The Variable Regions of *Lactobacillus rhamnosus* Genomes Reveal the Dynamic Evolution of Metabolic and Host-Adaptation Repertoires

Corina Ceapa[1,2,*], Mark Davids[3], Jarmo Ritari[4], Jolanda Lambert[1], Michiel Wels[5], François P. Douillard[6], Tamara Smokvina[7], Willem M. de Vos[2,4,6], Jan Knol[1,2], and Michiel Kleerebezem[8]

[1]Gut Biology & Microbiology Platform, Nutricia Research Centre, Utrecht, the Netherlands

[2]Laboratory of Microbiology, Wageningen University, the Netherlands

[3]Laboratory of Systems and Synthetic Biology, Wageningen University, the Netherlands

[4]RPU Immunobiology, Department of Bacteriology and Immunology, University of Helsinki, Finland

[5]NIZO Food Research, Ede, the Netherlands

[6]Department of Veterinary Biosciences, University of Helsinki, Finland

[7]Danone Research, Palaiseau, France

[8]Host-Microbe Interactomics Group, Department of Animal Sciences, Wageningen University and Research Centre, Wageningen, the Netherlands

*Corresponding author: E-mail: corinaceapa@uchicago.edu.

Abstract

Lactobacillus rhamnosus is a diverse Gram-positive species with strains isolated from different ecological niches. Here, we report the genome sequence analysis of 40 diverse strains of *L. rhamnosus* and their genomic comparison, with a focus on the variable genome. Genomic comparison of 40 *L. rhamnosus* strains discriminated the conserved genes (core genome) and regions of plasticity involving frequent rearrangements and horizontal transfer (variome). The *L. rhamnosus* core genome encompasses 2,164 genes, out of 4,711 genes in total (the pan-genome). The accessory genome is dominated by genes encoding carbohydrate transport and metabolism, extracellular polysaccharides (EPS) biosynthesis, bacteriocin production, pili production, the *cas* system, and the associated clustered regularly interspaced short palindromic repeat (CRISPR) loci, and more than 100 transporter functions and mobile genetic elements like phages, plasmid genes, and transposons. A clade distribution based on amino acid differences between core (shared) proteins matched with the clade distribution obtained from the presence–absence of variable genes. The phylogenetic and variome tree overlap indicated that frequent events of gene acquisition and loss dominated the evolutionary segregation of the strains within this species, which is paralleled by evolutionary diversification of core gene functions. The CRISPR-Cas system could have contributed to this evolutionary segregation. *Lactobacillus rhamnosus* strains contain the genetic and metabolic machinery with strain-specific gene functions required to adapt to a large range of environments. A remarkable congruency of the evolutionary relatedness of the strains' core and variome functions, possibly favoring interspecies genetic exchanges, underlines the importance of gene-acquisition and loss within the *L. rhamnosus* strain diversification.

Key words: comparative genomics, diversity, niche adaptation, probiotic, core and pan-genome.

Background

The genus *Lactobacillus* encompasses a phylogenetically diverse group of bacteria and currently amounts to more than 200 recognized species. Genomic approaches have provided insight in the evolutionary relationships among the *Lactobacillus* species, and have also revealed some genes and functions that explain their presence in particular ecological niches like the gastrointestinal tract (Lukjancenko et al. 2012), fermented dairy products (Broadbent et al. 2012), and plant-associated environments (Mahony et al. 2012). A combination of gene gain and loss plays a prominent role during niche adaptation, where particularly gene loss appears to be associated with the adaptation to the nutrient-rich dairy environment (Lukjancenko et al. 2012).

Together with *Lactobacillus plantarum, Lactobacillus reuteri*, and *Lactobacillus casei, Lactobacillus rhamnosus* has the largest genome among the species of the *Lactobacillus* genus or the lactic acid bacteria (LAB) in general. *Lactobacillus rhamnosus* is an anaerobic, facultative heterofermentative rod-shaped microorganism that is encountered in the fermentation of food and feed-raw materials (Bergey 2009), but also frequently found in the human and animal intestinal (Heilig et al. 2002), oral and uro-genital system (Hummelen and Fernandes 2010).

Among its most studied strains, *L. rhamnosus* GG (ATCC 53103) stands apart as the most documented probiotic strain (887 scientific articles and 58 patents by 2015—according to the Scopus Database). Different health and industrial benefits were associated with different strains of *L. rhamnosus*, but due to the lack of controlled comparative studies, it remains to be established if and to what extent all these effects are strain-specific (Douillard, Ribbera, Järvinen, et al. 2013).

While for some of the health benefits and niche adaptation factors, the mechanism of action of *L. rhamnosus* strains remains unclear, some of the molecular mechanisms that may underlay or contribute to these effects have been studied (van Bergenhenegouwen et al. 2014). For example, expression of mucus-binding pili in the intestine (Kankainen et al. 2009) and mucus-binding factor (MBF), an active mucus-specific surface adhesin (von Ossowski et al. 2011), play a role in the adherent mechanisms that contribute to intestinal colonization by *L. rhamnosus* GG. In addition, muramidases p40 and p75 were shown to inhibit apoptosis in mouse colon epithelial cells and cultured mouse colon explants treated with TNF-α, thereby positively influencing epithelial layer integrity and repair (Bäuerl et al. 2010), which could contribute to reinforcement of the epithelial barrier (Nermes et al. 2011). Moreover, the large surface protein MabA was shown to play a role in epithelial adhesion of *L. rhamnosus* cells, and its expression correlated with increased levels of bacterial adhesion to intestinal epithelial cells and tissues of the murine GIT (Vélez et al. 2010).

The present mechanistic insights do not fully explain the complex interaction that *L. rhamnosus* strains develop with their host. In this context, a detailed pan-genomic analysis of a substantial number of *L. rhamnosus* strains might potentially provide molecular discriminators that can be associated with functional variations within the species, including their differential capacity to affect the host's (mucosal) physiology. The previously reported sequence scanning of 100 strains of *L. rhamnosus* allowed the determination of the core genome of the *L. rhamnosus* species at high resolution (Douillard, Ribbera, Kant, et al. 2013). However, the exclusive mapping of the sequence information onto the *L. rhamnosus* GG template genome restricted the full species diversity analysis to functions that are present in *L. rhamnosus* GG.

Since the variable traits of the species, which are absent from *L. rhamnosus* GG can provide novel insights into the species' diversity, this study presents the in silico genomic and metabolic analysis of the diversity of the *L. rhamnosus* species using the genomic sequences of 40 strains of the species, that were isolated from various environmental niches. The main focus of the comparative genome analysis was targeted to the analysis of the distribution and function of the "variome" of the *L. rhamnosus* species, that is, the genes that do not belong to the previously reported core genome.

Materials and Methods

Functional Annotation

Strain Selection, Genome Sequencing, and Annotation

Thirty-six different *L. rhamnosus* strains from the Helsinki University and the Danone Nutricia Research Collection were selected based on origin information or AFLP classification. The strain selection aimed to cover as much as possible the diversity inherent to the species. The type strain ATCC7649 is part of the group of newly sequenced strains. A complete list of the selected *L. rhamnosus* strains and their origin can be found in supplementary table S1, Supplementary Material online. In addition, four public strains genomes (strains GG, HN001, Lc705, ATCC 21052) were downloaded and processed with the same bioinformatics pipeline as the newly sequenced genomes.

For DNA preparation, 2 ml of overnight culture was pelleted, washed and resuspended in 20 mM Tris-HCl, pH 8.0, 1 mM ethylenediaminetetraacetic acid, pH 8.0, 8% sucrose, 50 mM sodium chloride (TES) buffer. Cell lysis was performed with lysozyme (360 mg/ml) and mutanolysin (140 U/ml) during 2 h at 37 °C, then 300 µl water was added and 80 µl of 20% sodium dodecyl sulfate solution. DNA was extracted using phenol/chloroform (1:1) (3×). The DNA was precipitated with isopropanol and washed with 70% ethanol. RNAse treatment was performed using 100 ug/ml RNAse (Sigma) during 1 h at 37 °C.

