environmental microbiology

Environmental Microbiology (2016) **18**(12), 4974–4989 doi:10.1111/1462-2920.13455

Nomadic lifestyle of *Lactobacillus plantarum* revealed by comparative genomics of 54 strains isolated from different habitats

Maria Elena Martino,[1]* Jumamurat R. Bayjanov,[2] Brian E. Caffrey,[3] Michiel Wels,[4] Pauline Joncour,[1] Sandrine Hughes,[1] Benjamin Gillet,[1] Michiel Kleerebezem,[5†] Sacha A.F.T. van Hijum[2,4†] and François Leulier[1†]

[1]*Institut de Génomique Fonctionnelle de Lyon (IGFL), Ecole Normale Supérieure de Lyon, CNRS UMR 5242, Université Claude Bernard Lyon 1, Lyon, France.*

[2]*Center for Molecular and Biomolecular Informatics, Nijmegen Center for Molecular Life Sciences, Radboud UMC, P.O. Box 9101, 6500 HB, Nijmegen, The Netherlands.*

[3]*Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, Berlin, 14195, Germany.*

[4]*NIZO food research, P.O. Box 20, 6710 BA, Ede, The Netherlands.*

[5]*Host Microbe Interactomics Group, Wageningen University, De Elst 1, 6708WD, Wageningen, The Netherlands.*

Summary

The ability of bacteria to adapt to diverse environmental conditions is well-known. The process of bacterial adaptation to a niche has been linked to large changes in the genome content, showing that many bacterial genomes reflect the constraints imposed by their habitat. However, some highly versatile bacteria are found in diverse habitats that almost share nothing in common. *Lactobacillus plantarum* is a lactic acid bacterium that is found in a large variety of habitat. With the aim of unravelling the link between evolution and ecological versatility of *L. plantarum*, we analysed the genomes of 54 *L. plantarum* strains isolated from different environments. Comparative genome analysis identified a high level of genomic diversity and plasticity among the strains analysed. Phylogenomic and functional divergence studies coupled with gene–trait matching analyses revealed a mixed distribution of the strains, which was uncoupled from their environmental origin. Our findings revealed the absence of specific genomic signatures marking adaptations of *L. plantarum* towards the diverse habitats it is associated with. This suggests fundamentally similar trends of genome evolution in *L. plantarum*, which occur in a manner that is apparently uncoupled from ecological constraint and reflects the nomadic lifestyle of this species.

Received 30 May, 2016; accepted 13 July, 2016. *For correspondence. E-mail: maria-elena.martino@ens-lyon.fr. Tel: +33 (0)4 26 73 13 27; Fax: +33 (0) 4 26 73 13 70. †Co-senior authors.

Introduction

Lactic acid bacteria (LAB) are a group of Gram-positive acid-tolerant bacteria that occupy a wide range of niches. The largest and most diverse genus of LAB is *Lactobacillus*, which includes more than 200 species. The environmental distribution of *Lactobacillus* species is highly variable; some species are exclusively found in specific habitats (e.g., *Lactobacillus helveticus* and *Lactobacillus delbrueckii* ssp. *bulgaricus* in diary products, *Lactobacillus johnsonii* and *Lactobacillus gasseri* in vertebrate gastrointestinal tracts) and other species, such as *Lactobacillus plantarum* and *Lactobacillus casei*, are encountered in a variety of different environments (Cai *et al.*, 2009; Siezen *et al.*, 2010). Although *Lactobacillus* has traditionally been defined as a genus, its genetic diversity is larger than that of a typical family(Bringel *et al.*, 1996; Sun *et al.*, 2015). Several comparative genome studies have demonstrated the extraordinary genomic and metabolic diversity of lactobacilli, a feature that allows them to adapt and survive in diverse environments (Curk *et al.*, 1996; Makarova *et al.*, 2006; Sun *et al.*, 2015).

Lactobacillus plantarum is an extremely versatile LAB that has been isolated from a variety of habitats, such as plants, the gastro-intestinal tracts of human, animals, including poultry and insects, as well as food materials, such as meat, fish, vegetables and raw or fermented dairy products (Torriani *et al.*, 2001; Siezen *et al.*, 2010). *L. plantarum* is a facultative heterofermentative species that plays an important role in different food and health industries. It is one of the best-characterized vegetal-associated bacteria that transform a multitude of plant-derived raw

materials through fermentation (de Las Rivas *et al*., 2005; De Vuyst *et al*., 2009). Also, as some *L. plantarum* strains are naturally ocurring human commensals, they have been marketed as probiotics (Molenaar *et al*., 2005; Klarin *et al*., 2008; Siezen *et al*., 2010), and their potential beneficial effects on human or animal health have been recently reported as being able to promote juvenile growth in both drosophila and mouse in the presence of nutritional challenges (Molenaar *et al*., 2005; Connelly, 2008; Siezen *et al*., 2010; Storelli *et al*., 2011; Giri *et al*., 2013; Yang *et al*., 2014; Erkosar *et al*., 2015; Schwarzer *et al*., 2016). Its remarkable adaptability, wide industrial utility and potent impact on animal phyisiology have made *L. plantarum* an organism of significant interest to the scientific community.

Several studies have depicted the genetic diversity of *L. plantarum* strains through different phenotypic (Cai *et al*., 2009; Siezen *et al*., 2010) and genotypic approaches, such as AFLP, RAPD (Bringel *et al*., 1996; Torriani *et al*., 2001; Kleerebezem *et al*., 2003; Oh *et al*., 2010; Siezen *et al*., 2010; 2011; Frese *et al*., 2011; Sun *et al*., 2015), multi-locus sequence typing (MLST) (Curk *et al*., 1996; de Las Rivas *et al*., 2005; Makarova *et al*., 2006; Smokvina *et al*., 2013; Kafsi *et al*., 2014; Sun *et al*., 2015), and microarray-based comparative genome hybridization (CGH) (Torriani *et al*., 2001; Molenaar *et al*., 2005; Siezen *et al*., 2010). These approaches allowed the differentiation of *L. plantarum* at both inter- and intra-species level and investigated a potential link between genome and niche adaptation or fitness (de Las Rivas *et al*., 2005; Molenaar *et al*., 2005; De Vuyst *et al*., 2009; Siezen *et al*., 2010; Frese *et al*., 2011; Dutilh *et al*., 2013; Smokvina *et al*., 2013). The process of bacterial adaptation to a new habitat has usually been associated with changes in genome content and regulation (Kleerebezem *et al*., 2003; Molenaar *et al*., 2005; Klarin *et al*., 2008; Cai *et al*., 2009; Siezen *et al*., 2010; 2011). Notably, strains of different *Lactobacillus* species have been reported to adapt to defined environments by genome specialization driving niche-specific fitness. For example, *L. reuteri* strains isolated from different vertebrate intestinal tracts display host-adapted genome evolution paths, which translate into optimized ecological performance in their respective hosts (van Kranenburg *et al*., 2005; Molenaar *et al*., 2005; Connelly, 2008; Oh *et al*., 2010; Siezen *et al*., 2010; Frese *et al*., 2011; Storelli *et al*., 2011; Lukjancenko *et al*., 2012; Frese *et al*., 2013; Giri *et al*., 2013; Yang *et al*., 2014; Erkosar *et al*., 2015; Schwarzer *et al*., 2016); *L. paracasei* and *L. delbrueckii* strains adapt specifically to dairy environments through a process characterized by genome decay (Smokvina *et al*., 2013; Kafsi *et al*., 2014; Mendes-Soares *et al*., 2014). *L. plantarum* is a generalist species that encompasses phenotypically and genotypically diverse strains whose evolutionary relatedness and history is not clear. In particular, microarray-based CGH studies performed on *L. plantarum* clearly indicate the high genomic diversity of the species (Molenaar *et al*., 2005; Siezen *et al*., 2010; Frese *et al*., 2011). However, despite these CGH insights, whether a potential connection exists between a specific genomic background and a defined source of isolation remains an open question.

High-throughput genomic approaches provide a deep understanding of the evolution and ecology of microorganisms. Through analysing the genetic variability across different organisms, these approaches help to assign putative adaptive features as well as putative functional roles to a given species/strain in an ecosystem. The complete genome sequences of 10 *L. plantarum* strains (WCFS1, JDM1, ST-III, ZJ316, P8, 16, B21, 5-2, ZS2058 and HCF8) together with 30 draft genomes are currently available on NCBI (Geer *et al*., 2010). With the aim of gaining more insights into the functional capabilities and differences of *L. plantarum* strains, we sequenced the genomes of 43 additional *L. plantarum* strains that were isolated from a variety of food environments (such as fermented vegetables, dairy products, fruit and meat) and two natural animal hosts (human and *Drosophila melanogaster*). We next compared these 43 genomes to the existing genomes of 11 *L. plantarum* strains that were publicly available at the time we started our analyses (six complete and five draft genomes). Our analysis therefore includes a wide-range of diverse strains, which enables the in depth analysis of *L. plantarum* phylogenomics. We now report the comparative analysis of these genomes with the main objective of exploring the potential link between the intra-species genetic variability and environmental origin.

Results

Comparative analysis of 54 L. plantarum *strains*

A detailed description of the *L. plantarum* reference strain (WCSF1) has been previously reported (Kleerebezem *et al*., 2003). It consists of a circular chromosome of 3.3 Mb and three plasmids of 1.9 kb, 2.3 kb and 36.1 kb (van Kranenburg *et al*., 2005; Siezen and van Hylckama Vlieg, 2011). To broadly investigate the genomic diversity of *L. plantarum* species, we chose 54 *L. plantarum* strains whose origins encompassed a large spectrum of habitats: 17 strains isolated from fermented fruits and vegetables, 11 strains from human origin (oral cavity, urine and intestinal tract, faeces and one putatively from the spinal fluid), 7 strains isolated from silage, 6 strains of dairy origin, 6 strains isolated from meat products and 6 strains isolated from adult *D. melanogaster*'s midgut (Table 1).

The genomes of 39 *L. plantarum* strains were sequenced using Illumina Miseq sequencing technology, while the genomes of 4 strains were sequenced using Ion Torrent PGM (Life Technologies) (Supporting information Table S1). The draft genome sequences of the 43 *L.*

Table 1. *Lactobacillus plantarum* strains characterized in this study.

