC.

To confirm the *in silico* prediction of the putative cellulase-encoding genes, all four *L. mudanjiangensis* strains (AMBF197, AMBF209, AMBF249 and DSM28402^T) were grown on CMC agar, together with *L. plantarum* strain CMPG5300 as a negative control, as this strain lacks the endoglucanase E1 gene. Staining with Congo red, which binds to cellulose, revealed halo formation on plates containing *L. mudanjiangensis* strains AMBF197, AMBF209 and AMBF249 (Fig. 4). In contrast, no halo was formed on CMC agar plates containing *L. mudanjiangensis* DSM28402^T and *L. plantarum* CMPG5300. These results showed that while a putative cellulase-encoding gene was found in the genome sequences of all studied *L. mudanjiangensis* strains, unexpectedly, only three out of four strains showed cellulose-degrading activity on CMC agar plates. Together, these results thus pointed towards the presence of a novel cellulose-degrading enzyme in three *L. mudanjiangensis* strains.

Presence of a putative conjugative system in *L. mudanjiangensis*

To characterize potential novel cell surface macromolecules associated with this species, the cell surfaces of the four *L. mudanjiangensis* strains AMBF197, AMBF209, AMBF249 and DSM28402^T were visualized using SEM. This analysis revealed that three of the four strains (*L. mudanjiangensis* DSM28402^T, AMBF209 and AMBF249) formed pili or fimbriae, connecting different cells to each other as well as cells to an undefined structure (Fig. 5a).

In order to identify the genes encoding these pili, all genome sequences of *L. mudanjiangensis* were screened for the presence of genes associated with these kinds of phenotypes. These included the *spaCBA* gene cluster, as well as secretion systems based on pili, such as the type II and type IV secretion systems [12, 13, 49]. In this study, no *spaCBA* gene cluster was found. However, further exploration revealed the presence of a conjugation system in at least three of the four *L. mudanjiangensis* strains examined (AMBF209, AMBF249 and DSM 28402^T).

In general, a conjugation system consists of three major components: (i) a relaxase (Fig. 5c, green) that will bind and nick the DNA at the origin of replication and (ii) a coupling protein (T4CP; Fig. 5c, orange) that will couple the relaxase–DNA complex to (iii) a type IV secretion system (T4SS; Fig. 5c, blue), which ultimately transfers the whole complex to the recipient cell [22, 23, 49, 58]. Two complete conjugation systems containing all three mandatory parts (Fig. 5c) were found in *L. mudanjiangensis* AMBF209 and AMBF249, whereas one complete conjugation system was found in *L. mudanjiangensis* DSM 28402^T (Fig. 5b). All components of these conjugation systems were further classified based on the classification presented in Guglielmini *et al.* [58]. For all three *L. mudanjiangensis* strains, the relaxase gene (Fig. 5b, c, green) of this conjugation system was classified as a member of the MOB_Q class, whereas the coupling protein (Fig. 5b, c, orange)