Draft genome sequences of 36 *L. rhamnosus* strains were obtained (GATC Biotech, Germany) using 454 GS FLX sequencing at a sequence coverage ranging from 9 to 22×

(see complete sequencing statistics in supplementary table S1, Supplementary Material online). Genomes assembly was performed using Newbler 2.6 and 2.8 with standard settings. Contig sequences of all strains were annotated using the RAST pipeline (Overbeek et al. 2014). Genome sequences of *L. rhamnosus* strains taken from public databases were reanalyzed (gene assignments and annotation) using the same procedures to ensure consistency of the gene-function predictions for each genome used.

Metabolic Mapping—KEGG

The draft metabolic network and associated gene–reaction relationships of the pan, core and clades' genomes was constructed using a KEGG automatic annotation server (KAAS) (Moriya 2007), according to the KEGG database (Kanehisa et al. 2009) using the standard settings and mapping against 40 typical prokaryotes genomes. Then, manual checking of the metabolic maps obtained from the KAAS annotation identified the relevant metabolic diversity and gaps.

Substrate specificity of PEP group translocation (PTS) transporters was predicted based on homology to annotated PTS genes (the best hit in the NCBI nr database) and based on their genomic context, focusing on genetically linked genes with predicted functions related to enzyme and regulatory functions in carbohydrate metabolism.

Domain Annotation

Identification of proteins containing LPXTG signals (for sortase-dependent genes), mucus-, fibrinogen- or fibrin-binding, collagen- and Ig-like domains was performed using Interproscan (Zdobnov and Apweiler 2001) analysis.

Bacteriocin Detection

BAGEL3 was employed as an automated pipeline for identifying genes encoding class I and II bacteriocins using an identification approach that combines direct mining for the encoding gene and indirect mining of genetic context for accessory functions related to the production of these bacteriocins and related peptides (van Heel et al. 2013).

CRISPR Identification and Characterization

CRISPR (clustered regularly interspaced short palindromic repeat) loci and associated *cas* genes were identified using a combination of homology to previously identified *cas* genes and their corresponding CRISPR repeats (Brouns et al. 2008) and by de novo identification using the CRISPR Recognition Tool (CRT) (Bland et al. 2007). Spacer homologies to foreign genetic elements were assessed using BLASTN (Altschul et al. 1997) on two databases: one created with all *L. rhamnosus* genomes presented in this study and the NCBI complete nucleotide collection nr (Pruitt et al. 2007). Nucleotide conservation between CRISPR spacers and corresponding protospacers

in phages, plasmids, and chromosomal sequences were visualized using Ugene (Okonechnikov et al. 2012).

Genome Comparisons
Orthology Estimation

The orthology prediction was performed for all genomes (5 public and 35 sequenced by 454 GS FLX in this study) using sequence clustering and search program Usearch v. 6.0.307 (Edgar 2010). To define orthologous groups (OGs), all the predicted open reading frame (ORF) amino acid sequences of all the strains were first ordered by length using the "sortbylength" command, and then clustered with the "cluster_smallmem" command using identity threshold of 0.5 and E-value cut-off of $1e - 30$. Thus, the members of each resulting ORF sequence cluster shared at least 50% amino acid identity over their entire sequence lengths based on the longest sequence. The longest sequence in each cluster was chosen as the representative of the OG. The pan-genome was thus defined as the set of representative sequences of the OGs. The ortholog data were managed using R v. 2.15.3 with Biostrings package and in-house scripts.

Since most of the genomes are in a draft state, ORF prediction might have missed protein-coding sequences (CDSs). To detect the eventually missing genes, CDSs were aligned against each of the genomes using tBLASTn. CDSs that had a hit with 95% identity and query length or had the best alignment on the edge of a contig were classified as being present in the corresponding genome.

Search for *L. rhamnosus*-Specific Genes (Orphan OGs)

To identify *L. rhamnosus*-specific genes, we created a database of the protein sequences of all completely sequenced bacterial genomes present in the NCBI nr database. All known *L. rhamnosus* genomes were excluded from this set. The final database composed of 3,505,217 proteins from 1,047 genomes. We compared all the genes belonging to *L. rhamnosus* strains to this database using BLASTP with an initial identity threshold of 0.5 and E-value cut-off of $1e - 30$, the same value used to create the OG matrix. Genes that had no blast hit against any of the proteins in the database were considered to be *L. rhamnosus* specific. The genes specific for the species' core and pan gene sets were analyzed using the same approach.

Variome Classification Based on Presence–Absence of Genes

Genomes and OGs were clustered in R (R Development Core Team 2008) using the complete method based on the Manhattan distance of presence–absence matrix. A subset of 10% of core genes was added to the variable genome in order to identify the least variable OG clusters. The output was

visualized with heatmap.2 from the gplots package (Warnes 2012).

BRIGS Mapping on Reference Public Genomes

To identify the variable regions, all genomes were aligned against the public reference strain GG and LC705 genomes and visualized using BRIG (Alikhan et al. 2011). Rings were color coded according to the variome classification of the strains (supplementary fig. S4, Supplementary Material online). ABC and PTS transporters were visualized by aligning the genes against the genome and taking the best hit. Contigs mapping to multiple regions were all shown in the ring. Repetitive regions in the ring overlap on the reference and therefore do not display the repeat.

Phylogenetic Analysis

We define the phylogenetic relationship of the various strains from the patterns of single-amino acid substitutions of the proteins that belong to the core genome of the species and were present in a single copy in all strains (supplementary table S2, Supplementary Material online). The protein sequences of 1,008 OGs with a single member in each *L. rhamnosus* genome were aligned using MUSCLE (Edgar 2004), which delivered 5127 positions in the core protein sequences with altered amino acid content. These alignments were concatenated after which a maximum-likelihood tree was constructed using PHYML (Guindon et al. 2010).

Results and Discussion

General Features of L. rhamnosus Genomes

Strains of the species *L. rhamnosus* originate from a variety of ecological niches. In this study, we describe the sequencing and comparative genomics of 40 *L. rhamnosus* strains isolated from different sources like dairy fermented products, plants, human and animal intestine and human clinical samples (supplementary table S1, Supplementary Material online). A high number of isolates were obtained from the mammalian intestinal tract environment, that is, there are eight isolated from human feces, three from the intestine of healthy individuals and one from goat feces. In addition, *L. rhamnosus* strains of other human origins were also included, that is, 17 strains from hospitalized patients, including blood (infection) isolates and a vaginal isolate. The group is completed by seven dairy isolates and one beer isolate.

Genome size, G + C content and protein CDS counts vary among strains, independent of the strain's niche of isolation. The average genome size is 3 ±0.2 Mb, thereby placing the *L. rhamnosus* genomes among the largest within the *Lactobacillus* group (Canchaya et al. 2006). The molecular percentage G + C content varied among sequenced genomes from 46.5% to 46.8% (supplementary table S1,

Supplementary Material online). The mean percentage G + C content of 40 strains was comparable to the closely related *L. casei* species (46.6% compared with 46.3%) (Toh et al. 2013) but slightly higher than that of *L. plantarum* (44.5%) (Kleerebezem et al. 2003; Siezen et al. 2012) or the average for the lactobacilli that is estimated at 42.4% (Kant et al. 2011).

The newly sequenced genomes had average sequence coverage of approximately 14-fold (ranging from 9- to 22-fold) and could be assembled into 61–535 contigs (with an average number of contigs of 178). Due to the large number of strains analyzed, the number of gaps of each genome will have a lower influence on the overall OGs estimation (Li et al. 2011). This level of coverage is sufficient for comprehensive and comparative characterization of the pan-genome and variome diversity among these strains, as indicated by the asymptote of the new gene discovery (supplementary fig. S1, Supplementary Material online). A curve reaching saturation indicates there is no need for additional sequencing (Sims et al. 2014), and thus the data set that was generated provides an adequate reflection of the diversity of the species.