Strain	Isolation source	Geographical origin	Genome	No. of contigs	No. of OGs	Reference
16	Malt production steep water	n.a[a]	Complete	1	2862	Crowley *et al.* (2013)
19.1	Adult *Drosophila* midgut	Israël	Draft	42	3253	Sharon *et al.* (2010)
ATCC14917	Cabbage pickled	Denmark	Draft	39	2924	Tatusova *et al.* (2014)
CNW10	Adult *Drosophila* midgut	United States of America	Draft	88	2852	Newell *et al.* (2014)
ER	Surface sterilized adult *Drosophila*	France	Draft	37	3126	Storelli *et al.* (2011)
IPLA88	Sourdough	Italy	Draft	208	2930	Ladero *et al.* (2013)
JDM1	Grass silage	China	Complete	1	2799	Zhang *et al.* (2009)
NAB1	Adult *Drosophila* midgut	Switzerland	Draft	44	3184	Broderick *et al.* (2014)
NAB2	Adult *Drosophila* midgut	Switzerland	Draft	44	3229	Broderick *et al.* (2014)
NC8	Grass silage	Sweden	Draft	10	2801	Axelsson *et al.* (2012)
NIZO1837	Human colon	United Kingdom	Draft	33	2943	Siezen *et al.* (2010)
NIZO1838	Human stool	France	Draft	81	2721	Siezen *et al.* (2010)
NIZO1839	Sour cassava	South America	Draft	43	2814	Siezen *et al.* (2010)
NIZO1840	Cereal fermented (Ogi)	Nigeria	Draft	127	2843	Siezen *et al.* (2010)
NIZO2029	Raw cheese with rennet	Italy	Draft	46	2871	Pepe *et al.* (2004)
NIZO2256	Human stool	France	Draft	70	2686	Siezen *et al.* (2010)
NIZO2257	Human stool	France	Draft	114	2848	Siezen *et al.* (2010)
NIZO2258	Human urine	France	Draft	110	2832	Siezen *et al.* (2010)
NIZO2259	Human tooth abscess	France	Draft	78	2995	Siezen *et al.* (2010)
NIZO2260	Human intestine	United Kingdom	Draft	34	2926	Siezen *et al.* (2010)
NIZO2262	Silage	n.a.	Draft	25	2860	Siezen *et al.* (2010)
NIZO2263	Silage	n.a.	Draft	55	2912	Siezen *et al.* (2010)
NIZO2264	Silage	France	Draft	23	2754	Siezen *et al.* (2010)
NIZO2457	Pork pickled sour sausage	Vietnam	Draft	42	2962	Siezen *et al.* (2010)
NIZO2484	Pork pickled sour sausage	Vietnam	Draft	65	2996	Siezen *et al.* (2010)
NIZO2485	Pork pickled sour sausage	Vietnam	Draft	39	2974	Siezen *et al.* (2010)
NIZO2494	Pork pickled sour sausage	Vietnam	Draft	47	2934	Siezen *et al.* (2010)
NIZO2535	Orange fermented	Vietnam	Draft	35	3041	Siezen *et al.* (2010)
NIZO2726	Maize ensilage	n.a.	Draft	27	2853	Siezen *et al.* (2010)
NIZO2741	Cabbage kimchi	Japan	Draft	30	2959	Nishitani *et al.* (2004)
NIZO2753	Sourdough fermented	Italy	Draft	39	2769	Pepe *et al.* (2004)
NIZO2757	Sourdough fermented	Italy	Draft	62	2802	Siezen *et al.* (2010)
NIZO2766	Sourdough fermented	Italy	Draft	64	2820	Siezen *et al.* (2010)
NIZO2776	Cheese	n.a.	Draft	149	2909	Siezen *et al.* (2010)
NIZO2801	Turnip pickled	Japan	Draft	55	2936	Nishitani *et al.* (2004)
NIZO2802	Cheese	Japan	Draft	58	3001	Siezen *et al.* (2010)
NIZO2806	Sauerkraut	United Kingdom	Draft	33	2833	Vermeiren *et al.* (2003)
NIZO2814	Wine red grapes	Italy	Draft	30	2974	Siezen *et al.* (2010)
NIZO2830	n.a.	n.a.	Draft	23	2904	Siezen *et al.* (2010)
NIZO2831	Grass silage	United States of America	Draft	27	2899	Siezen *et al.* (2010)
NIZO2855	Pork pickled sour sausage	Vietnam	Draft	22	2863	Siezen *et al.* (2010)
NIZO2877	Hot dog	Vietnam	Draft	2	2853	Martino *et al.* (2015a)
NIZO2889	Banana fermented	Vietnam	Draft	66	2923	Siezen *et al.* (2010)
NIZO2891	Radish pickled	Vietnam	Draft	41	3064	Siezen *et al.* (2010)
NIZO3400	Milk	Senegal	Draft	87	2971	Siezen *et al.* (2010)
NIZO3892	Human spinal fluid	France	Draft	25	2976	Siezen *et al.* (2010)
NIZO3893	Human stool	France	Draft	98	2909	Siezen *et al.* (2010)
NIZO3894	Vegetables	n.a	Draft	29	2899	Siezen *et al.* (2010)
P8	Dairy product	China	Complete	1	2768	Tatusova *et al.* (2014)
ST-III	Kimchi	China	Complete	1	2820	Wang *et al.* (2011)
UCMA3037	Raw milk camembert cheese	France	Draft	68	2754	Naz *et al.* (2013)
WCFS1	Human saliva	England	Complete	1	2970	Siezen *et al.* (2011)
WJL_IGFL	Adult *Drosophila* posterior midgut	South Korea	Draft	13	3097	Martino *et al.* (2015b)
ZJ316	Newborn infant faeces	China	Complete	1	2946	Li *et al.* (2013)

[a]n.a: Not available.

plantarum strains were analysed together with 11 publicly available *L. plantarum* genomes (Table 1) and aligned to the *L. plantarum* WCFS1 reference genome. To report all the detailed information about the comparative analyses of these genomes, we have created an online database that presents all the results concerning the strain-specific gene

content of the 54 *L. plantarum* strains relative to the entire gene set of all strains, named pan-genome, including the comparative analysis of functionalities encoded in the individual genomes (http://igfl.ens-lyon.fr/equipes/f.-leulier-functional-genomics-of-host-intestinal-bacteria-interactions/l-plantarum/). The sizes of the sequenced genomes range from 3 to 3.6 Mb, and encompassed a set of 1957 Orthologous Groups (OGs) that are shared by all *L. plantarum* strains (core genome). The pan-genome size of the 54 genomes amounts to 7107 OGs (Table D1 available at http://igfl.ens-lyon.fr/equipes/f.-leulier-functional-genomics-of-host-intestinal-bacteria-interactions/l-plantarum/), which include the core genome and the variable genome (variome). The variome refers to genes not present in all strains, that is, genes present in two or more strains and genes unique to single strains. The large size of pan-genome of the *L. plantarum* strains is consistent with what has been previously reported for many *Lactobacillus* species (Molenaar *et al*., 2005; Siezen *et al*., 2010; Siezen and van Hylckama Vlieg, 2011; Lukjancenko *et al*., 2012; Mendes-Soares *et al*., 2014). Nevertheless, the pan-genome did not appear to approach saturation with the current strain collection, implying that the genetic repertoire of *L. plantarum* exceeds the current pan-genome estimate (Supporting information Fig. S1). The core genome of 1957 OGs (Table D2 available at http://igfl.ens-lyon.fr/equipes/f.-leulier-functional-genomics-of-host-intestinal-bacteria-interactions/l-plantarum/) covers 66% of *L. plantarum* WCFS1 annotated genes, and is of a similar size as has been reported for other *Lactobacillus* species (Siezen *et al*., 2010; Frese *et al*., 2011; Dutilh *et al*., 2013; Smokvina *et al*., 2013). Moreover, the core-genome size is in good agreement with that obtained by microarray-based CGH analysis of *L. plantarum* strains that predicted 2049 core genes(Siezen *et al*., 2010; Siezen and van Hylckama Vlieg, 2011), and used 20 strains, which could explain the somewhat larger size of the core-genome estimate. The *L. plantarum* core genome contains the anticipated shared genetic repertoire involved in replication, transcription and translation, as well as genes involved in energy production and amino acid- and carbohydrate-transport and metabolism (Supporting information Fig. S2). Notably, 226 OGs of the core genome are annotated to encode hypothetical proteins that are conserved with yet unknown function (Table D2 available at http://igfl.ens-lyon.fr/equipes/f.-leulier-functional-genomics-of-host-intestinal-bacteria-interactions/l-plantarum/). The variome in the *L. plantarum* collection contains 5150 OGs, among which 4500 OGs appeared to be scaffolded in the chromosome (Table D3 available at http://igfl.ens-lyon.fr/equipes/f.-leulier-functional-genomics-of-host-intestinal-bacteria-interactions/l-plantarum/). This estimate may be slightly high, since a fraction of the OGs represent fragments of mobile element (e.g., transposases of the IS elements) that can be part of scaffolds, or part of independently replicating plasmids. We estimate that there are 2686–3253 OGs per strain (Table 1), of which 2600–3000 OGs are on contigs in scaffolded chromosomes and 0–300 are non-scaffolded and may been coded by plasmids. Most of the sequenced *L. plantarum* strains appeared to have several plasmids (Table D4 available at http://igfl.ens-lyon.fr/equipes/f.-leulier-functional-genomics-of-host-intestinal-bacteria-interactions/l-plantarum/), with the exception of 6 strains (CNW10, NIZO2264, NIZO2855, NIZO2877, NC8 and JDM1). Our comparative genome analysis identified 4137 novel OGs that are not present in the reference genome WCFS1 (Supporting information Table S2).

We evaluated the degree of gene content variation among the 54 genomes by analysing the presence/absence of the OGs encoding proteins associated with central cellular processes and/or molecular functions. Metabolic maps of OGs were created using KEGG database (http://www.genome.jp/keg/pathway.html) (http://igfl.ens-lyon.fr/equipes/f.-leulier-functional-genomics-of-host-intestinal-bacteria-interactions/l-plantarum/). As reported by Molenaar and colleagues (Molenaar *et al*., 2005), the highest genetic conservation in *L. plantarum* strains is observed for OGs involved in energy metabolism, or biosynthesis or degradation of cellular structural components, such as nucleotides, proteins and lipids. Despite the high conservation of these biosynthetic pathways, *L. plantarum* shows high inter-strain genetic variability. Consistent with previous studies, large variable regions exist among the 54 *L. plantarum* genomes (Molenaar *et al*., 2005; Siezen *et al*., 2010; Siezen and van Hylckama Vlieg, 2011). These regions include genes involved in exopolysaccharide (EPS) biosynthesis, restriction-modification, sugar-importing phosphotransferase (PTS) systems and other transport functions, sugar metabolism, bacteriocin production, as well as the notoriously variable functions associated with prophages, insertion-sequence (IS) elements and transposases. A detailed analysis of the genes belonging to the conserved and variable regions of the 54 *L. plantarum* strains is available in the online database (http://igfl.ens-lyon.fr/equipes/f.-leulier-functional-genomics-of-host-intestinal-bacteria-interactions/l-plantarum/). The genes are grouped depending on their metabolic functions and each category presents a detailed description of gene distribution and activity, together with tables summarizing all gene information and their presence/absence among the strains, and the metabolic maps related to each category.