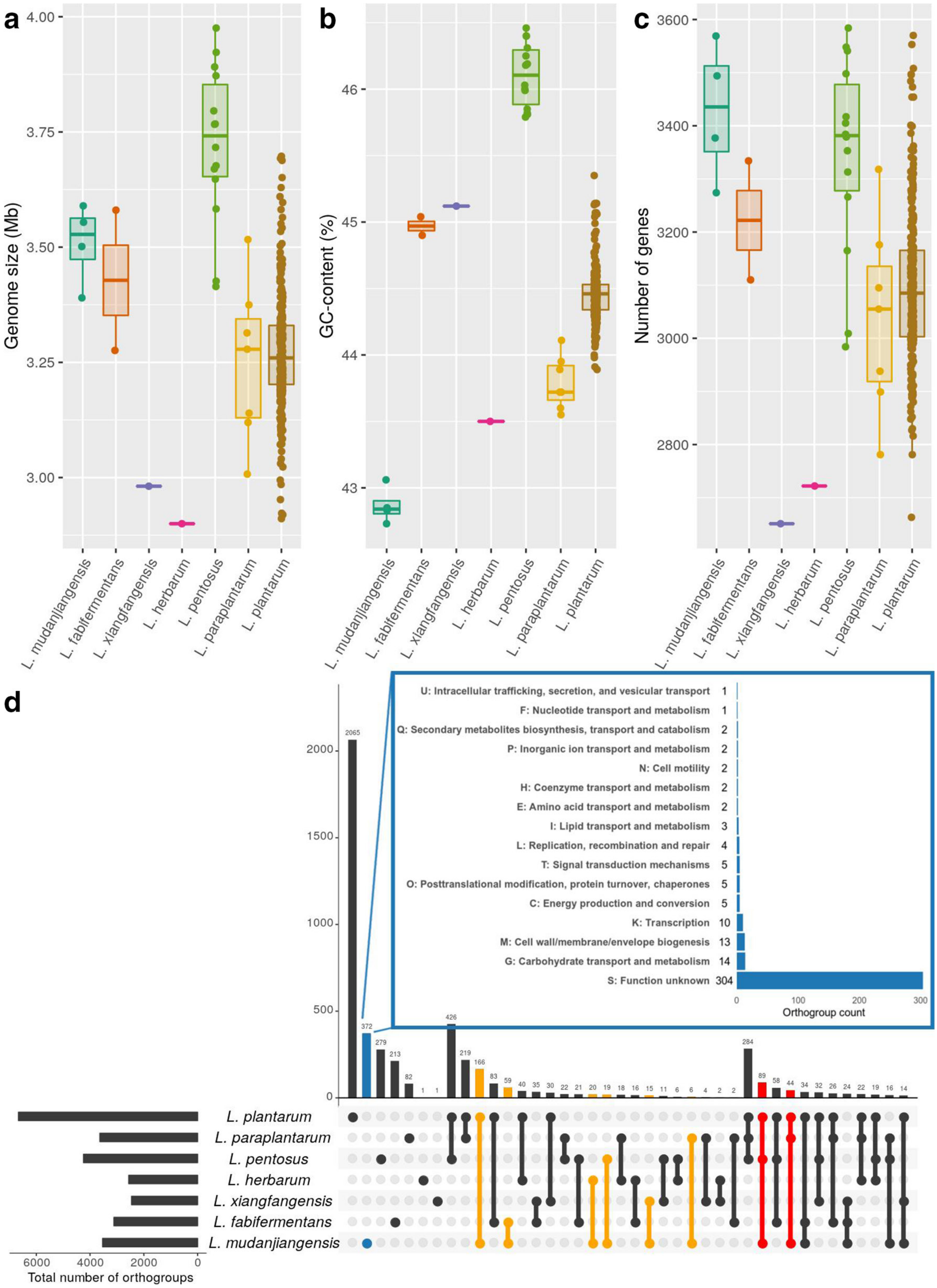

Fig. 3. Estimated genome sizes, GC content and gene counts for all genomes of the *L. plantarum* group species studied, and predicted functional capacity of all unique *L. mudanjiangensis* orthogroups. (a) Total genome size, (b) GC content and (c) gene counts for all genomes studied, coloured by species. (d) Upset plot comparing shared orthogroup counts between the eight *L. plantarum* group species. Species-specific orthogroups for *L. mudanjiangensis* are coloured in blue and the inset shows their functional category based on EggNOG classification. Uniquely shared orthogroups between *L. mudanjiangensis* and one other *L. plantarum* group member are coloured in orange, whereas uniquely shared orthogroups between *L. mudanjiangensis* and two other species are coloured in red.

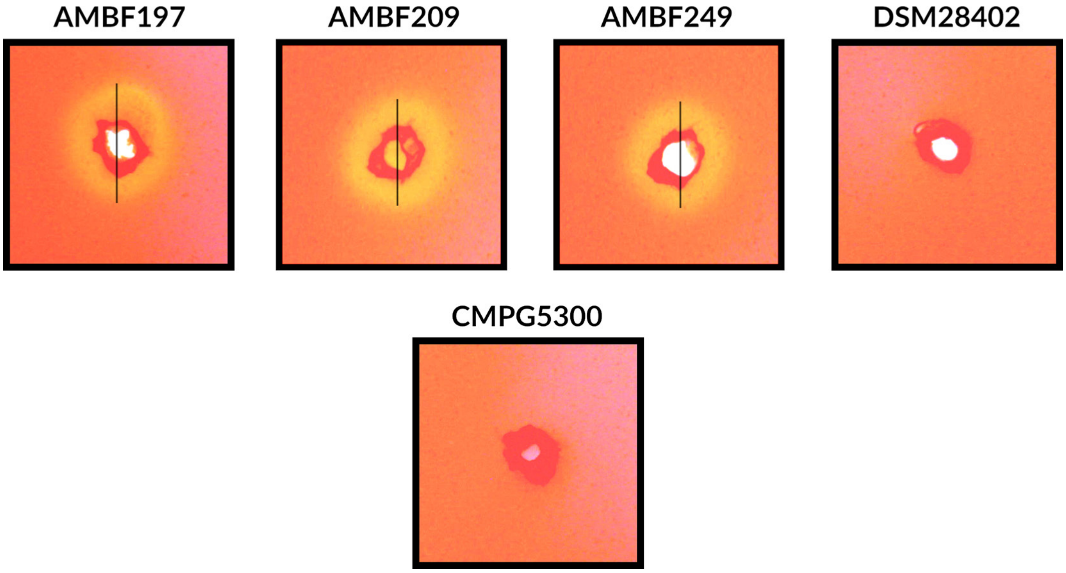

Fig. 4. Cellulose degradation of *L. mudanjiangensis*. *L. mudanjiangensis* AMBF197, AMBF209, AMBF249 and DSM28402^T were grown on carboxymethyl cellulose (CMC) agar and cellulose degradation was visualized by Congo red staining. *L. plantarum* CMPG5300 was used as a negative control.

was classified as T4CP2. The T4SS system, which harboured the genes possibly related to the observed pilus, was further classified as belonging to the class MPF$_{FATA}$, which groups the conjugation-related T4SS systems of Gram-positive bacteria [58]. The ATPase motor of this T4SS system was identified as a VirB4 homologue (first described in *Agrobacterium tumefaciens*). Furthermore, this T4SS system contained three accessory genes (*trsC*, *trsD* and *trsJ*) in *L. mudanjiangensis* AMBF209 and AMBF249, whereas four accessory genes were annotated in *L. mudanjiangensis* DSM 28402^T (*trsC*, *trsD*, *trsF* and *trsJ*) (Fig. 5b). For these accessory genes, homologues had already been identified previously for the genes *trsC* and *trsD*, with *trsC* coding for a VirB3 homologue, which is linked to the formation of the membrane pore, and *trsD* coding for another homologue of VirB4, the conjugation ATPase [58]. In contrast, both *trsF* and *trsJ* are poorly characterized.