Previous work has established the *L. rhamnosus* core genome (Douillard, Ribbera, Kant et al. 2013) by SOLID sequencing 100 *L. rhamnosus* clinical and dairy strains and mapping all reads to the *L. rhamnosus* GG (LGG) reference genome. Only sequences mapping to the LGG genome were analyzed further and the core genome consisted of all genes shared between all 100 strains (~80% of LGG genome). Due to the short read length and the methodology used, the genome collection produced by Douillard, Ribbera, Kant et al. is not suitable for analysis of the variable genome regions. The current work explores the species' variome to identify the basis of strain (and clade) variability, as well as potential evolutionary pathways for the species *L. rhamnosus*, while using the characteristics of the core genome only to highlight the relevance of certain predicted functional variations.

The total predicted genetic content encompassed 4,711 OGs (forming the pan-genome) with an average of 2,707 (±74) OGs per genome, out of which 2,164 are present in all genomes (Douillard, Ribbera, Kant, et al. 2013) (supplementary table S2, Supplementary Material online). The size of the pan-genome is therefore higher than the one reported for other related LAB species: around 2,800 OGs for *Oenococcus oeni* (Borneman et al. 2012), 4,200 OGs for *L. casei* (Broadbent et al. 2012) and *L. paracasei* (Smokvina et al. 2013). In our assessment, the core- and pan-genome estimates almost reach saturation (supplementary fig. S1, Supplementary Material online), implying that the total diversity inherent to the species is well represented within this data set. Putative biological functions were assigned to 2,761 (58%) of the predicted OGs and another 1,272 (27%) OGs share sequence conservation with conserved proteins of
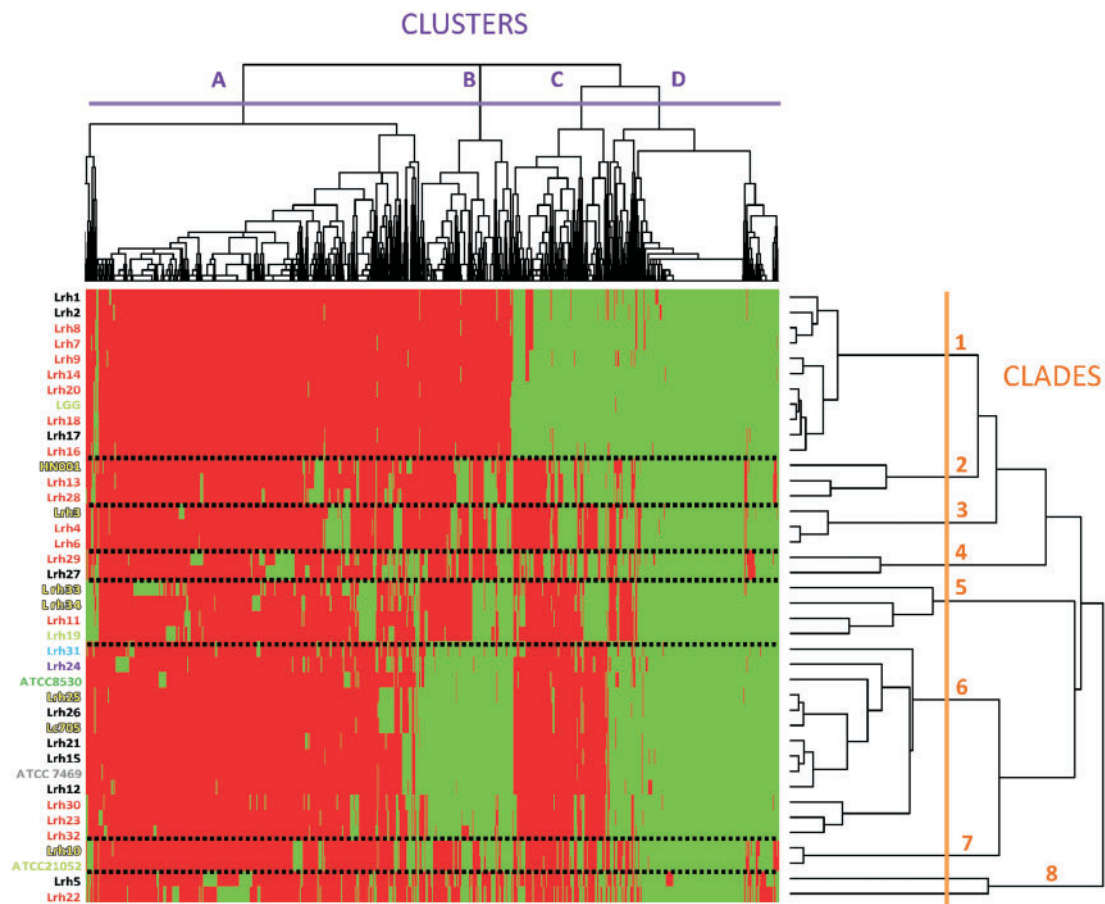
Fɪɢ. 1.—Hierarchical map and clustering of the 40 *L. rhamnosus* genomes based on presence–absence of genes (green, presence and red, absence of genes). OGs of genes could be classified in four groups noted cluster A–D. Cluster D includes 10% of core genes to allow a shared OGs skewed classification of the clusters. At strain level there are eight recognizable clades of strains (1–8). The niche from which the strains were isolated is indicated by the color-coding of the strain's ID-code: yellow: dairy; black: human feces; red: clinical (blood); light green: healthy intestine; blue: vagina; purple: goat feces; dark green: beer; gray: type strain (unknown).

unknown function in other organisms (supplementary table S2, Supplementary Material online). Among the conserved core genome OGs, 67% have their closest relative sequence in *L. casei* and *L. paracasei*, followed by *Lactobacillus zeae* (30%), *Lactobacillus plantarum* (1%), *Lactobacillus pentosus*, *O. oeni*, and *Streptococcus* species' (supplementary table S2, Supplementary Material online), which is in good agreement with previous taxonomic classification of *L. rhamnosus* (Toh et al. 2013).

In brief, the core genome contains genes needed for replication, transcription, translation, central, and cell wall metabolism, biosynthesis of most amino acids and metabolism of nucleotides, fatty acids, and phospholipids. Hypothetical proteins having a yet unknown function amount to more than 682 OGs of the core genome (32% of the core genome). The core genome also contains more than 18 complete sugar utilization gene clusters, and a variety of cell-surface components, discussed in more detail below.

The variable gene content of the analyzed genomes was investigated using the gene presence/absence matrix and visualized by creating a heatmap using R (fig. 1) and used to construct a genome-relatedness or "pan-genome" tree based on variable genome content (variome tree; fig. 1).

The eight clade-classification identifies *L. rhamnosus* unique and clade-specific OGs (supplementary table S3, Supplementary Material online). Species-specific orphans OGs (OOGs) that are uniquely present in *L. rhamnosus* and missing in all other bacteria (based on the entries in the NCBI nr database), amounted to 49% of the *L. rhamnosus* pan-genome. The present study identified 232 OOGs in the *L. rhamnosus* group, among which a substantial amount are small-sized ORFs (more than 10% have less than 75 amino acids), suggesting that the number of OOGs may be somewhat overestimated due to sequencing gaps and/or overprediction of ORFs, also supported by an observed larger size of the pan-genome compared to other LAB species (see above).

Notably, *L. rhamnosus* strains share 37 core OOGs, which consistently lack a functional prediction.