Gene content analysis does not reveal specialization of L. plantarum *for a specific environment*

The process of environmental adaptation may include events of gene gain and/or loss, or genome decay. Genes

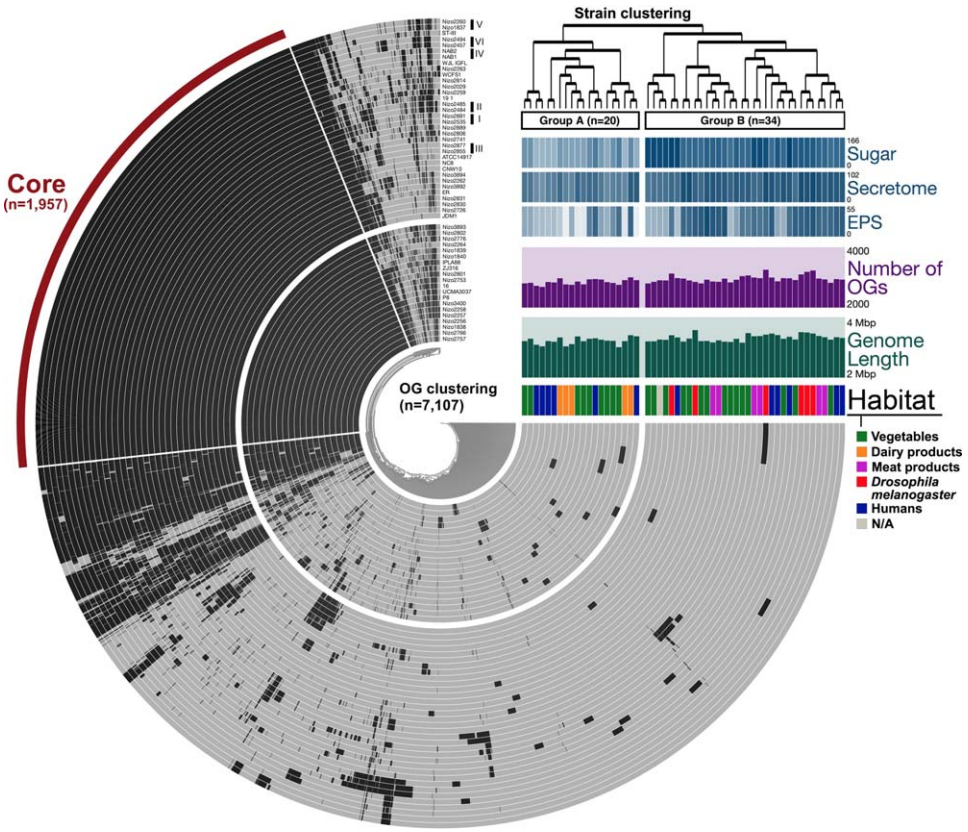

Fig. 1. OG distribution across *L. plantarum*. The center of the figure shows the hierarchical clustering of OGs based on their presence/absence. Each ring represents a *L. plantarum* strain and each layer displays the OG distribution. Black color stands for OG presence, whereas grey colour represents OG absence. The red bar groups the OGs belonging to the core genome. The top right panel reports additional metadata information about the dataset. The clustering present on the top right refers to strain grouping based on OG distribution. The occurrence of genes belonging to the three most variable genetic regions (exopolysaccharide biosynthesis (EPS), sugar cassettes and secretome) is shown. The colour intensity of each bar reflects the increasing number of OGs present. Each strain (layer) is labelled depending on the source of isolation. The figure has been generated using Anvi'o software (Eren *et al.*, 2015).

whose functions are dispensable for a strain's fitness in a particular environment can be lost during adaptation. We therefore analysed the potential link between strain origin and their gene content. We first sorted strains by comparing the total number of OGs they encode (Supporting information Fig. S3), and found no strain grouping dependent on the origin of isolation, although it is quite remarkable that 5 out of the 6 strains isolated from *D. melanogaster*'s midgut (19.1, ER, NAB1, WJL_IGFL, NAB2; Supporting information Fig. S3) encode the highest number of OGs. Next, we generated a dendrogram of the 54 *L. plantarum* strains based on the presence/absence of each OG in their pan-genome (Fig. 1). Two major groups could be distinguished: strains isolated from dairy products belong to group A, strains isolated from meat products and *D. melanogaster* cluster in group B, while strains of human and plant origin are spread across the two groups. Notably, in each group, strains isolated from the same habitat do not appear to be closely related (i.e., within the same subcluster), indicating that gene distribution poorly reflects strain origin. As shown in Figure 1, the main differences in OG distribution between the two clusters mainly relate to genes involved in sugar metabolism and EPS biosynthesis. In addition, the 2-cluster separation results from the distribution of OGs belonging to amino-acid metabolism and cell wall biosynthesis (Table D1 available at http://igfl.ens-lyon.fr/equipes/f.-leulier-functional-genomics-of-host-intestinal-bacteria-interactions/l-plantarum/).

Next, using the PhenoLink tool with its default parameters (Bayjanov *et al.*, 2012), we performed a Gene Trait Matching (GTM) analysis, a statistical method based on the Random Forest algorithm that is used to correlate gene presence and absence to phenotypes (Breiman, 2001; Liaw and Wiener, 2002; Dutilh *et al.*, 2013). PhenoLink ranks gene-to-phenotype associations based on random forest generated local importance scores and combines this with gene's presence/absence in strains with given phenotype to provide faster screening of associations (Bayjanov *et al.*, 2012). We used the strains' origins as phenotypes and OG presence/absence as feature values for the GTM. Among the most

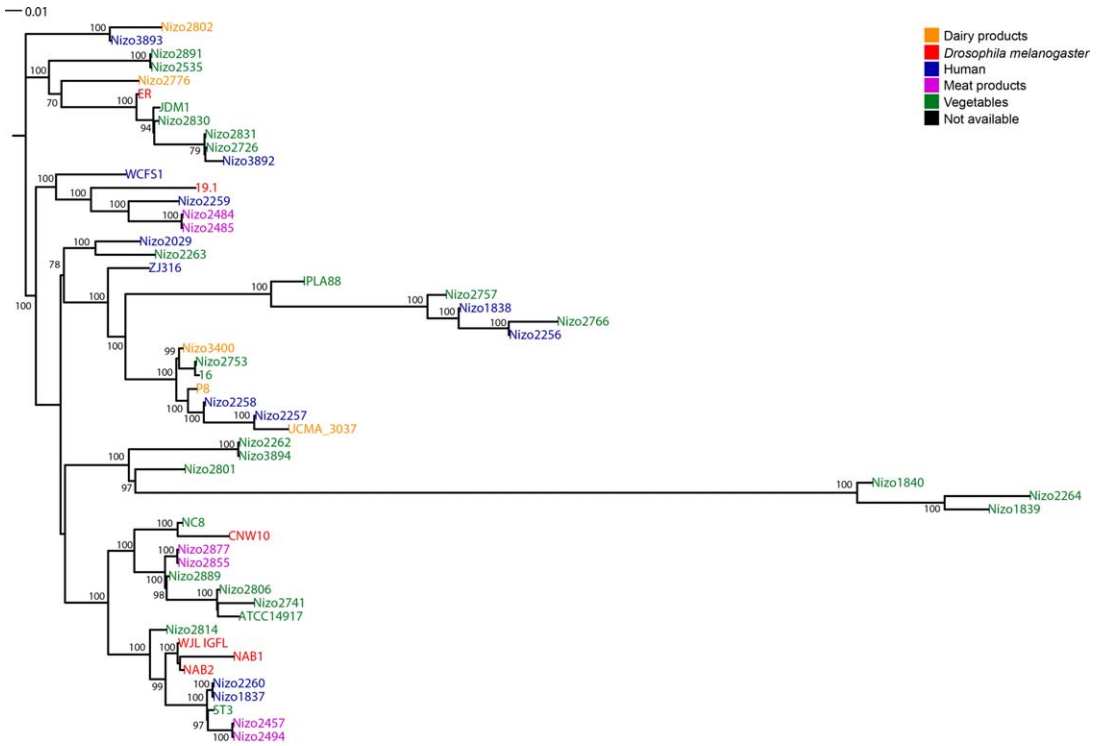

Fig. 2. Phylogeny of *L. plantarum*. The evolutionary history of the core genome of the54 *L. plantarum* strains was inferred by using an approximately maximum likelihood method (Price *et al*., 2010). The concatenated protein sequences of the OGs belonging to the core genome were used as input. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown on nodes scoring >50%. Each strain is colour coded by origin of isolation.

important associations, we identified groups of genes whose presence or absence associated with the origins isolated from vegetables, meats, humans and *D. melanogaster*. No gene association could be pinpointed for the dairy origin (Supporting information Table S3). However, when we defined the environmental signature as genes mostly present in strains isolated from a habitat but mostly absent in all the others strains (at least 75% of the strains) or *viceversa*, we found few genes: mostly hypothetical proteins or genes related to conjugation in the *D. melanogaster* strains, bear such a robust habitat signature. However, currently it is difficult to assign specific functions to these genes in terms of habitat adaptation. Therefore, these observations indicate that gene presence/absence is a poor indicator of origin for the considered *L. plantarum* strains.

Phylogeny and analysis of allelic and functional divergence in L. plantarum *confirm the absence of environmental-specialization*

Bacterial adaptation to environment may not just depend on gene absence or presence, but can also be the result of accumulation of adaptive specific allelic variations of conserved genes. To gain insights into the potential role of such allelic variation in environmental specialization and to reveal the evolutionary relationships of the *L. plantarum* strains studied, we performed a phylogenetic analysis the 54 strains. A phylogenetic tree of the 16S ribosomal RNA gene was initially generated to have a first glance of nucleotide divergence among the strains (Supporting information Fig. S4). The overall mean distance was calculated using the number of differences model in MEGA7 software (Kumar *et al*., 2016). It resulted to be 96.7%, which is slightly lower than the similarity score used as cutoff for species discrimination (Reller *et al*., 2007). A phylogenetic analysis was also carried out on the entire core genome of the 54 strains (Fig. 2) and the average amino-acid identity (AAI) was found to be 82.4%. The core genome phylogeny clearly shows a mixed distribution of strains isolated from similar sources. Thus, no phylogenetic cluster related to a defined habitat appears to be present, but pairs of strains sharing more similar core genomic OGs were identified (NIZO2891 and NIZO2535, NIZO2484 and NIZO2485, NIZO2877 and NIZO2855, NIZO2260 and NIZO1837, NIZO2494 and NIZO2457, NIZO2831 and NIZO2726, NIZO2262 and NIZO3894). These strains are not clonal but share a highly similar set of core genome associated OGs; however, they have also the same geographical origin (Table 1), therefore their phylogenetic relatedness cannot be solely attributed to a process of environmental adaptation.