Further analysis of the genes surrounding the annotated conjugation genes showed that this genomic region contained 18 to 19 open reading frames, most of them annotated as hypothetical proteins (Fig. 5b and Table S3). However, a conserved domain analysis revealed a bacteriophage peptidoglycan hydrolase domain in orthogroup OG0002812, annotated as a hypothetical protein, in both *L. mudanjiangensis* AMBF209 conjugation region 1 (AMBF209_CR1) and *L. mudanjiangensis* AMBF249 conjugation region 2 (AMBF249_CR2), making it a VirB1-like protein [58]. In *A. tumefaciens*, the VirB1 protein provides localized lysis of the peptidoglycan cell wall to allow insertion of the T4SS [59]. A similar domain, also known to harbour peptidoglycan lytic activity, was found in *L. mudanjiangensis* DSM28402_CR1 (orthogroup OG0002812). Finally, another conserved domain was found in all five gene regions clustered in orthogroup OG0003012, annotated as T4SS_CagC, which was shown to be a VirB2 homologue. VirB2 is the major pilus component of the type IV secretion system of *A. tumefaciens*, which is the main building block for extension and retraction of the conjugative pilus [60–62]. Taken together, these results showed the presence of pili in three *L. mudanjiangensis* strains (AMBF209, AMBF249 and DSM 28402^T), which after genomic analysis were hypothesized to be part of a conjugation system that is poorly characterized.

Finally, genome analysis of all other *L. plantarum* group members showed that the presence of a complete conjugation system was not unique to *L. mudanjiangensis* (Table S1). All 3 necessary genes were also found in 58 out of 275 *L. plantarum* strains, 2 out of 7 *L. paraplantarum* strains and 4 out of 14 *L. pentosus* strains. In contrast, the system was completely absent in *L. herbarum*, *L. xiangfangensis* and *L. fabifermentans*.

Plasmid reconstruction from genome data

Many conjugation systems are coded on plasmids [25]. Therefore, all four *L. mudanjiangensis* genomes were screened for plasmid presence. Plasmids were only found in two out of four genome assemblies, namely *L. mudanjiangensis* AMBF209 and AMBF249 (Fig. S3). Both strains harboured a plasmid of 27.3 Kb with 33 predicted genes, and after pairwise alignment of the plasmid they were found to be exactly the same. Subsequently, the presence of a conjugation system on these plasmids was confirmed using CONJScan. Further examination showed that the plasmid exactly matched the above-described AMBF209_CR2 and AMBF249_CR1 gene regions (Fig. 5b). Regarding annotated genes, 13 out of 33 gene products were predicted to be hypothetical proteins by Prokka. Further annotation using the EggNOG database revealed that most genes were mapped to category S (function