Clade-specific OGs are present in all members of a specific variome-based genetic clade and absent in all other genomes. As can be expected from the variome-based clade prediction (fig. 1), there were large differences between the numbers of clade-specific OGs (in order: clade 1: 18.2%; clade 2:13.0%; clade 3: 4.8%; clade 4:10.0%; clade 5: 4.4%; clade 6: 21.3%; clade 7: 0.8%; and clade 8:27.5%). Clades with a lower number of strains did not necessarily contain fewer specific genes: clades 1 and 4 with 2 and, respectively, 4 strains retained almost the highest number of specific genes among all clades. Clade 7 has the lowest number of clade-specific OGs (7) followed by clade 3 (41), whereas clade 8 has the highest, 235 OGs (supplementary table S4, Supplementary Material online). Global function scanning among the clade-specific gene sets, indicated some clade-specific functional (supplementary table S4, Supplementary Material online) as well as metabolic (supplementary table S8, Supplementary Material online) enrichments: clades 1, 2, 5, and 7, energy metabolism functions (1.34% of total clade genes); clade 1, metabolism of cofactors and vitamins (1.01%) as well as signal transduction (0.56%); clades 3, 6, and 8, carbohydrate metabolism (5.07%), transport (2.34%), and glycan degradation (0.34%); clades 3, 6, and 8, amino acid metabolism (2.88%) and restriction-modification; and clades 2, 3, 5, 7, and 8, membrane transport (2.34%). Mobile elements are abundant in clades 1 (especially phage-associated genes), 6 and 8 but missing from clades 3 and 7, whereas strains of clades 4 and 7 appear to lack plasmids (fig. 4). Notably, the CRISPR-Cas system appeared to be entirely missing from strains in clades 2 and 8 (supplementary fig. S2, Supplementary Material online), whereas this system is consistently present in all other strains of the other clades. Below, we will describe in detail a selection of functional characteristics that appear to be distributed in a clade-specific manner.

The number of strain-specific OGs fluctuates from 0 to 100 per strain and amounted to 706 OGs (15%) in total (supplementary table S6, Supplementary Material online), which is most probably a somewhat overestimated number, due to the inclusion of small ORFs (see also above), and the potential over assignment of ORFs in general. Nevertheless, some of the strain-specific genes are predicted to encode functions that may be acquired recently and/or confer advantages in certain environments or represent specific genetic entities. The strains Lrh22, Lrh33, and Lrh24 appear to encode more than 50 unique genes each, and thereby standout from all other strains analyzed in this study. Notably, in strain Lrh22 among those unique genes is a plasmid replication associated gene, which may imply that some of the other strain-specific genes may also be located on a plasmid. Analogously, other strains (Lrh3, Lrh23, and Lrh32) also appear to contain plasmid replication genes, but no further evidence for presence of plasmids was found. Notably, the only beer isolate, ATCC8530, contains only two strain-specific functions (pediocin pepB, and thioredoxin) that display very high level homology with *L. pentosus* and could play a role in the elevated resistance to hops in the presence of ethanol [79] reported for this strain. Strain-specific genes appear to reflect gene exchanges with other species, like the presence of an apparently complete prophage and an arsenic resistance operon in strain Lrh33, which appear to be almost identical to genetic entities in *L. casei*. Analogously, strain Lrh24 encodes five prophage pi1 genes and a complete polysaccharide cluster (ten genes) with very high identity with genes found in *L. paracasei*.

Genomic Prediction of L. rhamnosus Metabolic Diversity

Efficient use of environmental resources, especially carbohydrates, is relevant for the survival of bacteria in different habitats, including in the intestine, which is characterized by dynamic diet and host-dependent nutrient availability. Metabolism genes assigned within the variome are dominated by carbohydrate transport and associated metabolic pathways, supporting the idea that sugar utilization is important for the evolution and adaptation of *L. rhamnosus* as a species (Douillard, Ribbera, Järvinen, et al. 2013), which is in agreement with observations made in other *Lactobacillus* species such as *L. reuteri* (Hüfner et al. 2008) and *L. casei* (Douillard, Ribbera, Järvinen, et al. 2013) (supplementary table S8, Supplementary Material online).

Carbohydrate Import and Utilization

One of the largest functional classes encoded by *L. rhamnosus* genomes—transport proteins (in total 418 OGs)—is dominated by proteins predicted to be involved in carbohydrate import (44% of transporters with a predicted substrate) (fig. 2, supplementary table S8, Supplementary Material online). Notably, for 68 transport-associated OGs (62 ABC and 6 PTS OGs) no substrate specificity could be predicted, and these functions could further expand the substrate utilization capacities of *L. rhamnosus* strains.

The average number of transport-associated genes per genome is 315 ± 12, a number comparable to the number of transport functions encountered in the genome of the versatile *L. plantarum* WCFS1 (Kleerebezem et al. 2003). A large fraction of these transporters (~40%) have a clear distribution pattern among the different strains and clades, enabling the discrimination of clades on basis of these functions (supplementary table S7, Supplementary Material online). This observation is congruent with other studies that already identified strain-specific carbohydrate utilization capacities as a means of functional classification of *L. rhamnosus* strains (Ceapa et al. 2015). In the genome of strain GG (Morita et al. 2009), the
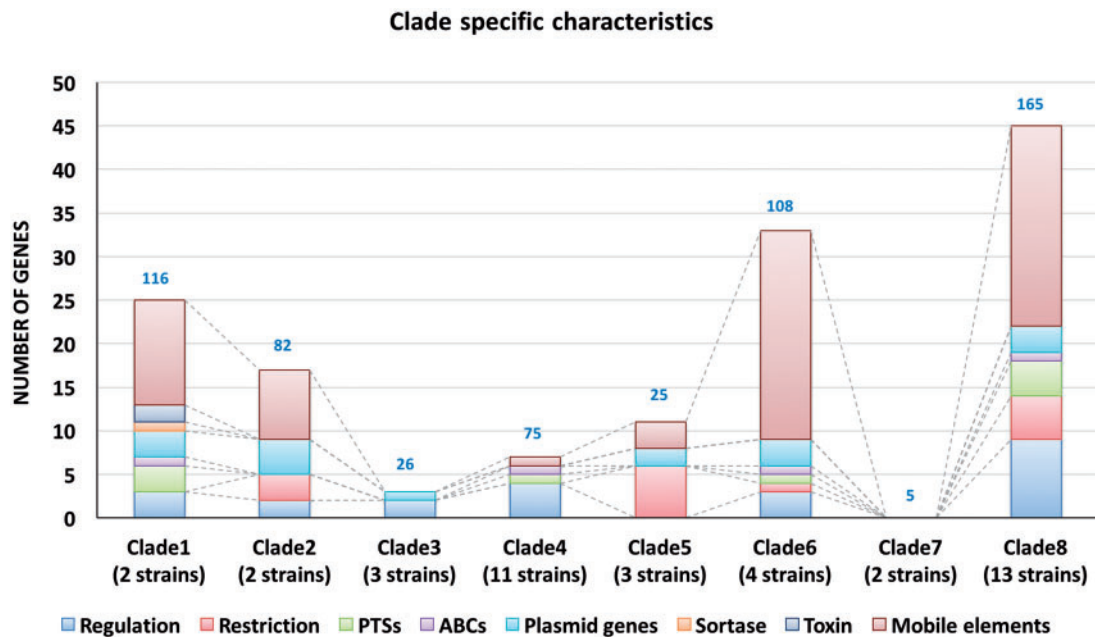
FIG. 2.—Number of clade-specific genes and their functionality. Annotated genes are displayed; the number of clade-specific genes with unknown functionality is added on top of each bar.