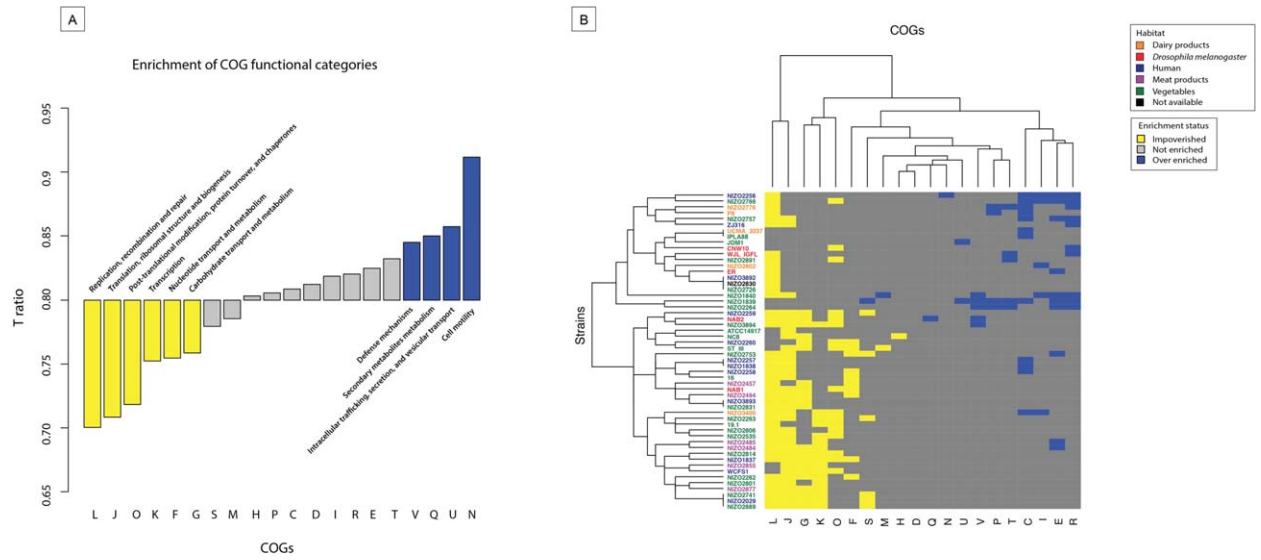

Fig. 3. Analysis of functional divergence in *L. plantarum*. (a) Functional divergence of clusters of orthologous groups (COGs) was classified into three categories; impoverished (yellow, 6 groups), enriched (blue, 4 groups) or neither impoverished nor enriched (grey, 10 groups).The *y*-axis represents the T ratio = *x*/*y*, where *x* is the number of proteins with functional divergence and *y* is the total number of proteins tested. (b) Heatmap showing the hierarchical clustering of the main patterns of functional divergence in the 54 *L. plantarum* strains (tree on the left-hand side, and strain identifiers on the right-hand side), as well as COG category clustering (shown on the top). Each strain is colour coded by origin of isolation.

To further investigate the evolutionary history of *L. plantarum* strains and find potential habitat-adaptation signatures, we tested if the phylogenetic relatedness of the strains is coupled to any functional divergence, i.e., the process by which new genes and functions originate through modification of existing ones. To probe for functional divergence, we employed the software tool developed by Caffrey *et al.* (2012) to pinpoint enrichment or impoverishment of amino-acid replacements within the predicted coreproteome; enrichment may indicate specialization, wherease impoverishment signifies fixation. Subsequently, correlating such functional divergence detected in the individual proteins to the source of isolation can reveal whether amino-acid replacement(s) may have influenced the evolution of new functions within a group of strains originating from the same source. Importantly, this analysis takes into account the potential occurrence of horizontal gene transfer, assuming that the phylogenies of individual proteins may differ from the whole genome phylogeny, and provides a different representation of the evolutionary relationships among the strains. The proteins belonging to the core proteome of the 54 strains were assigned to Clusters of Orthologous Groups (COGs). Each COG consists of a group of proteins found to be orthologous and likely corresponds to an ancient conserved domain. We discarded the proteins with no COG annotation and those that had ambiguous COG annotations. In addition, all proteins including less than 9 strains were discarded. As a result, we found that 80% of the COGs was either impoverished or not enriched for functional divergence among strains compared with the background (Fig. 3A). The impoverished COGs (30%) are almost all related to genome information storage and processing such as DNA replication, recombination and repair, transcription, translation, ribosomal structure and biogenesis, protein turnover and chaperones. Caffrey *et al.* previously made similar observations based on the analysis of 750 complete bacterial proteomes (Caffrey *et al.*, 2012) but their study also included COGs related to transport and metabolism of carbohydrates and nucleotides. The remaining 20% of COGs were found to be enriched in functional divergence (Fig. 3A): they include cellular defence mechanisms (V), cell motility (N), secondary metabolites biosynthesis, transport and catabolism (Q) and intracellular trafficking, secretion and vesicular transport (U) (Supporting information Table S4). CAFS software was also used to verify whether each strain exhibited significant functional divergence in a particular COG (Fig. 3B). No clear habitat-dependent pattern was found: the distribution of both impoverished and enriched categories resulted to be related to strain clustering, rather than to the environmental origin of the isolates (Fig. 3B). Notably, the functional divergence revealed for the enriched COGs (Fig. 3A) is represented in a very limited number of strains (C: 5, U: 2, Q: 1, N: 1) (Fig. 3B) that nevertheless were sufficient to statistically support the divergence of those COGs. However, it is important to state that gray squares might be slightly but not statistically divergent on their own, but

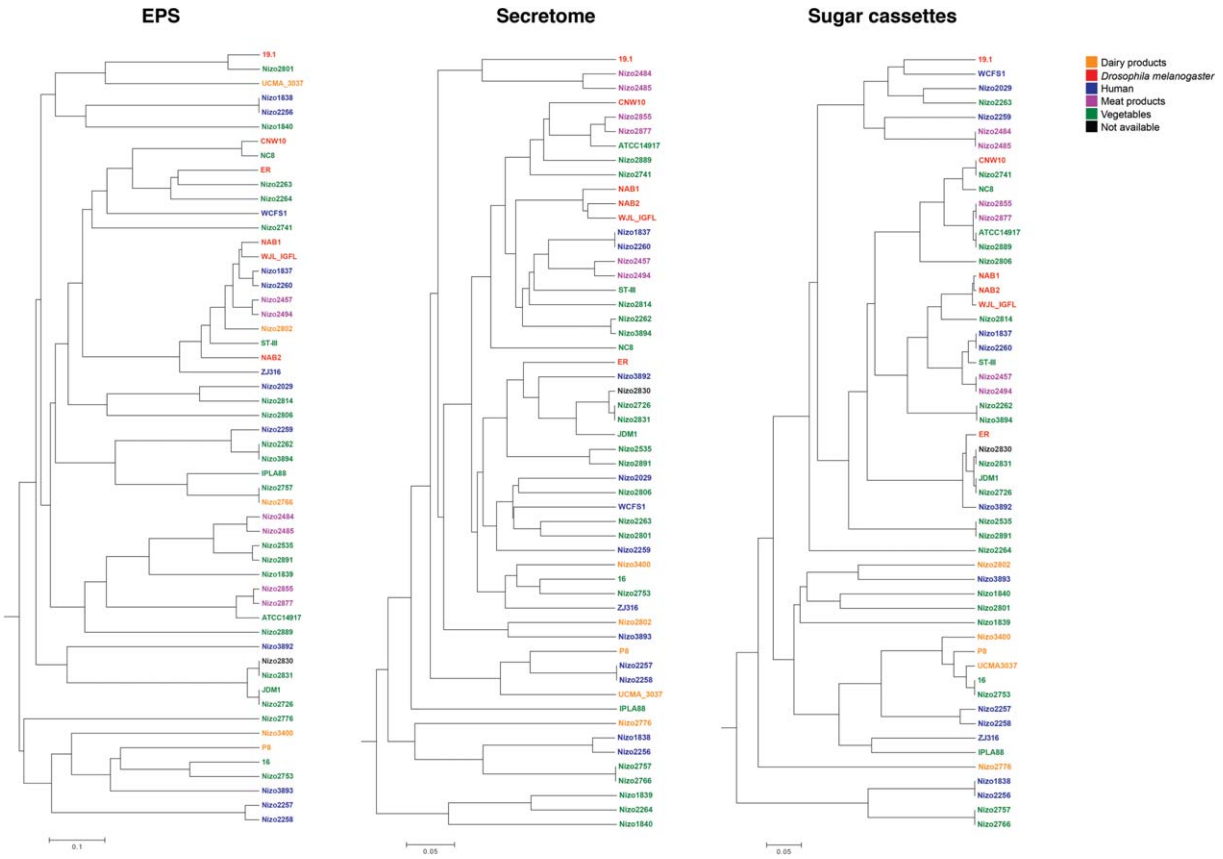

Fig. 4. Hierarchical clustering of the 54 *L. plantarum* strains based on the most variable regions. The 54 *L. plantarum* strains were clustered based on the presence/absence of the genes belonging to the EPS, sugar metabolism and secretome categories.The dendrograms are derived from UPGMA analysis of binary data generated from the 54 *L. plantarum* strains. Each strain is colour coded by origin of isolation.

finally contributing to the divergence of the whole category, which can explain why the 4 COGs were found to be significantly enriched.Considering the two enriched COGs including more than one isolate (C, U), the strains have different environmental origins (Fig. 3B), implying that source of isolation is not related to the functional enrichment of those COGs. This conclusion was further supported by clustering the strains based on the functional divergence of each COG group, which is important to verify whether the strains eventually cluster according to their source of isolation (Fig. 3B). Taken together, these results indicate that the functional divergence of the *L. plantarum* strains is not correlated with their source of isolation.

The analysis of highly variable regions in L. plantarum *supports the lack of environmental adaptation processes*

Thus far, the analysis of the gene content, the phylogenetic study of the core genome and the analysis of the proteome for functional divergence failed to identify any environmental signatures in the genome of the 54 *L. plantarum* strains. However, the phylogenetic analysis reported above does not take into account the variome of the 54 *L. plantarum* strains. In addition, when we clustered pan-genome OG distribution based on the presence/absence of OGs belonging to the most variable functionalities (secretome, sugar metabolism and EPS), we found a potential signature of clustering into Group A or B (Fig. 1). Therefore, the core genome phylogeny depicts the evolution of the genes that are presumably necessary for all strains, but the distribution pattern of the variome might reflect a specific evolutionary trajectory taken by the respective strains. Therefore, we decided to focus on the variable functional categories and analyse each of them separately to evaluate whether the OG content of a particular functional category may predict the strains' origins. We performed independent hierarchical clustering of 3 of the functional categories that displayed the highest variability in terms of OG presence and absence among the 54 *L. plantarum* strains (i.e., EPS, sugar utilization cassettes and secretome) (Fig. 4). In addition, we performed

independent phylogenetic analyses (Supporting information Figs S5–S7) within the shared gene sequences of each category.