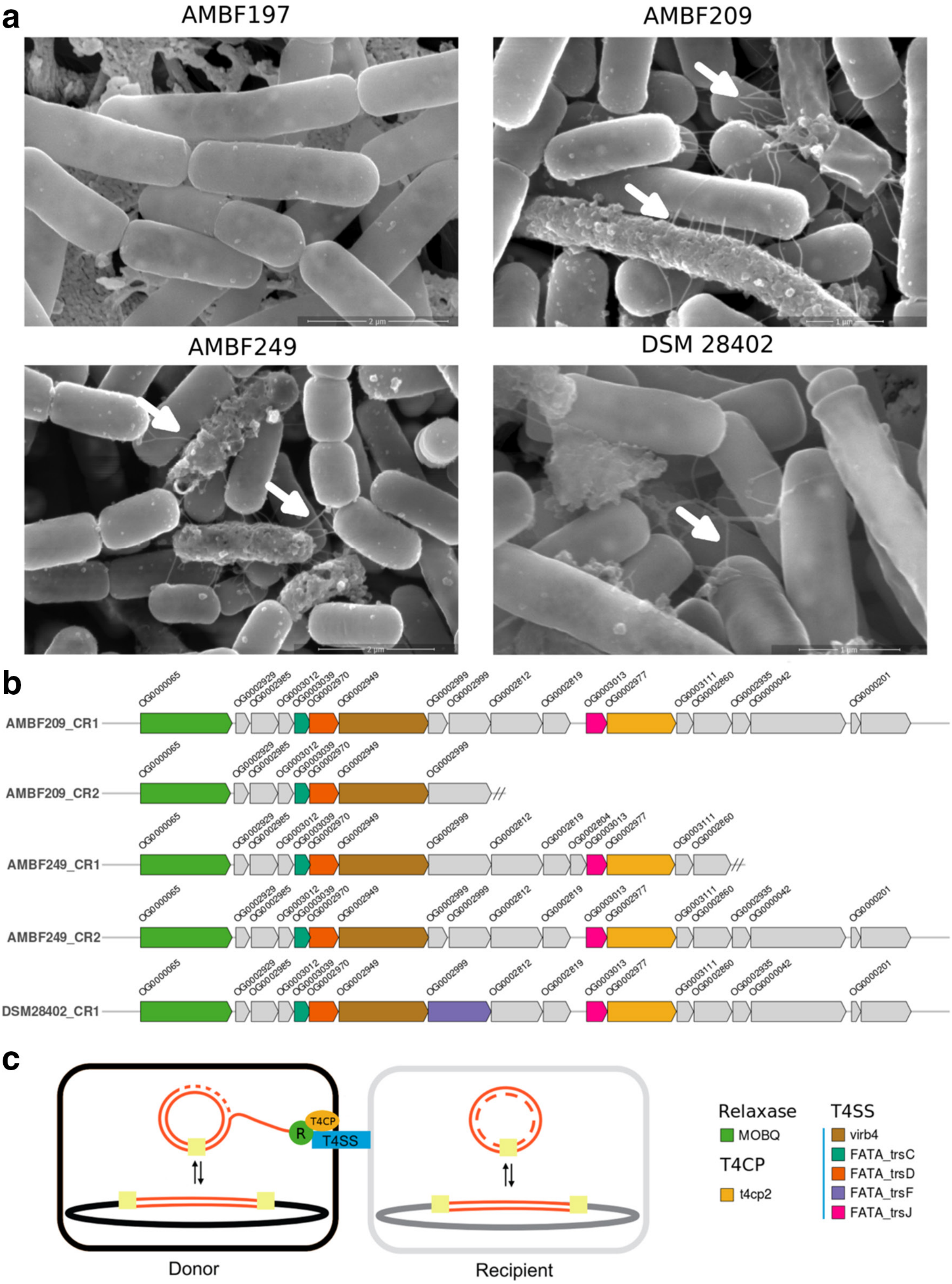
a
AMBF197
AMBF209
AMBF249
DSM 28402
2 µm
1 µm
2 µm
1 µm
b
AMBF209_CR1
OG0000065 OG0002929 OG0002985 OG0003012 OG0003039 OG0002970 OG0002949 OG0002999 OG0002999 OG0002812 OG0002819 OG0003013 OG0002977 OG0003111 OG0002860 OG0002935 OG0000042 OG0000201
AMBF209_CR2
OG0000065 OG0002929 OG0002985 OG0003012 OG0003039 OG0002970 OG0002949 OG0002999
AMBF249_CR1
OG0000065 OG0002929 OG0002985 OG0003012 OG0003039 OG0002970 OG0002949 OG0002999 OG0002812 OG0002819 OG0002804 OG0003013 OG0002977 OG0003111 OG0002860
AMBF249_CR2
OG0000065 OG0002929 OG0002985 OG0003012 OG0003039 OG0002970 OG0002949 OG0002999 OG0002999 OG0002812 OG0002819 OG0003013 OG0002977 OG0003111 OG0002860 OG0002935 OG0000042 OG0000201
DSM28402_CR1
OG0000065 OG0002929 OG0002985 OG0003012 OG0003039 OG0002970 OG0002949 OG0002999 OG0002812 OG0002819 OG0003013 OG0002977 OG0003111 OG0002860 OG0002935 OG0000042 OG0000201
c
T4CP
R
T4SS
Donor
Recipient
Relaxase
🟩 MOBQ
T4CP
🟧 t4cp2
T4SS
🟫 virb4
🟩 FATA_trsC
🟧 FATA_trsD
🟪 FATA_trsF
🟥 FATA_trsJ

Fig. 5. Scanning electron microscopy (SEM) and genes related to conjugation. (a) SEM images of all four *L. mudanjiangensis* strains studied. White arrows indicate putative conjugative pili. (b) Gene clusters encoding a putative conjugation system, coloured according to their potential function, as classified by CONJSCAN. The text above each gene shows its matching orthogroup. (c) Schematic model representing the process of bacterial conjugation with all three mandatory elements. The conjugation system is coloured based on the figure legend of (b). (R, relaxase; T4CP, type IV coupling protein; T4SS, type IV secretion system.) Adapted from [22].

unknown), followed by category L (replication, recombination and repair) and category C (energy production and conversion). Finally, a Mash search was performed against the PLSDB database to explore whether a similar plasmid was already described in the literature. Three plasmids showed a high similarity (Mash distance score <0.05) to the predicted *L. mudanjiangensis* plasmid, namely *Leuconostoc carnosum* MFPC16A2803 plasmid pMFPC16A2803B (GenBank accession number: LT991587), *L. carnosum* JB16 plasmid pKLC4 (NC_018699) and *Leuconostoc mesenteroides* SRCM103356 plasmid unnamed1 (NZ_CP035140). Out of 33 *L. mudanjiangensis* plasmid genes, 18 genes were shared with pMFPC16A2803B and unnamed1, while 21 genes were shared with pKLC4 (Table S4). All plasmid-borne conjugation genes (AMBF209_CR2 and AMBF249_CR1) were also detected in all three reference plasmids. Taken together, these results showed that *L. mudanjiangensis* strains AMBF209 and AMBF249 carried the same conjugative plasmid, for which the encoded gene functions are poorly characterized.