genes encoding ATP-binding cassette (ABC) transporters appear to be evenly distributed across the bacterial chromosome, whereas a regional enrichment is observed for phosphotransferase systems (PTSs), encoding carbohydrate transport functions, in the region at the 5′-side of the origin of replication (supplementary fig. S3, Supplementary Material online). This functional enrichment in a particular region of the chromosome is analogous to the positioning of the so-called "sugar-island" of *L. plantarum* WCFS1 (Kleerebezem et al. 2003). These regions also display a high degree of variability among strains of the species (Molenaar et al 2005), indicating local genome plasticity and implying that PTSs transporters might be more frequently acquired and/or lost by horizontal gene transfer (HGT) as compared with ABC transporters in both these *Lactobacillus* species. The main driver is probably that PTS systems which are always carbohydrate transport systems, while ABC transporters are also known to transport other compounds (AA, metals, small molecules). This may imply also that ABC transport functions have less redundancy and may play more essential roles in a variety of environments as compared to the PTSs that could serve to some extent expendable or redundant functions depending on the environmental availability of specific carbohydrates. Nevertheless, 20 of the 134 PTS associated OGs (20 PTSs), appear to be conserved among all strains and are predicted to transport the monosaccharides glucose, galactose, mannose, fructose, the polyols sorbitol, and mannitol, and the disaccharides cellobiose, trehalose, sucrose, or oligosaccharides (glucosides). The import and utilization of this panel of carbohydrate sources for

growth is consistent with the phenotypic analyses of strains of the *L. rhamnosus* species (Douillard, Ribbera, Kant, et al. 2013; Ceapa et al. 2015). The remaining PTS encode transporters predicted to import the additional carbohydrates maltose, arbutin, sorbose, N-acetylgalactosamine, galactosamine, and lactose, but include also systems that have a predicted substrate redundancy relative to the core-genome associated PTS functions (supplementary table S7, Supplementary Material online). Of the overall 163 ATP-binding cassette (ABC) transporters (the gene number varying between 133 and 148 genes per genome; supplementary table S7, Supplementary Material online), 72% belong to the core genome functions (118 OGs forming between 99 and 108 units). These core genome ABC transport systems are predicted to encode both importers and exporters, with substrate assignments that include carbohydrates, amino acids, lipids, and metals (supplementary table S7, Supplementary Material online). The carbohydrate ABC transporters encoded within the variome are predicted to mediate maltose/maltodextrin transport in strains of clade 1 and ribose/D-xylose importers in all clades except clade 4.

Variation in the transporters encoded by any strains can be explained by strain-specific gene loss in comparison to a common ancestor (Toh et al. 2013) or HGT acquisition from bacteria that coinhabit the same niche (Levin and Cornejo 2009; Richards et al. 2011; Barraclough et al. 2012). In some cases the latter evolutionary process can be recognized in the clade distribution of these functions. For example, the sorbose PTS transport function that encompasses eight genes,

is only present in strains belonging to clades 5–8, and shares very high sequence identity (>80%) with the sorbose transporter PTSs identified in *L. casei* and *L. paracasei* species, which could imply that *L. rhamnosus* strains acquired these genes from these close relative species. The selective loss of these OGs in several strains of all three species, is likely to illustrate that this event occurs relatively frequently in representatives of the *L. casei* taxonomic group (Toh et al. 2013). Analogously, an operon encoding five OGs annotated as a ribose ABC transporter is absent in all strains of clade 4 (except strain Lrh2), whereas a taurine ABC transporter is absent in several strains of clades 6 and 8. The OGs belonging to these functions have close homologues in the genomes of *L. casei* and *L. zeae*, and may, therefore, also represent genetic traits that are exchanged among these species.

Several of the transporters do not appear to follow a clade-specific distribution and appear to be present in various strains of different clades. An N-acetyl galactosamine PTS (GalNAc) transporter is present in all strains, but a functionally redundant and genetically distinct PTSs appears to be acquired by various strains of clades 2, 3, 6, 7, and 8. Gene acquisition by HGT of the arbutin PTS, which is only found in the two strains of clade 1 is most likely explained by HGT, since the best homologues of these genes are found in *Lactobacillus farciminis* but not the taxonomically related species *L. casei* or *L. paracasei*. Another example of HGT is the operon responsible for taurine transport (LGG_00171 to LGG_00177) that is lacking from strains of clades 6 and 8. A cladogram obtained from the multiple alignment of all similar sequences from NCBI's nr database (supplementary fig. S4, Supplementary Material online) for the ABC transporter showed that it is shared by some strains of the *L. casei* group, with the closest homologue in *L. farciminis*, suggesting that this species could be the source of multiple HGT events in *L. rhamnosus*.

A recent comparative genomics study reported that *L. rhamnosus* GG (clade 4) has the highest number of predicted transport proteins among the *L. casei* group representatives (Toh et al. 2013). However, the current analysis revealed that strains of clades 3 and 4 encode even a larger number of transport genes, making these strains of *L. rhamnosus* the richest transport encoding strains of this taxonomic group (supplementary table S7, Supplementary Material online).

The large variety of carbohydrates *L. rhamnosus* strains are adapted to grow on is apparent in the percentage of genes of the pan-genome (23%) and especially the variome (34%) assigned to functions related to carbohydrate digestion, transport, and metabolism (supplementary table S8, Supplementary Material online). A more detailed view at the carbohydrate utilization patterns of *L. rhamnosus* strains in vitro has been published (Ceapa et al. 2015). In the context of glycan degradation, all strains were predicted to be able to release glucosyl, galactosyl, and mannosyl reducing-end residues of saccharide polymers, as well as cleave 1–6 linked fucose and 1–3 and 2–6 linked N-acetyl-glucosamine

(GlucNAc) residues from (host associated cell-surface) glycosides. Strains of clades 2 and 6 have a reduced ability to degrade glycans compared with the other strains and, for example, are predicted to be unable to release protein O-linked glycans.

The differential presence of extracellular glycosidases, transport systems, and metabolic pathways for the release and utilization of L-fucose, galactose, GalNAc, which are particularly present in clades 4, 5, and 6, supports the idea that some *L. rhamnosus* strains could be adapted to grow on either human milk oligosaccharides or mucus (Reunanen et al. 2012). Notably, *L. rhamnosus* GG (clade 4), was shown to adhere to mucus in vitro (Kankainen et al. 2009) and appears to encode the capacity to use mucus-associated saccharides as carbon and energy source for growth (Sánchez et al. 2010). Strains of clade 8, which is also dominated by human-derived isolates, lack these extracellular glycosidases, but share the capacity to produce other galactosidases and phospho-glucosidases, which may also reflect their intestinal adaptation.

Fermentative Capacity, Pyruvate Dissipation

Lactobacillus rhamnosus is a facultative heterofermentative organism whose main carbohydrate fermentation pathway is glycolysis through the Embden–Meyerhof–Parnas (EMP) pathway [52]. A partial pentose phosphate pathway also appears to be present (missing enzymes: gluconolactonase, 2-keto-3-deoxygluconate aldolase, 3-hexulose-6-phosphate synthase) and is likely to support biosynthetic routes as well as pentose utilization.

Lactobacillus rhamnosus grows preferentially on glucose that is degraded via the EMP pathway leading to pyruvate, which is then transformed into D- and L-lactate by the corresponding lactate dehydrogenases, encoded by *ldhL* and *ldhD* genes that are part of the core genome. These functions appear redundant in the *L. rhamnosus* core genome, with three copies of *ldhD* and seven copies of *ldhL*, supporting the critical role of this function in the overall metabolism of this bacterium. Nevertheless, *L. rhamnosus* strains cultivated in cheese-like environments displayed relatively low lactic acid production and produced significant amounts of acetic acid, illustrative of the heterofermentative character of the species [64]. Besides D- and L-lactic acid and acetic acid pathways, additional pyruvate-dissipation pathways are encoded by the genome leading to formic acid, acetaldehyde, ethanol, oxaloacetate, and acetoin (supplementary table S8, Supplementary Material online). Production of all these compounds is industrially relevant in the preservation of food raw-materials (Bove et al. 2012), and cheese ripening (Bove et al. 2011), but also for substrate production in bio-plastic manufacturing (Flieger et al. 2003).