EPS/CPS biosynthesis genes. The gene clusters involved in the EPS/CPS biosynthesis vary considerably in size, composition, sequence and gene order across the 54 *L. plantarum* strains. A detailed description and comparative analysis of these regions is reported in the database we have developed online (http://igfl.ens-lyon.fr/equipes/f.-leulier-functional-genomics-of-host-intestinal-bacteria-interactions/l-plantarum-on-line-material/eps-text-description). The distribution and genetic divergence analysis of the EPS/CPS-assigned OGs (Fig. 4 and Supporting information Fig. S5, respectively) are significantly correlated (p value = 9.99001e−04), revealing that the clustering of the strains on basis of the genetic divergence of the OGs in this category is strongly driven by the presence/absence distribution of these OGs. However, unlike for what has been observed for *L. paracasei* (Smokvina *et al*., 2013), no correlation was found between the source of isolation and the presence of specific (subsets of) EPS/CPS-assigned OGs (Supporting information Fig. S5). The strains with the fewest EPS/CPS-assigned OGs as well as those with the most EPS/CPS-assigned OGs are of various origins (human, dairy products and vegetables) (Table D9 available at http://igfl.ens-lyon.fr/equipes/f.-leulier-functional-genomics-of-host-intestinal-bacteria-interactions/l-plantarum/database-for-comparative-genomics-of-54-l-plantarum-strains-7-7). Moreover, strains isolated from the same habitat appear to be scattered across the phylogenetic tree (Supporting information Fig. S5), and strains that contain a similar set of EPS/CPS-assigned OGs were isolated from different habitats. These results demonstrate that the strain-specific distribution of genes that were assigned to the functional category of EPS/CPS biosynthesis does not reflect the origin of isolation of *L. plantarum*.

Secretome. The OGs assigned to the secretome were sorted and categorized into functional categories (Table D11 available at http://igfl.ens-lyon.fr/equipes/f.-leulier-functional-genomics-of-host-intestinal-bacteria-interactions/l-plantarum/database-for-comparative-genomics-of-54-l-plantarum-strains-7-7). Consistent with previous studies (Frese *et al*., 2011; Douillard *et al*., 2013; Smokvina *et al*., 2013), the most variable sub-categories of the secretome included cell-surface complexes (*csc*), ATP-binding cassette (ABC) transporters and bacteriocins.The distribution of secretome-assigned OGs among the 54 *L. plantarum* strains revealed no correlation with the origin of isolation of the strains (Fig. 4). Analogous to what was observed for the genetic divergence analysis of the EPS/CPS assigned OGs, the genetic divergence analysis of the secretome-assigned OGs appeared to be quite strongly driven by the presence/absence of these genes among the strains. However, the refinement of this clustering by the genetic divergence did not reveal an association with the origin of isolation (Supporting information Fig. S6).

Sugar metabolism. Sugar metabolism genes are highly variably distributed among LAB (Molenaar *et al*., 2005; Cai *et al*., 2009; Siezen *et al*., 2010; Smokvina *et al*., 2013). The presence or absence of these genes presumably reflects adaptation to the availability of substrates for growth in different habitats. Previous comparative studies conducted on other LAB species showed a strong correlation between thesource of isolation and sugar metabolism (Frese *et al*., 2011; Douillard *et al*., 2013). *L. plantarum* has been reported to contain a highly variable repertoire of genes related to sugar import and utilization, which appears to cluster in a so-called sugar utilization island of the genome (Kleerebezem *et al*., 2003; Molenaar *et al*., 2005; Cai *et al*., 2009; Siezen *et al*., 2010). However, to date, no clear relationship has been deduced between the sugar utilization repertoire of different strains of *L. plantarum* and their source of isolation. The genomic analysis of the 54 *L. plantarum* strains confirmed the high plasticity and diversity of the genetic repertoire related to sugar metabolism in this species. The distribution of all OGs involved in sugar metabolism in the 54 *L. plantarum* genomes is reported in the database that we developed online (Table D12 available at http://igfl.ens-lyon.fr/equipes/f.-leulier-functional-genomics-of-host-intestinal-bacteria-interactions/l-plantarum/database-for-comparative-genomics-of-54-l-plantarum-strains-7-7).

The OGs assigned to carbohydrate utilization are generally organized in ‘cassette’-like genomic loci, whose distribution among the strains are hypervariable and largely explain the differences found in their genome size. This finding is similar to what has been reported previously about the genomic analysis of 37 *L. paracasei* strains (Hottes *et al*., 2013; Smokvina *et al*., 2013).

The initial pan-genome analysis already revealed a clustering of the 54 *L. plantarum* strains in two main groups, which appeared to include a prominent involvement of the distribution of genes belonging to sugar metabolism (Fig. 1). The hierarchical clustering of the strains based on the distribution of the OGs belonging to this category did not cluster the strains of the same source of isolation (Fig. 4). We were able to find a few weak correlations between the origin of isolation and sugar utilization genomic repertoires; 5 out of the 7 strains bearing the fewest sugar cassettes (28-31) are of human origin (NIZO1838, NIZO2256, NIZO2257, NIZO2258, NIZO3893), while 3 of the 6 strains with the most sugar cassettes (50-51) were isolated from silage (NIZO2726, NIZO2831, JDM1) (Table D12 available at http://igfl.ens-lyon.fr/equipes/f.-leulier-functional-genomics-of-host-intestinal-bacteria-interactions/l-plantarum/

database-for-comparative-genomics-of-54-l-plantarum-strains-7-7). Therefore, no clear correlation between sugar metabolism and environmental origin could be found in *L. plantarum*. Most of the sugar utilization cassettes are spread among the genomes of strains with different origins. Analogously, the genetic divergence of the genes assigned to the sugar utilization cassettes did not segragate the strains according to their origin of isolation (Supporting information Fig. S7).Overall, in contrast to the general concept proposed for several LAB species, our study based on 54 *L. plantarum* strains strongly suggest that both the presence/absence of genetic cassettes involved in sugar utilization and their genetic divergence do not reflect strain adaptation to a particular habitat involving utilizing specific substrates for growth. The ability of most *L. plantarum* strains to utilize many different sugars for energetic needs might explain why it is difficult to find environmental determinants in this functional group.

Finally, to measure the degree of congruence between the phylogenetic trees obtained from the 54 *L. plantarum* genomes, we compared the phylogeny of the core genome (Fig. 2) to that of the three variable regions (Supporting information Figs S5–S7). The phylogenetic analyses of the secretome and sugar metabolism functions displayed partial correlation with the core genome phylogeny, whereas the analysis of EPS/CPS showed very limited or no congruency with the core genome phylogeny (Fig. 5), indicating a high degree of genomic plasticity of this functional category.

Taken together, none of the genomic clustering approaches used in our studyidentified a genomic signature that reflects the origin of isolation among the of the 54 *L. plantarum* strains. The only apparent exceptions are the 7 pairs of strains that consistently appeared as close relatives in all phylogenetic analyses (NIZO2891 and NIZO2535, NIZO2484 and NIZO2484, NIZO2877 and NIZO2855, NIZO2260 and NIZO1837, NIZO2494 and NIZO2457, NIZO2831 and NIZO2726, NIZO2262 and NIZO3894) and shared their origin of isolation but geographical origins as well (Table 1). Such sharing of geographic origin prevents us from attributing a shared evolutionary history of environmental adaptation to such paring.

Discussion

We report for the first time a comprehensive sequence-based pan-genome analysis of *L. plantarum* strains isolated from different sources with the aim to further investigate their evolution and find potential genomic signatures that may reflect strain-specific environmental adaptation. The processes of bacterial adaption to specific environments have been chacterized in detail (Alm *et al.*, 2006; De Vuyst *et al.*, 2009). During evolution, most bacteria get rid of useless functions and enrich others that can increase fitness and survival in a particular habitat (Cai *et al.*, 2009; Frese *et al.*, 2011; Douillard *et al.*, 2013; Hottes *et al.*, 2013). However, *L. plantarum* does not seem to follow this trend. Unlike most bacterial species, the evolutionary history of *L. plantarum* does not appear to be related to environmental features, such as the habitats where they were isolated from. Our study suggests that *L. plantarum* is a diverse and versatile species that acquires and retains functional capacities independently of its habitat, thus representing a typical example of a nomadic bacterial species. Its evolutionary history appears complex and not related to environmental adaptation: its variable and flexible genetic composition helps the bacteria maintain and employ a 'universal' set of genes to thrive in many different environments. Although the conventional and mainstream understanding dictates that the presence of specific gene cassettes is often encountered in bacterial adaptation to a particular habitat, *L. plantarum* appears to represent an exception. We performed different analyses targeting different genomic features and parameters (GTM, allelic and functional divergence, hierarchical clustering and phylogenetic analysis, using core genome and variome separately) but failed to identify genomic signatures that reflect environmental adaptation. Concerning sugar metabolism, *L. plantarum* can grow on a large variety of carbohydrates, but our results disclosed the uncoupling of sugar metabolism from strain's origin of isolation. This result differs from what has been proposed to be a common evolutionary feature in other species of *Lactobacillus* (Cai *et al.*, 2009; Frese *et al.*, 2011; Douillard *et al.*, 2013). Strains belonging to *L. reuteri* have been shown to follow different trends of genome evolution depending on their source of isolation (i.e., rodents or humans) (Frese *et al.*, 2011). Two distinct geno-phenotypes were identified in *L. rhamnosus* species by Douillard and colleagues (Douillard *et al.*, 2013), although a habitat enrichment was confirmed only for strains isolated from the human intestinal tract by Ceapa and coworkers (Ceapa *et al.*, 2015). Some *Lactobacillus* species isolated from the human vaginal tract, such as *L. crispatus*, *L. gasseri*, *L. jensenii* and *L. iners*, have been reported to contain smaller genomes than those of the non-vaginal species (Mendes-Soares *et al.*, 2014). Caffrey *et al.* also demonstrated fundamentally different evolutionary trends between host-associated species and their free-living relatives (Caffrey *et al.*, 2012). One bacterial species that seems to be closer to *L. plantarum* in terms of its evolutionary relatedness and environmental adaptation is *L. paracasei*. Smokvina and colleagues performed a broad comparative analysis across several *L. paracasei* genomes and also failed to identify specific evolutionary signatures related to environmental adaptation, although the dairy isolates appeared to cluster together. We therefore conclude that *L. plantarum* is likely a nomadic species