Since only two out of five conjugation regions (Fig. 5, AMBF209_CR2 and AMBF249_CR1) were plasmid-encoded, an additional analysis was performed to assess whether the other three conjugation systems could be part of an ICE. For this, all four *L. mudanjiangensis* genomes were analysed using a similar method to one that was recently described [24]. However, ICE regions usually contain repeats, such as transposases, leading to fragmentation of these ICE regions, if short-read sequencing technology is used [24]. Therefore, these analysis methods usually require a complete genome for proper ICE identification. The assembly state of the four genomes thus made it hard to correctly interpret the results obtained.

DISCUSSION

In this study, the genome sequence of the *L. mudanjiangensis* type strain DSM 28402^T was presented together with the genomes of three new *L. mudanjiangensis* strains, AMBF197, AMBF209 and AMBF249, which were isolated from three different spontaneous carrot juice fermentations [11]. To gain more insight into this understudied species, the genome sequences were compared with all publicly available genome sequences of the closely related species belonging to the *L. plantarum* group. This resulted in the discovery of a putative cellulose-degrading enzyme, annotated as endoglucanase E1, in all four *L. mudanjiangensis* strains and three uncharacterized *Lactobacillus* sp. genomes. To date, such an enzyme has never been found in any other LGC genome. Moreover, cellulose-degrading enzymes have never previously been found in lactic acid bacteria (LAB). Cellulose is the most abundant organic polymer on Earth, the most important skeletal component in plants in general [63] and the most abundant crude fibre in carrots [64]. From an industrial perspective, degradation of this polysaccharide is seen as being valuable for the production of high value-added products such as biofuels and lactic acid [65]. For this reason and because of their widespread industrial use, much effort has been devoted to the construction of recombinant cellulolytic LAB strains [65]. In particular, *L. plantarum* strains have been genetically modified to express cellulases due to their added benefit of high acid and ethanol tolerance [65–72]. This means that the discovery of a natural cellulolytic species, namely *L. mudanjiangensis*, which is a member of the *L. plantarum* group, may be valuable for industrial purposes. However, it should be noted that efficient hydrolysis of lignocellulose requires the cooperation of multiple enzymes [65] and therefore further characterization of the detected endoglucanase E1 is necessary. For example, as this enzyme was found in four *L. mudanjiangensis* strains but only three strains showed CMC degradation on CMC agar plates, future studies should focus on unravelling the expression conditions.

SEM analysis revealed the presence of pili or fimbriae in three of the four *L. mudanjiangensis* strains studied. In this study, the observation of pili in *L. mudanjiangensis* was associated with bacterial conjugation. Conjugation is one of the main drivers of horizontal gene transfer and is commonly associated with conjugative plasmids [23, 24]. Here, two of the five conjugative regions found were plasmid-associated and the two plasmids found were exactly the same for both *L. mudanjiangensis* AMBF209 and AMBF249, although these strains were isolated from different household carrot juice fermentations [11]. Previous studies also identified and described conjugative plasmids in other *Lactobacillus* species, such as *Lactobacillus brevis* [73], *Lactobacillus casei* [74], *Lactobacillus gasseri* [75], *Lactobacillus hokkaidonensis* [76], *L. plantarum* [77] and *Lactobacillus reuteri* [78]. Genes on these plasmids often code for proteins involved in detoxification, virulence, antibiotic resistance and ecological interactions [23], which could give them a fitness advantage in certain environments. Here, apart from the conjugation-related genes, many genes were annotated as hypothetical proteins on the conjugative plasmid, showing that the detected plasmid was poorly characterized. However, since this plasmid showed high similarity and a high number of shared genes with a plasmid from a *L. carnosum* strain, which was isolated from fermented kimchi [79], it could potentially harbour genes that are beneficial for survival on plants or in a fermented vegetable environment.

In conclusion, in this study, the genome sequences of four *L. mudanjiangensis* strains were studied in relation to the closely related members of the *L. plantarum* group. Comparative genome analysis showed that *L. mudanjiangensis* harboured one of the largest genomes and the highest gene counts of the *L. plantarum* group. Furthermore, a cellulose-degrading enzyme was detected in all four strains studied, but only three of these showed *in vitro* cellulase activity. Finally, three of the four *L. mudanjiangensis* strains studied showed the presence of pili on SEM images, and these were linked to conjugative gene regions. For two strains, *L. mudanjiangensis* AMBF209 and AMBF249, these regions were plasmid-associated. Further experimental studies, such as phenotypic growth curve-based screenings, conjugation experiments and the creation of knock-out mutants, are necessary to characterize the plasmid found and to confirm the link between the pili observed and this conjugation gene region.

Funding information

Some computational resources and services used in this work were provided by the Flemish Supercomputer Center (VSC), funded by the Research Foundation – Flanders (FWO) and the Flemish Government – department EWI. Additional computational work was performed on Rhambo and LAMBorGENEi. We acknowledge financial support from the Agency for Innovation by Science and Technology in Flanders (IWT-SB 141198) and PhD grants from the Research Foundation Flanders (FWO-SB 1S 035 16 N and FWO-Asp 11A0618N).