All *L. rhamnosus* strains contain only a partial tricarboxylic acid cycle (TCA cycle) (Golbach et al. 2007) and harbor genes predicted to encode the pathway required for succinate to

citrate conversion. Nevertheless, the utilization of molecular oxygen as a terminal electron acceptor appears possible via the production of menaquinones, or vitamin K, which is a capacity that appears unique for *L. rhamnosus* among the *Lactobacillus* genus (Hess et al. 1979).

Nitrogen Metabolism

The niches from which the *L. rhamnosus* strains used in this study were isolated encompass both protein-rich (e.g., dairy) and protein-poor environments (e.g., soil and plants) (Douillard, Ribbera, Kant, et al. 2013) (supplementary table S1, Supplementary Material online). The *L. rhamnosus* core genome encodes import functions (dominated by the ABC-family transporters) for several amino acids, for which the biosynthesis pathway is not encoded by the genome, including alanine, isoleucine, leucine, valine, phenylalanine, threonine, and tyrosine. Notably, within the variome an ABC transporter is predicted to import taurine.

The main source of amino acids and peptides produced by LAB that live in milk, are derived from caseins, the most abundant protein in milk. Hydrolysis of caseins by LAB is initiated by cell-envelope proteinases (CEP) that degrade the protein into oligopeptides. The *L. rhamnosus* core genome encodes two CEP, PrtP, and PrtR, that are responsible for hydrolysis of caseins. Other core protein degradation enzymes include protease PepS16, aminopeptidases PepA, and MAP, endopeptidases PepS, PepS24, and PepM16 and proline-specific peptidases PIP and PepQ.

While the genomic content of amino acid and peptide transport differs considerably between strains, all strains seem to be equipped with a similar cytoplasmic peptidase repertoires, of which only four out of the 36 predicted peptidases vary among strains (supplementary table S5, Supplementary Material online). These encompass three metallopeptidases and a serine peptidase. The intracellular metallopeptidase encoded by the *abgB* gene is present in all strains of clusters 2 and 4 (enriched for human isolates) and might provide an advantage to these strains in the intestinal niche since this function is predicted to support aminobenzoyl-glutamate utilization for folic acid recycling (de Crécy-Lagard et al. 2007). For the additional metallopeptidases (DppA and ZmpB) no information is available about their specific function, although the observation that ZmpB is secreted may imply that it exerts its function in the bacterial cell envelope.

Complementary to this extensive protein degradation machinery, all *L. rhamnosus* genomes encode complete pathways for the biosynthesis of 12 amino acids: arginine, asparagine, aspartic acid, cysteine, glutamine, glutamic acid, glycine, histidine, lysine, methionine, proline, serine, and tryptophan. However, the operon encoding serine acetyltransferase, cystathionine gamma-lyase, and cystathionine beta-synthase that is involved in the transformation of serine to cysteine, appears to be absent from strains in clades 4 and 7. Intriguingly, the sequence of this operon in the other strains displays high-level identity with similar operons encountered in *Lactobacillus brevis/L. buchneri*, suggesting it has been acquired by HGT.

Regulation and Environmental Interaction

Regulatory Functions. Bacteria have developed sophisticated mechanisms for the regulation of both catabolic and anabolic pathways. Generally, bacteria do not synthesize degradative (catabolic) enzymes unless the substrates for these enzymes are present in their environment (Deutscher 2008). Proteins controlling cellular transcription (regulatory proteins) constitute 6% of *L. rhamnosus* pan-genome proteins. Among them, 123 OGs are annotated as core regulator genes and 84 OGs display variation among the strains. Two sigma factor encoding genes are core functions in *L. rhamnosus*, including the housekeeping sigma factor 70 encoded by *rpoD*, and a sigma 24-like factor for which the target genes remain to be determined, but could include stress response as has been observed in *Escherichia coli* (Haldenwang 1995). Strains belonging to the same clade share a variable number of regulator OGs with different predicted functions. For example, 28 regulator OGs of clade 1 include the predicted regulators of arabinose, rhamnose, fucose, fructose, and tagatose transport as well as the LexA protease, and the MazF and YafQ toxin–antitoxin systems. Clade 2 regulator genes are predicted to control transport of ribose, galactitol, glycerol, and oligopeptides, as well as a Zn-dependent protease. OGs representing regulator proteins for clade 5 control the transport of mannose, tagatose, galactitol, cellobiose, mannitol, sorbitol, sorbose, pentitol, xylulose, and arabinose, and illustrate the regulatory variability among the different clades. Notably, although clades 7 and 8 encode 32 and 28 variome-associated regulator OGs, none of these appears to be specific for either of these clades (fig. 2, supplementary table S4, Supplementary Material online summarizing clade specific predicted protein functions). A few regulators that could impact bacterial adaptation to the environment are shared among several strains. Notably, among these is a Prck including quorum-sensing system that was shown to respond to environmental homoserine lactones in *Salmonella enterica* (Michael et al. 2001) and is also found in *L. casei* (Zhang et al. 2010), is present in most members of clades 1, 5, 6, 7, and 8 and is genetically linked to an extracellular polysaccharides (EPS) biosynthesis operon. In addition, in strains of clades 2, 3, and 4 a LysR family regulator is predicted to control the synthesis of rhamnose-containing polysaccharides (Péant et al. 2005), suggesting that environmental adaptation could include the adjustment of extracellular exposed glycan-polymers.

Transcription regulation in response to environmental stimuli frequently involves two-component regulatory systems (TCSs) in many bacteria. In *L. rhamnosus* all seven TCSs

belong to the core genome and are predicted to control envelope stress signal transduction (BaeS-BaeR and LiaS-LiaR), cell-wall metabolism and corresponding antibiotic resistance (VicK-VicR, CiaH-CiaR, and VraR-VraS), membrane lipid fluidity in response to low-temperature growth (DesK-DesR), secreted protein production (AgrC-AgrA), and citrate/malate metabolism (DcuR-DcuS) (supplementary table S11, Supplementary Material online).

Antimicrobial Production and Immunity

Bacteriocins and Bacterial Immunity Proteins. Many LAB produce small antimicrobial peptides or bacteriocins that can provide a competitive advantage by inhibiting or killing competing (close relative) bacterial species. Bacteriocins can display both narrow and broad spectrum inhibitory activities (Riley and Wertz 2002). The *L. rhamnosus* genomes include one to three operons predicted to encode bacteriocin production (supplementary table S9, Supplementary Material online), which is in agreement with the notion that various *L. rhamnosus* strains were shown produce varying antimicrobial activities (De Keersmaecker et al. 2006; Douillard, Ribbera, Kant, et al. 2013; Pithva et al. 2014; Vong et al. 2014).

The *L. rhamnosus* bacteriocin encoding genes in the variome do not have a clear clade distribution. A set of four antimicrobial peptide-encoding genes are distributed among the strains in an apparent random manner with at least one member of each clade containing one up to all four peptides. Another antimicrobial peptide resembles a Class II pediocin-like bacteriocin (Motlagh et al. 1992), which is only present in the beer isolate ATCC8530 and is closely related to the pediocin-like bacteriocin PA-1 identified in *L. pentosus* (Henderson et al. 1992).

Surface Proteins. The capacity to interact with environmental surfaces, including host cells, is an important factor in niche colonization, and persistence. Several molecules that play a role in *L. rhamnosus* cross-talk with host cells have been identified and are among the core capacities of the species, including the p40/p75 muramidases (Bäuerl et al. 2010), and lipoteichoic acid modification D-alanylation genes (*dlt* operon) (Abi Khattar et al. 2009) (supplementary table S10, Supplementary Material online), which have been proposed to interaction with the host mucosa by binding to mucus, influencing epithelial and immune functions, respectively (Lee et al. 2013). In the core genome of *L. rhamnosus*, we identified 37 genes that encode sortase-dependent surface proteins (characterized by the conserved LPxTG motif) (Supplementary table S10, Supplementary Material online), including 17 proteins with a domain composition that may be predictive for a role in host interaction, including the recognition of fibronectin- and collagen-binding domains as well as a bacterial Ig-like domain (fig. 3). For instance, the third-largest protein—LGG_02923 (OG1951) contains a signal

peptide, an LPxTG anchor, and a domain with four leucine rich repeats that is not frequently found in bacteria, but was proven to play a role in infection by some pathogens (Courtemanche and Barrick 2008). In 21 genomes, the copy number per genome is 1, and it varies from 2 to 5 in the others.