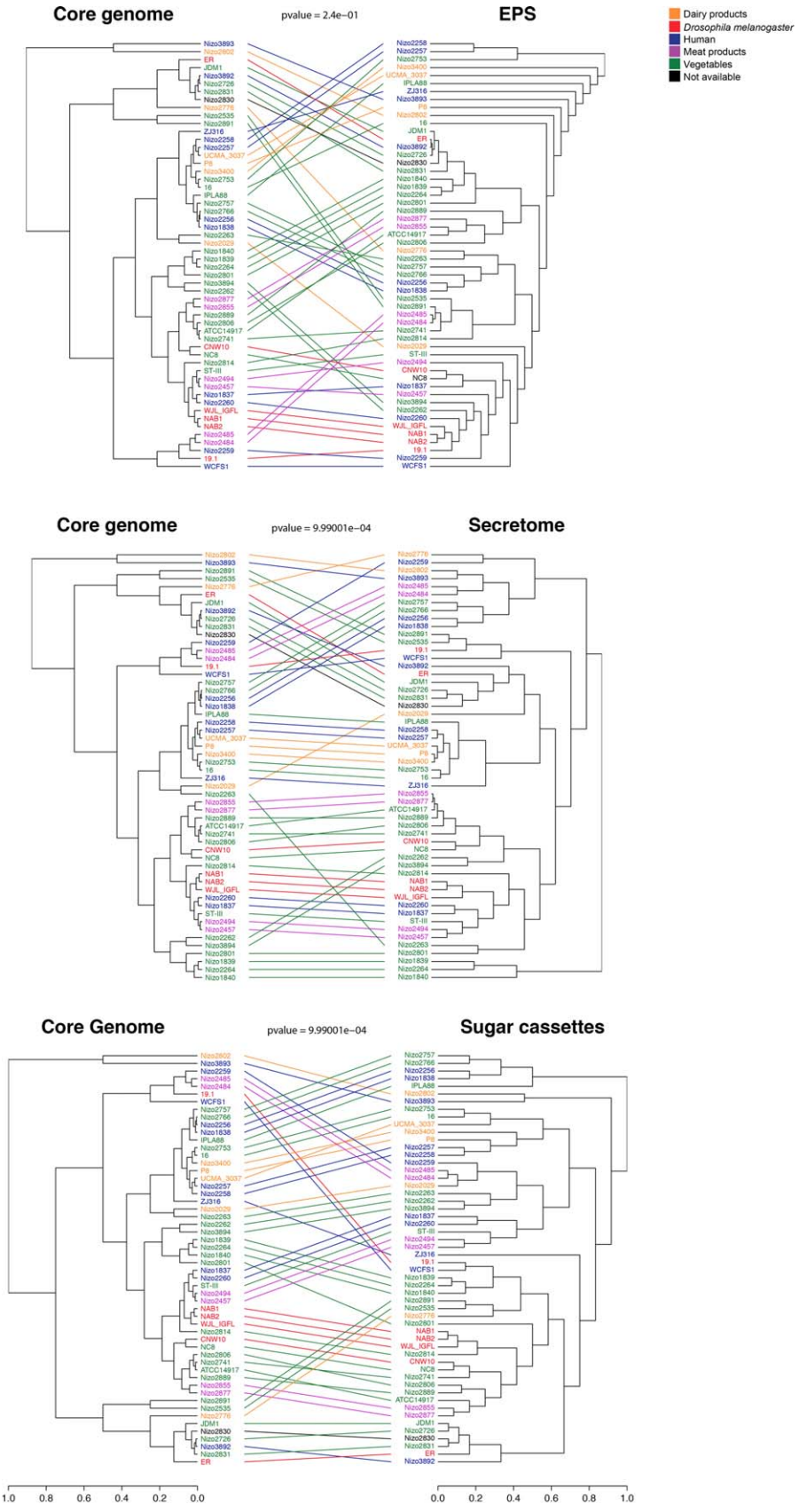

Fig. 5. Comparison of core genome and variome phylogenies.
The phylogenetic tree of the core genome was inferred by using an approximately maximum likelihood method (Price *et al.*, 2010). The phylogenetic trees of the three variome categories (EPS, Secretome, Sugar cassettes) were built using the Neighbor-Joining method (Saitou and Nei, 1987). Coloured strings connecting the same strain of both trees highlight the degree of similarities between the phylogenies. The *p* value reported for each tree represents the result of the permutation test conducted to compare the two methods. Each strain is colour coded by origin of isolation.

like *L. paracasei*, and its genomic adaptation may be driven by alternative selective pressures other than specific environmental adaptation. Nomadic lactobacilli maintain genomic flexibility that enables them to grow efficiently in a variety of environments, while the specialized lactobacilli evolved to an 'evolutionary end' by specializing to a particular habitat. This phenomenon might be explained by the presence of an initial common gene pool shared among the members of a bacterial species, which eventually mutates in strains that go through environmental adaptation. Our findings suggest that for *L. plantarum*, the evolutionary trajectory of the genome cannot be informatively derived from the origin of isolation, thus highlighting the potential capability of *L. plantarum* to frequently migrate across different environments. This feature probably stems from the metabolic flexibility of this bacterial species, which would buffer the selection pressure imposed by a fluctuating environment and permits *L. plantarum* to survive in variable environmental habitats without accumulating massive genetic variations. At the same time, the process of environmental adaptation is very complex and not strictly related to the presence of genetic markers; it might also rely on alternative regulatory mechanisms such as gene expression and protein stability.

Nevertheless, we suspect that some aspects of the approaches adopted in this study may have influenced our findings. First of all, it is always possible to encounter sampling bias in comparative genome studies. We chose to analyse *L. plantarum* strains isolated from the main environments where this species has been commonly found. However, it is important to realize that the isolation of a strain from a particular habitat does not only reflect strain adaptation to that environment. Finding a strain from a given environment might be casual, and does not necessarily indicate that this habitat is the environment to which the species has adapted. For example, the strains isolated from dairy or meat products might have a completely different origins and are occasionally inoculated in those matrices. Also, the plant-derived isolates were not isolated from a single homogenous habitat but instead from several sources which differ in terms of chemical conditions and sugar content (such as fermented fruits, different vegetables and silages). Hence, this might explain why those strains do not share a common genotype. However, we did not identify any specific signatures of environmental isolation in the subgroups belonging to this category. Secondly, so far *L. plantarum* has not been demonstrated to be a host-restricted species, thus implying that the strains isolated from the animals might have another origin. In particular, we only used strains isolated from the intestinal tract for the human-derived strains. The gut is an ecosystem open to food-derived microbes. Therefore, those strains might originate from food matrices. The selection of more specific strains, such as starter cultures used in dairy and meat products and in fermented vegetables, together with strains isolated from human and animal non-gastrointestinal sites could complement our study, and bring an added value in understanding *L. plantarum* genetics and evolution.

To conclude, this study represents a broad, extensive and in-depth comparative genome analysis of 54 *L. plantarum*strains. The analysis of the 54 *L. plantarum* genomes reveals the extreme versatility of this species. The lack of habitat-specialization in these 54 strains reflects the nomadic lifestyle of *L. plantarum*. This might be due to the fact that this species does not persist in specific environments, which in turn prohibits the long-term genomic adaptation and habitat-specialization. Therefore, *L. plantarum* probably has acquired and retained functional capacities that enable it to effectively migrate between different habitats, thus representing a prototypical example of a nomadic bacterial species that is characterized by its dynamic and flexible lifestyle.

Experimental procedures

Lactobacillus plantarum *isolate collection and DNA isolation*

Forty-three *L. plantarum* isolates used in this study were obtained from various institutions and universities (Table 1). Eleven *L. plantarum* genomes, including *L. plantarum* reference strain WCFS1, were obtained from the National Center for Biotechnology Information (NCBI) database (Geer *et al.*, 2010). All isolates were grown in standing MRS broth (Difco Laboratories, BD, US) at 37°C. Genomic DNA from each isolate was extracted using UltraClean Microbial DNA Isolation Kit (MoBio Laboratories, Carlsbad, CA), following manufacturer's instructions.

Genome sequencing and annotation

Genomes of the 38 *L. plantarum* strains obtained from the NIZO culture collection, together with the WJL_IGFL strain, were sequenced using Illumina Miseq technology, while sequencing of 4 *L. plantarum* strains (ER, NAB1, NAB2, 19.1) obtained from different institutions were accomplished using Ion Torrent PGM technology (Life Technologies). The reads of each strain were assembled into contigs using Ray (Boisvert *et al.*, 2012). The RAST annotation server (Aziz *et al.*, 2008) was used to find open reading frames (ORFs) that could code for proteins and to provide an automatic annotation of the encoded functions. A RAST annotation was also done for the 5 *L. plantarum* partial genomes available in the NCBI database to allow for a straightforward comparison of annotations across all genomes. Protein sequences were aligned using blastp version 2.2.25 (Altschul *et al.*, 1990) and alignment results were used by OrthoMCL to group orthologous genes into ortholog groups (OGs) using the inflation parameter of 1.5 (Li *et al.*, 2003). When an OG contained more than one gene per strain (i.e., paralogues), the OG was manually split into separate OGs containing only one gene per strain (except for transposase and mobile elements).The obtained contigs were

aligned and ordered based on OGs using 11 published *L. plantarum* genomes (six complete and five draft genomes) (Table 1) as templates. Whole genome comparisons and pseudogene analyses were completed using the Artemis Comparison Tool (ACT) (Carver *et al.*, 2005) and NCBI BLAST (Altschul *et al.*, 1990). Metabolic pathways were analysed using KEGG pathway (http://www.genome.jp/keg/pathway.html). The CRISPRs Finder tool (http://crispr.u-psud.fr/Server/). The CRISPRs Finder tool (http://crispr.u-psud.fr/Server/) was used to search for CRISPR direct repeats and spacers in the 54 *L. plantarum* strains.

Plasmid prediction

Contigs that represent fragments of putative plasmids were predicted based on one or more of the following criteria: (1) they do not map to the reference chromosomes, (2) they encode typical plasmid functions, (3) they map to published *L. plantarum* plasmids, (4) they have considerably lower GC content (i.e.,<40% GC) than typical *L. plantarum* chromosome (44.5%), (5) they have at least 2× higher sequence coverage than the chromosomal contigs, (6) they appear to be circular, (7) they contain many mobile element proteins (transposases, recombinases, etc.).

Gene trait matching (GTM)

To correlate observed phenotypes with the presence/absence of particular genes and to extend the previous analyses conducted on *L. plantarum* using CGH data (Siezen and van Hylckama Vlieg, 2011), a GTM approach was performed using Phenolink, a web tool that associates bacterial phenotypes to omics data (Bayjanov *et al.*, 2012). Habitat data were divided into 5 classes (vegetables, human, dairy products, meat products, not available). Phenolink performs GTM analyses for the chosen phenotype and generates a single table of OGs correlating to the classes.