Acknowledgements

We thank Eline Oerlemans, Ilke De Boeck, Ines Tuyaerts and all other members of the ENdEMIC group for their general assistance and/or fruitful discussions. We also thank Charlotte Claes and Arvid Suls (Centre of Medical Genetics, University of Antwerp, Antwerp, Belgium) for their valuable input regarding whole-genome sequencing. Furthermore, we would like to thank the Antwerp Centre for Advanced Microscopy (ACAM, University of Antwerp) for processing and imaging of the SEM samples and the Electron Microscopy for Material Science group (EMAT, University of Antwerp) for the use of the environmental SEM (Quanta 250 FEG).

Conflicts of interest

The authors declare that there are no conflicts of interest.

Data bibliography

1. Wuyts S. European Nucleotide Archive. Study Accession number: PRJEB29655

References

1. Sun Z, Harris HMB, McCann A, Guo C, Argimón S *et al*. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nat Commun* 2015;6:8322.
2. Claesson MJ, van Sinderen D, O'Toole PW. *Lactobacillus* phylogenomics- towards a reclassification of the genus. *Int J Syst Evol Microbiol* 2008;58:2945–2954.
3. Salvetti E, Torriani S, Felis GE. The genus *Lactobacillus*: a taxonomic update. *Probiotics Antimicrob Proteins* 2012;4:217–226.
4. Salvetti E, Harris HMB, Felis GE, O'Toole PW. Comparative genomics reveals robust phylogroups in the genus *Lactobacillus* as the basis for reclassification. *Appl Environ Microb* 2018;84:AEM.:00993–18.
5. Duar RM, Lin XB, Zheng J, Martino ME, Grenier T *et al*. Lifestyles in transition: evolution and natural history of the genus *Lactobacillus*. *FEMS Microbiol Rev* 2017;41:S27–S48.
6. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A *et al*. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 2018;36:996–1004.
7. Zheng J, Ruan L, Sun M, Gänzle M. A genomic view of *Lactobacilli* and *Pediococci* demonstrates that phylogeny matches ecology and physiology. *Appl Environ Microbiol* 2015;81:7233–7243.
8. Mao Y, Chen M, Horvath P. *Lactobacillus herbarum* sp. nov., a species related to Lactobacillus plantarum. *Int J Syst Evol Microbiol* 2015;65:4682–4688.
9. Miyashita M, Yukphan P, Chaipitakchonlatarn W, Malimas T, Sugimoto M, Tanaka N, Tanasupawat S, Kamakura Y *et al*. Lactobacillus plajomi sp. nov. and Lactobacillus modestisalitolerans sp. nov., isolated from traditional fermented foods. *Int J Syst Evol Microbiol* 2015;65:2485–2490.
10. Gu CT, Li CY, Yang LJ, Huo GC. *Lactobacillus mudanjiangensis* sp. nov., *Lactobacillus songhuajiangensis* sp. nov. and *Lactobacillus nenjiangensis* sp. nov., isolated from Chinese traditional pickle and sourdough. *Int J Syst Evol Microbiol* 2013;63:4698–4706.
11. Wuyts S, Van Beeck W, Oerlemans EFM, Wittouck S, Claes IJJ *et al*. Carrot juice fermentations as man-made microbial ecosystems dominated by lactic acid bacteria. *Appl Environ Microbiol* 2018;84:e00134-18–18.
12. Lebeer S, Verhoeven TLA, Francius G, Schoofs G, Lambrichts I *et al*. Identification of a gene cluster for the biosynthesis of a long, Galactose-Rich exopolysaccharide in *Lactobacillus rhamnosus* GG and functional analysis of the priming glycosyltransferase. *Appl Environ Microbiol* 2009;75:3554–3563.
13. Kankainen M, Paulin L, Tynkkynen S, von Ossowski I, Reunanen J *et al*. Comparative genomic analysis of *Lactobacillus rhamnosus* GG reveals pili containing a human- mucus binding protein. *Proc Natl Acad Sci U S A* 2009;106:17193–17198.
14. Douillard FP, Ribbera A, Kant R, Pietilä TE, Järvinen HM *et al*. Comparative genomic and functional analysis of 100 *Lactobacillus rhamnosus* strains and their comparison with strain GG. *PLoS Genet* 2013;9:e1003683.
15. Douillard FP, Mora D, Eijlander RT, Wels M, de Vos WM. Comparative genomic analysis of the multispecies probiotic-marketed product VSL#3. *PLoS One* 2018;13:e0192452.
16. Yu X, Jaatinen A, Rintahaka J, Hynönen U, Lyytinen O *et al*. Human Gut-Commensalic *Lactobacillus ruminis* ATCC 25644 displays Sortase-Assembled surface piliation: phenotypic characterization of its fimbrial operon through in silico predictive analysis and recombinant expression in *Lactococcus lactis*. *PLoS One* 2015;10:e0145718–0145731.
17. Kant R, Palva A, von Ossowski I, Ossowski von I. An in silico pan-genomic probe for the molecular traits behind *Lactobacillus ruminis* gut autochthony. *PLoS One* 2017;12:e0175541.
18. Harris HMB, Bourin MJB, Claesson MJ, O'Toole PW. Phylogenomics and comparative genomics of *Lactobacillus salivarius*, a mammalian gut commensal. *Microb Genom* 2017;3.
19. Mandlik A, Swierczynski A, Das A, Ton-That H. Pili in Gram-positive bacteria: assembly, involvement in colonization and biofilm development. *Trends Microbiol* 2008;16:33–40.
20. Filloux A. A variety of bacterial pili involved in horizontal gene transfer. *J Bacteriol* 2010;192:3243–3245.
21. Muschiol S, Balaban M, Normark S, Henriques-Normark B. Uptake of extracellular DNA: competence induced pili in natural transformation of *Streptococcus pneumoniae*. *Bioessays* 2015;37:426–435.
22. Guglielmini J, Quintais L, Garcillán-Barcia MP, de la Cruz F, Rocha EPC. The repertoire of ice in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet* 2011;7:e1002222.
23. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, de la Cruz F. Mobility of plasmids. *Microbiol Mol Biol Rev* 2010;74:434–452.
24. Cury J, Touchon M, Rocha EPC. Integrative and conjugative elements and their hosts: composition, distribution and organization. *Nucleic Acids Res* 2017;45:8943–8956.
25. Johnson CM, Grossman AD. Integrative and conjugative elements (ICEs): what they do and how they work. *Annu Rev Genet* 2015;49:577–601.
26. Delavat F, Miyazaki R, Carraro N, Pradervand N, van der Meer JR. The hidden life of integrative and conjugative elements. *FEMS Microbiol Rev* 2017;41:512–537.
27. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
28. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.
29. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;25:1043–1055.
30. Wickham H. *ggplot2: Elegant Graphics for Data Analysis [Internet]*. New York: Springer-Verlag; 2009.

31. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.

32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.

33. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421.

34. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 2015;16:157.

35. R Core Team. *R: a Language and Environment for Statistical Computing [Internet]*. Austria: Vienna; 2015.

36. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 2017;33:2938–2940.

37. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol Biol Evol* 2017;34:2115–2122.

38. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.

39. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2017;8:28–36.

40. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Analytical Methods* 2016;8:12–24.

41. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P *et al*. Dna-Dna hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 2007;57:81–91.

42. Yamada KD, Tomii K, Katoh K. Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees. *Bioinformatics* 2016;32:3246–3251.

43. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ *et al*. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res* 2017;45:D200–D203.

44. Wittouck S, Wuyts S, Meehan CJ, van NV, Lebeer S. A genome-based species taxonomy of the *Lactobacillus* genus complex. *bioRxiv* 2019;537084.

45. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 2014;42:D490–D495.

46. Yin Y, Mao X, Yang J, Chen X, Mao F *et al*. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* 2012;40:W445–W451.

47. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 2011;39:W29–W37.

48. Zhang H, Yohe T, Huang L, Entwistle S, Wu P *et al*. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* 2018;46:W95–W101.

49. Abby SS, Cury J, Guglielmini J, Néron B, Touchon M *et al*. Identification of protein secretion systems in bacterial genomes. *Sci Rep* 2016;6:23080.

50. Abby SS, Néron B, Ménager H, Touchon M, Rocha EPC. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS One* 2014;9:e110726.

51. Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E *et al*. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* 2016;53:btw651.

52. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A *et al*. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics* 2016;151:btw493.

53. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH *et al*. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;17:132.

54. Galata V, Fehlmann T, Backes C, Keller A. PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res* 2019;47:D195–202.

55. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* 2009;106:19126–19131.

56. Yennamalli RM, Rader AJ, Kenny AJ, Wolt JD, Sen TZ. Endoglucanases: insights into thermostability for biofuel applications. *Biotechnol Biofuels* 2013;6:136.

57. Aspeborg H, Coutinho PM, Wang Y, Brumer H, Henrissat B. Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol Biol* 2012;12:186.

58. Guglielmini J, Néron B, Abby SS, Garcillán-Barcia MP, de la Cruz F, la Cruz Fde *et al*. Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res* 2014;42:5715–5727.

59. Zupan J, Hackworth CA, Aguilar J, Ward D, Zambryski P. VirB1* promotes T-Pilus formation in the vir-Type IV secretion system of *Agrobacterium tumefaciens*. *J Bacteriol* 2007;189:6551–6563.

60. Schröder G, Lanka E. The mating pair formation system of conjugative plasmids-A versatile secretion machinery for transfer of proteins and DNA. *Plasmid* 2005;54:1–25.

61. Alvarez-Martinez CE, Christie PJ. Biological diversity of prokaryotic type IV secretion systems. *Microbiol Mol Biol Rev* 2009;73:775–808.

62. Shariq M, Kumar N, Kumari R, Kumar A, Subbarao N *et al*. Biochemical analysis of cage: a VirB4 homologue of *Helicobacter pylori* Cag-T4SS. *PLoS One* 2015;10:e0142606.

63. Klemm D, Heublein B, Fink H-P, Bohn A. Cellulose: fascinating biopolymer and sustainable RAW material. *Angewandte Chemie International Edition* 2005;44:3358–3393.

64. Sharma KD, Karki S, Thakur NS, Attri S. Chemical composition, functional properties and processing of carrot-a review. *J Food Sci Technol* 2012;49:22–32.