Proteins predicted to be exported or surface exposed within the *L. rhamnosus* variome do not appear to follow a clade-specific distribution (supplementary table S10, Supplementary Material online), which may illustrate recent and variable gene acquisition or loss of these functions among the *L. rhamnosus* strains. Some of these proteins have been investigated in vitro for their possible role in the interaction with human epithelial or immune cells, including the adhesion and biofilm stimulating factor encoded by *mabA* (Vélez et al. 2010), the secreted-docking protein pair encoded by the *spcABCD* (Gagic et al. 2013), and the main surface protease and anti-inflammatory molecule PrtP (Habimana et al. 2007) and the pilin encoding *spaCBA–srtC1* operon (Reunanen et al. 2012). These proteins are all large, multidomain proteins that contain the sortase-dependent LPxTG anchoring motif, as well as several potential adhesion domains. This observation prompted us to investigate other large multidomain proteins that are predicted to be secreted or surface exposed. The largest among these genes is encoded by OG1717 (> 3,500 amino acid residues), and is shared among 21 of the strains analyzed here, and entirely missing from clades 2 to 5. Notably, the strains that encode this gene appear to have a duplicated copy of this gene that encodes a protein with a signal sequence, one or more adhesion domains and 16 collagen binding domains (fig. 3).

The capacity to adhere to mucus, the extracellular matrix, and/or intestinal epithelial cells is interesting properties with regard to probiotic features such as colonization of the gastrointestinal tract and interaction with the host. Among the variome surface exposed proteins, there is only one predicted mucus-binding protein: OG2369 (605 aa). The protein contains a cell wall anchor repeat and two MucBP mucus-binding domains and has a close resemblance to *L. paracasei* surface proteins. It is only present in two strains: Lrh33 originating from a fermented dairy product and Lrh24, isolated from animal (goat) feces.

Extracellular Polysaccharides. Bacterial EPS can play various roles in environmental interactions, including mammalian host interactions, for example by preventing recognition of the bacteria by the host immune system (Fanning et al. 2012). Six EPS biosynthesis gene clusters were identified in the *L. rhamnosus* pan-genome (supplementary table S9, Supplementary Material online), and each strain contains two to four of these without an obvious clade-specific pattern being apparent. Only EPS cluster "5" appears to belong to the core genome (Péant et al. 2005; Lebeer et al. 2009), and its inactivation was shown to reduce the production of rhamnose-rich EPS (Lebeer et al. 2009). The EPS cluster "6,"
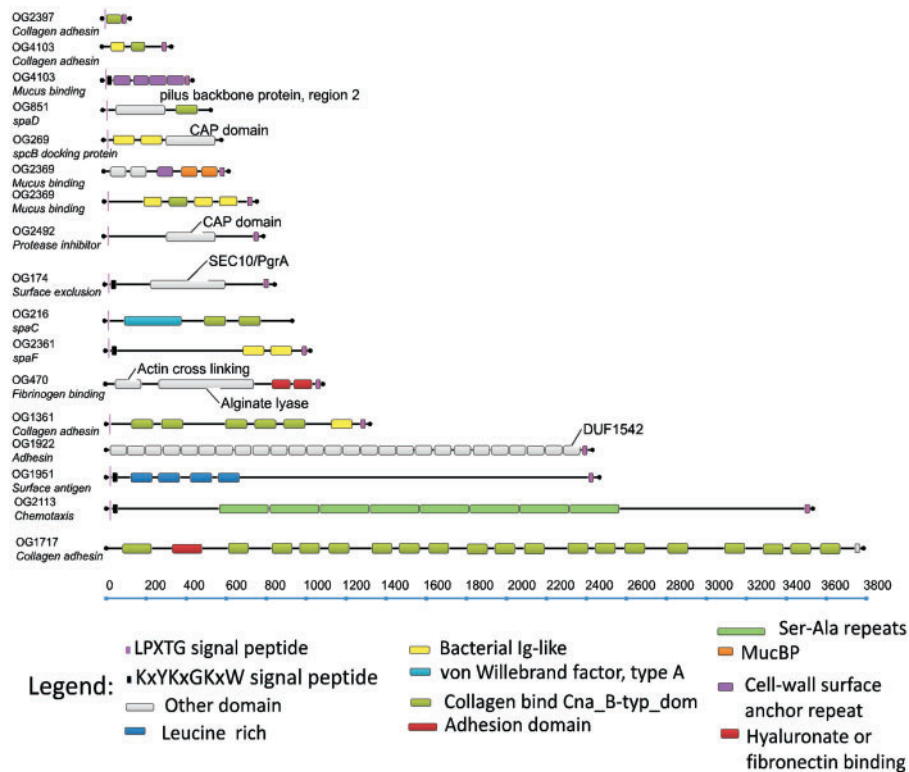
F ɪ ɢ. 3.—LPxTG proteins with host interaction potential. LPxTG proteins not included in this figure: sortases, hydrolases, lyases, proteases.

containing several predicted mannosyl-glycosyltransferases may be responsible for mannosyl-EPS production in *L. rhamnosus* which was previously proposed on basis of ConA surface-probing of *L. rhamnosus* GG using atomic force microscopy (Francius et al. 2008). The remaining EPS do not have clearly predicted specificities for the glycosyltransferases they encompass and clusters "1" and "2" were present in only a single strain Lrh24, and HN001, respectively.

Mobile Genetic Elements, and Defense against Foreign DNA.

Mobile genetic elements in bacteria like conjugative plasmids, insertion sequences (IS) elements, transposons and (pro-) phages (Frost et al. 2005) play a prominent role in genetic mobility within a strain's genome (intracellular mobility), but also between bacterial genomes of different strains (intercellular mobility) and form one of the driving forces in the evolution of organisms (Liu et al. 2009).

Plasmids. Three plasmids of *L. rhamnosus* have previously been described, pLC001 (strain LC705), pLR001 and pLR002 (strain HN001) (Morita et al. 2009). Next to the search for these known plasmids, novel plasmid identification relied on the detection of contigs that contain plasmid-associated genes, particularly focusing on replication functions. The latter approach failed to identify novel plasmid entities in the

genome database created in this study. Notably, strains belonging to clades 4 and 7 did consistently appear to lack plasmids, which was also the case for individual strains in the other clades (Lrh5 from clade 1, Lrh11 from clade 6, and ATCC7469, ATCC8530, Lrh23, Lrh12, and Lrh15 from clade 8; fig. 4). The distribution of the previously identified plasmids among the current strain collection appeared not to follow a clade-specific distribution.

Bacteriophages. Fourteen regions were identified that encode a total of 328 phage-related OGs. Seven phage-related regions that encompass more than 20 genes each, appear to be part of the core genome, and appear to be consistently inserted in the same chromosomal locus. Additional phage-related loci were variably present in the *L. rhamnosus* clades and strains (fig. 4) and resembled previously recognized phages of *Lactobacillus* origin (e.g., Lc-Nu [Tuohimaa et al. 2006], Lrm1 [Durmaz et al. 2008], A2 [Alvarez et al. 1999], Lb338-1 [Alemayehu et al. 2009], phi adh [Altermann et al. 1999], phi AT3 [Lo et al. 2005]). The gene composition of these phage-associated loci is congruent with the canonical mosaic-like composition of (pro)phage genomes, which exemplifies the high evolutionary rate of phage diversification (Abedon 2009).