Hierarchical clustering, phylogeny and cluster analysis of functional shifts (CAFS)

The phylogenetic analysis of the 16S rRNA gene was conducted using Neighbor-Joining method (Saitou and Nei, 1987). The DNA sequences were aligned using Clustal W (Thompson *et al.*, 1994). Bootstrapping was performed (1000 replicates) to determine the reliability of obtained topologies. All positions containing gaps and missing data were eliminated.

The core genome of the 54 *L. plantarum* strains was analysed using an approximately maximum likelihood tree based on concatenated amino acids that differ between the aligned core proteins. The protein sequences were aligned using blastp (version 2.2.25) (Altschul *et al.*, 1990) and average AAI of these concatenated sequences was calculated using a mean of AAI identity of all blastp hits covering at least 50% length of both query and reference sequences, where query and reference sequences belong to different strains. A phylogenetic tree was created using FastTree2 with its default parameters for amino-acid sequences, such as 1000 bootstrap samples and JTT + CAT amino-acid evolution model with 20 categories (Price *et al.*, 2010).

The hierarchical clustering analyses of the three variable categories (EPS, secretome, sugar cassettes) were generated using Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm. The presence/absence of the OGs belonging to the three variome categories (EPS, secretome and sugar cassettes) has been used as input data (0: absent OG, 1: present OG). Cluster analysis was conducted using DendroUPGMA (http://genomes.urv.cat/UPGMA/) applying Jaccard's coefficient with default settings to compare between the set of variables (Garcia-Vallvé *et al.*, 1999).

The phylogenetic analyses of the three variable categories were conducted using the concatenated sequences of the OGs belonging to those categories that were shared across the 54 *L. plantarum* strains as input. The trees were built using the Neighbour-Joining method with translated amino-acid sequences (Saitou and Nei, 1987) in MEGA v7 software (Kumar *et al.*, 2016). Genetic distances were computed using the Kimura two-parameter model (Kimura, 1980). A bootstrap test was used to test the reliability of trees (Felsenstein, 1985).

Correlation tests between trees have been performed using the R package *dendextend* and its function *cor_corphenetic* that gives the cophenetic correlation coefficient for two trees. A dedicated R script has been created to perform a permutation test and obtain a pvalue (<0.05) indicating if the given coefficient could have been obtained by chance.

The analysis of functional divergence across the 54 *L. plantarum* strains has been conducted using CAFS software (Caffrey *et al.*, 2012). As a first step the orthologous groups were selected and a dendrogram was calculated for each gene. Consequently, functional divergence is identified as the potential departure of the derived protein function from its ancestral one as a result of amino acid changes at important functional sites (Caffrey *et al.*, 2012). The software identifies amino acid positions within a protein, which show radical and statistically significant substitutions among clusters. This analysis was conducted on each tree by comparing the amino acid composition between two clusters to that of an outgroup using a BLOSUM62 amino acid substitution matrix (Henikoff and Henikoff, 1992). Next, the software performs enrichment tests to identify strains and categories of genes that went through significantly more (enriched) or significantly less (impoverished) functional divergence compared with a background level. The enrichment status of each category is then calculated. Finally, a heatmap is generated from the enrichment status of functional categories within strains.

Nucleotide sequence accession numbers

All DNA sequences were deposited at DDBJ/ENA/GenBank under the following accession numbers LUXM00000000, LUWN00000000, LUXL00000000, LUXN00000000, LUXO00000000, LTAU00000000, LUWA00000000, LUWB00000000, LUWC00000000, LUWD00000000, LUWE00000000, LUWF00000000, LUWG00000000, LUWH00000000, LUWI00000000, LUWJ00000000, LUWK00000000, LUWL00000000, LUWM00000000, LUWO00000000, LUWP00000000, LUWQ00000000,

LUWR00000000, LUWS00000000, LUWT00000000, LUWU00000000, LUWV00000000, LUWW00000000, LUWX00000000, LUWY00000000, LUWZ00000000, LUXA00000000, LUXB00000000, LUXC00000000, LUXD00000000, LUXE00000000, LUXF00000000, LUXG00000000, LUXH00000000, LUXI00000000, LUXJ00000000, LUXK00000000.

Acknowledgements

The authors are grateful to Dr Roland Siezen for his help during genomes annotations and comparative analysis, to Dr Tom Delmont for the help in using Anvi'o software and to Dr Dali Ma for proofreading. The work was funded by an ERC starting grant (FP7/2007-2013-N°309704). MEM was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement N°659510. The lab of FL is sponsored by the EMBO YIP program, the ATIP/AVENIR program, the fondation FINOVI and the 'Fondation Schlumberger pour l'Education et la Recherche'. The authors declare no competing interests.

References

Alm, E., Huang, K., and Arkin, A. (2006) The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLoS Comput Biol* **2:** e143.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Axelsson, L., Rud, I., Naterstad, K., Blom, H., Renckens, B., Boekhorst, J., *et al.* (2012) Genome sequence of the naturally plasmid-free *Lactobacillus plantarum* strain NC8 (CCUG 61730). *J Bacteriol* **194:** 2391–2392.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., *et al.* (2008) The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9:** 75.

Bayjanov, J.R., Molenaar, D., Tzeneva, V., Siezen, R.J., and van Hijum, S.A.F.T. (2012) PhenoLink—a web-tool for linking phenotype to ∼omics data for bacteria: application to gene–trait matching for *Lactobacillus plantarum* strains. *BMC Genomics* **13:** 170.

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., and Corbeil, J. (2012) Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol* **13:** R122.

Breiman, L. (2001) Random forests. *Mach Learn* **45:** 5–32.

Bringel, F., Curk, M.C., and Hubert, J.C. (1996) Characterization of lactobacilli by Southern-type hybridization with a *Lactobacillus plantarum* pyrDFE probe. *Int J Syst Bacteriol* **46:** 588–594.

Broderick, N.A., Buchon, N., and Lemaitre, B. (2014) Microbiota-induced changes in *Drosophila melanogaster* host gene expression and gut morphology. *MBio* **5:** e01117–14.

Caffrey, B.E., Williams, T.A., Jiang, X., Toft, C., Hokamp, K., and Fares, M.A. (2012) Proteome-wide analysis of functional divergence in bacteria: exploring a host of ecological adaptations. *PLoS One* **7:** e35659.

Cai, H., Thompson, R., Budinich, M.F., Broadbent, J.R., and Steele, J.L. (2009) Genome sequence and comparative genome analysis of *Lactobacillus casei*: insights into their niche-associated evolution. *Genome Biol Evol* **1:** 239–257.

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G., and Parkhill, J. (2005) ACT: the Artemis Comparison Tool. *Bioinformatics* **21:** 3422–3423.

Ceapa, C., Lambert, J., van Limpt, K., Wels, M., Smokvina, T., Knol, J., and Kleerebezem, M. (2015) Correlation of *Lactobacillus rhamnosus* genotypes and carbohydrate utilization signatures determined by phenotype profiling. *Appl Environ Microbiol* **81:** 5458–5470.

Connelly, P. (2008) *Lactobacillus plantarum*—a literature review of therapeutic benefits. *J Aust Tradit Med Soc* **14:** 79–82.

Crowley, S., Bottacini, F., Mahony, J., and van Sinderen, D. (2013) Complete genome sequence of *Lactobacillus plantarum* strain 16, a broad-spectrum antifungal-producing lactic acid bacterium. *Genome Announc* **1:** pii: e00533–13

Curk, M.C., Hubert, J.C., and Bringel, F. (1996) Lactobacillus paraplantarum sp. now., a new species related to *Lactobacillus plantarum*. *Int J Syst Bacteriol* **46:** 595–598.

De Vuyst, L., Vrancken, G., Ravyts, F., Rimaux, T., and Weckx, S. (2009) Food Microbiology. *Food Microbiology* **26:** 666–675.

Douillard, F.P., Ribbera, A., Kant, R., Pietilä, T.E., Järvinen, H.M., Messing, M., *et al.* (2013) Comparative genomic and functional analysis of 100 *Lactobacillus rhamnosus* strains and their comparison with strain GG. *PLoS Genet* **9:** e1003683.

Dutilh, B.E., Backus, L., Edwards, R.A., Wels, M., Bayjanov, J.R., and van Hijum, S.A.F.T. (2013) Explaining microbial phenotypes on a genomic scale: GWAS for microbes. *Brief Funct Genomics* **12:** 366–380.

Eren, A.M., Esen, ÖC., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont, T.O. (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3:** e1319.

Erkosar, B., Storelli, G., Mitchell, M., Bozonnet, L., Bozonnet, N., and Leulier, F. (2015) Pathogen virulence impedes mutualist-mediated enhancement of host juvenile growth via inhibition of protein digestion. *Cell Host Microbe* **18:** 445–455.

Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39:** 783–791.

Frese, S.A., Benson, A.K., Tannock, G.W., Loach, D.M., Kim, J., Zhang, M., *et al.* (2011) The evolution of host specialization in the vertebrate gut symbiont *Lactobacillus reuteri*. *PLoS Genet* **7:** e1001314.

Frese, S.A., MacKenzie, D.A., Peterson, D.A., and Schmaltz, R. (2013) Molecular characterization of host-specific biofilm formation in a vertebrate gut symbiont. *PLoS Genet* **9:** e1004057.

Garcia-Vallvé, S., Palau, J., and Romeu, A. (1999) Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in *Escherichia coli* and *Bacillus subtilis*. *Mol Biol Evol* **16:** 1125–1134.

Geer, L.Y., Marchler-Bauer, A., Geer, R.C., Han, L., He, J., He, S., *et al.* (2010) The NCBI BioSystems database. *Nucleic Acids Res* **38:** D492–D496.

Giri, S.S., Sukumaran, V., and Oviya, M. (2013) Potential probiotic *Lactobacillus plantarum* VSG3 improves the growth, immunity, and disease resistance of tropical freshwater fish, *Labeo rohita*. *Fish Shellfish Immunol* **34:** 660–666.

Henikoff, S., and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89:** 10915–10919.

Hottes, A.K., Freddolino, P.L., Khare, A., Donnell, Z.N., Liu, J.C., and Tavazoie, S. (2013) Bacterial adaptation through loss of function. *PLoS Genet* **9:** e1003617.

Kafsi, E.H., Binesse, J., Loux, V., Buratti, J., Boudebbouze, S., Dervyn, R., *et al*. (2014) *Lactobacillus delbrueckii* ssp. lactis and ssp. bulgaricus: a chronicle of evolution in action. *BMC Genomics* **15:** 407.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16:** 111–120.

Klarin, B., Molin, G., Jeppsson, B., and Larsson, A. (2008) Use of the probiotic *Lactobacillus plantarum* 299 to reduce pathogenic bacteria in the oropharynx of intubated patients: a randomised controlled open pilot study. *Crit Care* **12:** R136.

Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O.P., Leer, R., *et al*. (2003) Complete genome sequence of *Lactobacillus plantarum* WCFS1. *Proc Natl Acad Sci U S A* **100:** 1990–1995.

van Kranenburg, R., Golic, N., Bongers, R., Leer, R.J., de Vos, W.M., Siezen, R.J., and Kleerebezem, M. (2005) Functional analysis of three plasmids from *Lactobacillus plantarum*. *Appl Environ Microbiol* **71:** 1223–1230.

Kumar, S., Stecher, G., and Tamura, K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol Biol Evol* **33:** 1870–1874.

Ladero, V., Alvarez-Sieiro, P., Redruello, B., Del Rio, B., Linares, D.M., Martin, M.C., *et al*. (2013) Draft genome sequence of *Lactobacillus plantarum* strain IPLA 88. *Genome Announc* **1:** e00524–13

de Las Rivas, B., Marcobal, A., and Muñoz, R. (2005) Development of a multilocus sequence typing method for analysis of *Lactobacillus plantarum* strains. *Microbiology* **152:** 85–93.

Li, L., Stoeckert, C.J., and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13:** 2178–2189.

Li, X., Gu, Q., Lou, X., Zhang, X., Song, D., Shen, L., and Zhao, Y. (2013) Complete genome sequence of the probiotic *Lactobacillus plantarum* strain ZJ316. *Genome Announc* **1:** e0009413i.

Liaw, A., and Wiener, M. (2002) Classification and regression by randomForest. *R news* **2 3:** 18–22.

Lukjancenko, O., Ussery, D.W., and Wassenaar, T.M. (2012) Comparative genomics of *Bifidobacterium*, *Lactobacillus* and related probiotic genera. *Microb Ecol* **63:** 651–673.

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., *et al*. (2006) Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci U S A* **103:** 15611–15616.

Martino, M.E., Bayjanov, J.R., Joncour, P., Hughes, S., Gillet, B., Kleerebezem, M., *et al*. (2015a) Nearly complete genome sequence of *Lactobacillus plantarum* strain NIZO2877. *Genome Announc* **3:** e01370–15.

Martino, M.E., Bayjanov, J.R., Joncour, P., Hughes, S., Gillet, B., Kleerebezem, M., *et al*. 2015b) Resequencing of the *Lactobacillus plantarum* strain WJL genome. *Genome Announc* **3:** e01382–15

Mendes-Soares, H., Suzuki, H., Hickey, R.J., and Forney, L.J. (2014) Comparative functional genomics of *Lactobacillus* spp. reveals possible mechanisms for specialization of vaginal lactobacilli to their environment. *J Bacteriol* **196:** 1458–1470.

Molenaar, D., Bringel, F., Schuren, F.H., de Vos, W.M., Siezen, R.J., and Kleerebezem, M. (2005) Exploring *Lactobacillus plantarum* genome diversity by using microarrays. *J Bacteriol* **187:** 6119–6127.

Naz, S., Tareb, R., Bernardeau, M., Vaisse, M., Lucchetti-Miganeh, C., Rechenmann, M., and Vernoux, J.P. (2013) Genome sequence of *Lactobacillus plantarum* strain UCMA 3037. *Genome Announc* pii: e00251–13.

Newell, P.D., Chaston, J.M., Wang, Y., Winans, N.J., Sannino, D.R., Wong, A.C., *et al*. (2014) *In vivo* function and comparative genomic analyses of the *Drosophila* gut microbiota identify candidate symbiosis factors. *Front Microbiol* **5:** 576.

Nishitani, Y., Sasaki, E., Fujisawa, T., and Osawa, R. (2004) Genotypic analyses of lactobacilli with a range of tannase activities isolated from human feces and fermented foods. *Syst Appl Microbiol* **27:** 109–117.

Oh, P.L., Benson, A.K., Peterson, D.A., Patil, P.B., Moriyama, E.N., Roos, S., and Walter, J. (2010) Diversification of the gut symbiont *Lactobacillus reuteri* as a result of host-driven evolution. *ISME J* **4:** 377–387.

Pepe, O., Blaiotta, G., Anastasio, M., Moschetti, G., Ercolini, D., and Villani, F. (2004) Technological and molecular diversity of *Lactobacillus plantarum* strains isolated from naturally fermented sourdoughs. *Syst App Microbiol* **27:** 443–453.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5:** e9490.

Reller, L.B., Weinstein, M.P., and Petti, C.A. (2007) Detection and identification of microorganisms by gene amplification and sequencing. *Clin Infect Dis* **44:** 1108–1114.

Saitou, N., and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406–425.

Schwarzer, M., Makki, K., and Storelli, G. (2016) *Lactobacillus plantarum* strain maintains growth of infant mice during chronic undernutrition. *Science* **351:** 854–857.

Sharon, G., Segal, D., Ringo, J.M., Hefetz, A., Zilber-Rosenberg, I., and Rosenberg, E. (2010) Commensal bacteria play a role in mating preference of *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* **107:** 20051–20056.

Siezen, R.J., and van Hylckama Vlieg, J.E.T. (2011) Genomic diversity and versatility of *Lactobacillus plantarum*, a natural metabolic engineer. *Microb Cell Fact* **10:** S3.

Siezen, R.J., Francke, C., Renckens, B., Boekhorst, J., Wels, M., Kleerebezem, M., and van Hijum, S.A.F.T. (2011) Complete resequencing and reannotation of the *Lactobacillus plantarum* WCFS1 genome. *J Bacteriol* **194:** 195–196.

Siezen, R.J., Tzeneva, V.A., Castioni, A., Wels, M., Phan, H.T.K., Rademaker, J.L.W., *et al*. (2010) Phenotypic and genomic diversity of *Lactobacillus plantarum*strains isolated from various environmental niches. *Environ Microbiol* **12:** 758–773.

Smokvina, T., Wels, M., Polka, J., Chervaux, C., Brisse, S., Boekhorst, J., *et al*. (2013) *Lactobacillus paracasei* comparative genomics: towards species pan-genome definition and exploitation of diversity. *PLoS One* **8:** e68731.

Storelli, G., Defaye, A., Erkosar, B., Hols, P., Royet, J., and Leulier, F. (2011) *Lactobacillus plantarum* promotes *Drosophila* systemic growth by modulating hormonal signals through TOR-dependent nutrient sensing. *Cell Metab* **14:** 403–414.

Sun, Z., Harris, H.M.B., McCann, A., Guo, C., Argimón, S., Zhang, W., *et al.* (2015) Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nat Commun* **6:** 8322.

Tatusova, T., Ciufo, S., Fedorov, B., O'neill, K., and Tolstoy, I. (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res* **42:** D553–D559.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix. *Nucleic Acids Res* **22:** 4673–4680.

Torriani, S., Clementi, F., Vancanneyt, M., Hoste, B., Dellaglio, F., and Kersters, K. (2001) Differentiation of *Lactobacillus plantarum*, L. pentosus and L. paraplantarum species by RAPD-PCR and AFLP. *Syst Appl Microbiol* **24:** 554–560.

Vermeiren, L., Devlieghere, F., and Debevere, J. (2003) Evaluation of meat born lactic acid bacteria as protective cultures for the biopreservation of cooked meat products. *Int J Food Microbiol* **96:** 149–164.

Wang, Y., Chen, C., Ai, L., Zhou, F., Zhou, Z., Wang, L., *et al.* (2011) Complete genome sequence of the probiotic *Lactobacillus plantarum* ST-III. *J Bacteriol* **193:** 313–314.

Yang, K.M., Jiang, Z.Y., Zheng, C.T., Wang, L., and Yang, X.F. (2014) Effect of *Lactobacillus plantarum* on diarrhea and intestinal barrier function of young piglets challenged with enterotoxigenic *Escherichia coli* K88. *J Anim Sci* **92:** 1496–1503.

Zhang, Z.Y., Liu, C., Zhu, Y.Z., Zhong, Y., Zhu, Y.Q., Zheng, H.J., *et al.* (2009) Complete genome sequence of *Lactobacillus plantarum* JDM1. *J Bacteriol* **191:** 5020–5021.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site.

Fig. S1. Pan-genome and core-genome size. The number of pan-genome OGs (A) and core genome OGs (B) is shown as a function of genomes added to the pan-genome.
Fig. S2. Cluster of orthologous groups belonging to the core genome and their relative presence across the 54 *L. plantarum* strains.
Fig. S3. OGs number per strain.
Fig. S4. 16S rRNA gene phylogeny. Phylogenetic analysis of the 16S rRNA gene of the 54 *L. plantarum* strains. The evolutionary history of the 16S rRNA gene was inferred using the Neighbor-Joining method (Saitou and Nei, 1987). The optimal tree with the sum of branch length- = 0.00502859 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown on nodes scoring >50% (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Each strain is color coded by origin of isolation.
Fig. S5. Phylogenetic tree of OGs belonging to EPS functional category. The 54 *L. plantarum* strains were clustered based on the nucleotide sequences of the genes belonging to the EPS category shared among all strains.The tree was built using the Neighbor-Joining method (Saitou and Nei, 1987).The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown on nodes scoring >50% (Felsenstein, 1985). The heatmap next to the tree shows the distribution of the OGs belonging to the EPS functional category across the 54 *L. plantarum* strains. Each row corresponds to one strain and each column corresponds to one OG. Each strain is color coded by origin of isolation.
Fig. S6. Phylogenetic tree of OGs belonging to secretome functional category. The 54 *L. plantarum* strains were clustered based on the nucleotide sequences of the genes belonging to the secretome category shared among all strains. The tree was built using the Neighbor-Joining method (Saitou and Nei, 1987). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown on nodes scoring >50% (Felsenstein, 1985). The heatmap next to the tree shows the distribution of the OGs belonging to the secretome functional category across the 54 *L. plantarum* strains. Each row corresponds to one strain and each column corresponds to one OG. Each strain is color coded by origin of isolation.
Fig. S7. Phylogenetic tree of OGs belonging to sugar metabolism functional category. The 54 *L. plantarum* strains were clustered based on the nucleotide sequences of the genes belonging to the sugar metabolism category shared among all strains. The tree was built using the Neighbor-Joining method (Saitou and Nei, 1987).The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown on nodes scoring >50% (Felsenstein, 1985). The heatmap next to the tree shows the distribution of the OGs belonging to the sugar metabolism functional category across the 54 *L. plantarum* strains. Each row corresponds to one strain and each column corresponds to one OG. Each strain is color coded by origin of isolation.
Table S1. Sequencing statistics of the *L. plantarum* strains sequenced in this study.
Table S2. List of the 4137 novel OGs that are not present in *L. plantarum* WCFS1 reference genome.
Table S3. Gene-trait matching analysis by Random Forest classification. Genotype-phenotype linkage analysis on origin of isolation.
Table S4. List of the OGs belonging to the enriched COGs resulted from CAFS analysis.