65. Mazzoli R, Bosco F, Mizrahi I, Bayer EA, Pessione E. Towards lactic acid bacteria-based biorefineries. *Biotechnol Adv* 2014;32:1216–1236.

66. Rossi F, Rudella A, Marzotto M, Dellaglio F. Vector-free cloning of a bacterial endo-1,4-beta-glucanase in *Lactobacillus plantarum* and its effect on the acidifying activity in silage: use of recombinant cellulolytic *Lactobacillus plantarum* as silage inoculant. *Antonie van Leeuwenhoek* 2001;80:139–147.

67. Moraïs S, Shterzer N, Grinberg IR, Mathiesen G, Eijsink VGH *et al*. Establishment of a simple *Lactobacillus plantarum* cell Consortium for cellulase-xylanase synergistic interactions. *Appl Environ Microbiol* 2013;79:5242–5249.

68. Moraïs S, Shterzer N, Lamed R, Bayer EA, Mizrahi I. A combined cell-consortium approach for lignocellulose degradation by specialized *Lactobacillus plantarum* cells. *Biotechnol Biofuels* 2014;7:112–115.

69. Stern J, Moraïs S, Ben-David Y, Salama R, Shamshoum M *et al*. Assembly of synthetic functional cellulosomal structures onto the cell surface of *Lactobacillus plantarum*, a potent member of the gut microbiome. *Appl Environ Microbiol* 2018;84:1–14.

70. Okano K, Zhang Q, Yoshida S, Tanaka T, Ogino C *et al*. D-Lactic acid production from cellooligosaccharides and beta-glucan using L-LDH gene-deficient and endoglucanase-secreting *Lactobacillus plantarum*. *Appl Microbiol Biotechnol* 2010;85:643–650.

71. Bates EE, Gilbert HJ, Hazlewood GP, Huckle J, Laurie JI *et al*. Expression of a Clostridium thermocellum endoglucanase gene in *Lactobacillus plantarum*. *Appl Environ Microbiol* 1989;55:2095–2097.

72. Scheirlinck T, Mahillon J, Joos H, Dhaese P, Michiels F. Integration and expression of alpha-amylase and endoglucanase genes in the *Lactobacillus plantarum* chromosome. *Appl Environ Microbiol* 1989;55:2130–2137.

73. Fukao M, Oshima K, Morita H, Toh H, Suda W *et al*. Genomic analysis by deep sequencing of the probiotic *Lactobacillus brevis* KB290 harboring nine plasmids reveals genomic stability. *PLoS One* 2013;8:e60521.
74. Zhang W, Yu D, Sun Z, Chen X, Bao Q *et al*. Complete nucleotide sequence of plasmid plca36 isolated from *Lactobacillus casei* Zhang. *Plasmid* 2008;60:131–135.
75. Ito Y, Kawai Y, Arakawa K, Honme Y, Sasaki T *et al*. Conjugative plasmid from *Lactobacillus gasseri* LA39 that carries genes for production of and immunity to the circular bacteriocin gassericin a. *Appl Environ Microbiol* 2009;75:6340–6351.
76. Tanizawa Y, Tohno M, Kaminuma E, Nakamura Y, Arita M. Complete genome sequence and analysis of *Lactobacillus hokkaidonensis* LOOC260T, a psychrotrophic lactic acid bacterium isolated from silage. *BMC Genomics* 2015;16:1–11.
77. van Kranenburg R, Golic N, Bongers R, Leer RJ, de Vos WM *et al*. Functional analysis of three plasmids from *Lactobacillus plantarum*. *Appl Environ Microbiol* 2005;71:1223–1230.
78. Kim D-H, Jeon Y-J, Chung M-J, Seo J-G, Ro Y-T. Complete sequence and gene analysis of a cryptic plasmid pLU4 in *Lactobacillus reuteri* strain LU4 (KCTC 12397BP). *Appl Biol Chem* 2017;60:145–153.
79. Jung JY, Lee SH, Jeon CO. Complete genome sequence of *Leuconostoc carnosum* strain JB16, isolated from kimchi. *J Bacteriol* 2012;194:6672–6673.
80. De Bruyne K, Camu N, De Vuyst L, Vandamme P. *Lactobacillus fabifermentans* sp. nov. and *Lactobacillus cacaonum* sp. nov., isolated from Ghanaian cocoa fermentations. *Int J Syst Evol Microbiol* 2009;59:7–12.
81. Curk MC, Hubert JC, Bringel F. *Lactobacillus paraplantarum* sp. now., a new species related to *Lactobacillus plantarum*. *Int J Syst Bacteriol* 1996;46:595–598.
82. Zanoni P, Farrow JAE, Phillips BA, Collins MD, pentosus L. *Lactobacillus pentosus* (Fred, Peterson, and anderson) sp. nov., nom. rev. *Int J Syst Bacteriol* 1987;37:339–341.
83. Pederson CS. A study of the species *Lactobacillus plantarum* (Orla-Jensen) Bergey, *et al*. *Journal of bacteriology* 1936;31:217.
84. Gu CT, Wang F, Li CY, Liu F, Huo GC *et al*. *Lactobacillus xiangfangensis* sp. nov., isolated from Chinese pickle. *Int J Syst Evol Microbiol* 2012;62:860–863.
85. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M *et al*. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2012;28:1647–1649.