Restriction Modification. Restriction modification systems (RMS) are known to be variable among strains of a species
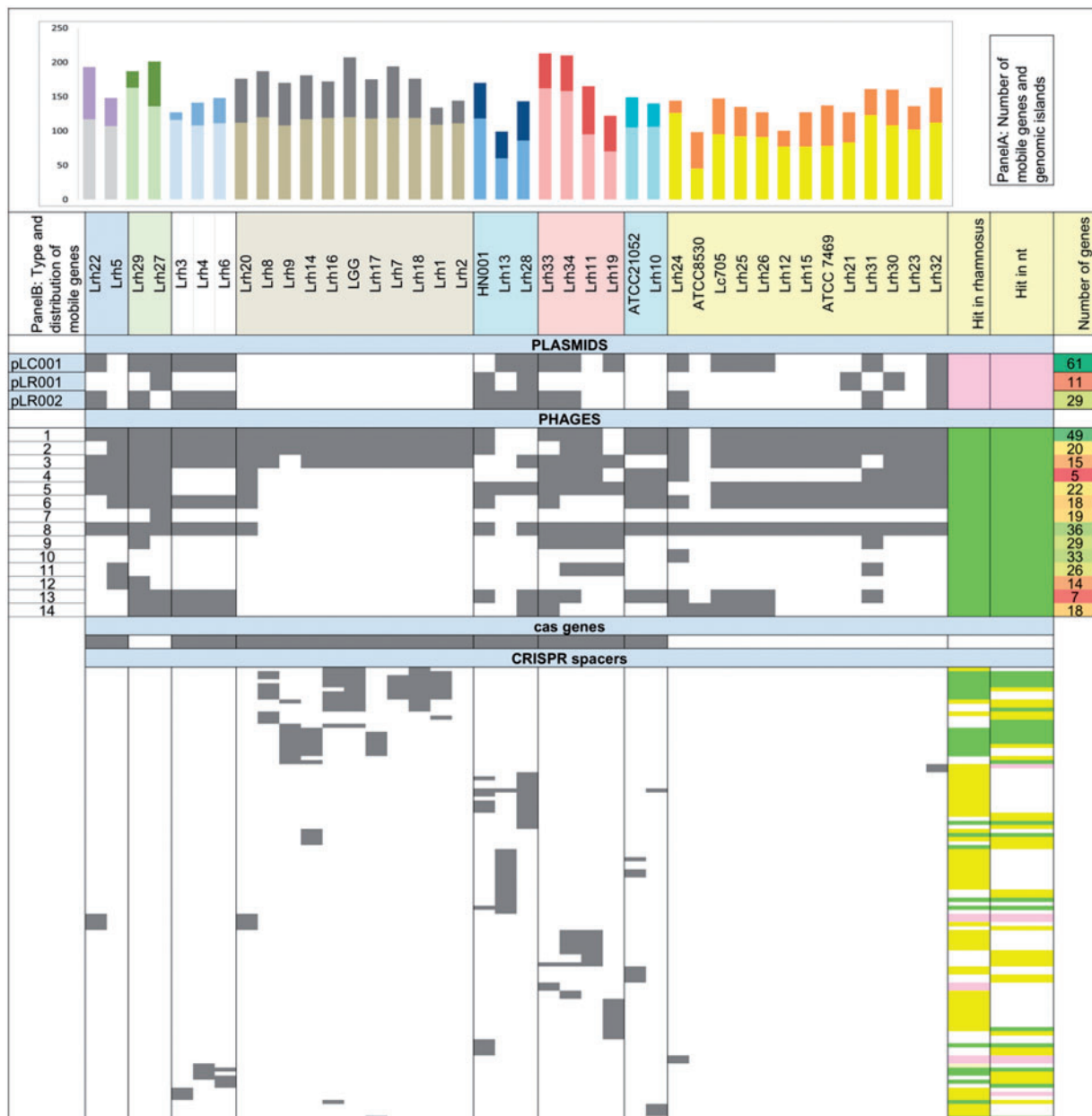
FIG. 4.—Summary of mobile elements present in *L. rhamnosus* strains. Panel 1: Number of genes annotated as mobile elements in the pan-genome, including plasmids, phages, integrases, transposases (lighter colors), and number of genomic islands (dark colors). Panel 2: Type and distribution of mobile elements in each genome: plasmids, phages, cas genes and CRISPR spacers. Gray represents gene presence and white gene absence. For the spacers' analysis, only spacers that present a hit in any of the databases are represented. In the columns Hits in rhamnosus and Hits in nt, the colors represent the type of hit: yellow: unknown; pink: plasmid; green: phage genes; white: not found. Strains are organized by genetic clades separated by vertical lines.

(Xu et al. 2000) and play an important role in protection against foreign DNA, including phage attack. Therefore, RMS system presence and diversity were evaluated in the *L. rhamnosus* genomes (supplementary table S2, Supplementary Material online), revealing that there are no core-genome as-sociated RMS, and the RMS within the variome display a large interstrain variation. Among the most common systems are the Mrr system (Waite-Rees et al. 1991) in strains of clades 1 and 8 (except Lrh31), the Type 1 RMS (Loenen, Dryden, Raleigh Wilson, et al. 2014) shared between the majority of clades 2, 3, 7, and 8, and the Type II and Type III RMS (Loenen, Dryden, Raleigh, Wilson, Murray, et al. 2014) characteristic for clade 4 and shared by some strains of clade 5. In addition, some RMS are present in certain strains of which the closest

relative lacks them. Notably, all strain-specific RMS appeared to be intact, which was also seen in other organisms like *Helicobacter pylori* (Lin et al. 2001), whereas the majority of the shared RMS (> 7 RMS) are predicted to be incomplete or inactivated by mutation(s). This could relate to positive selection for recently acquired strain-specific RMS, a process similar to *H. pylori* (Lin et al. 2001), where through constant acquisition of new RMS and the selective inactivation or removal of the older ones, the defense against foreign DNA, including phages, is constantly updated.

CRISPR-Cas.   HGT can be beneficial for bacteria by bringing in new functionalities as long as it does not disrupt fitness in the ecosystem. This balance is partly maintained by the CRISPR-Cas adaptive immunity system, which uses CRISPR and *cas* (CRISPR-associated) genes. The CRISPR loci are partially palindromic repeats separated by short stretches of spacer DNA that are acquired from invasive bacteriophage or plasmid DNA. These spacer sequences allow cells to recognize and cleave invasive DNA identical to the included sequence (Brouns et al. 2008).

The CRISPR locus in the *L. rhamnosus* genomes is of Type II-A (Lsal1 family) with a CRISPR repeat sequence of 5′- GTCTCA GGTAGATGTCAGATCAATCAGTTCAAGAGC-3′. Lsal1-type CRISPR loci were identified in 25 out of the 40 strains (fig. 4), and could play a prominent role in *L. rhamnosus* evolution by preventing intraspecies HGT. Notably, the presence of the CRISPR-Cas system follows the clade distribution in several ways, which is clearly illustrated by the observation that all clade 2 and 8 strains lack a CRISPR-Cas system. Both the sequence of the *cas* genes of in *L. rhamnosus* and their genetic organization resembles that of *L. casei* (Smokvina et al. 2013), suggesting its inheritance from a common ancestor. Notably, the presence of the CRISPR-Cas system does not seem to correlate with the number of genomic islands recognized in these different strains.

Noteworthy, groups of spacer sequences within the repeat locus are specific for each of the clades 3, 4, 5, and 6 (fig. 4, supplementary fig. S2, Supplementary Material online) and their predicted targets (on basis of 100% sequence identity to genes in the NCBI nr database) belong to *Lactobacillus* phages (Lc-Nu, A2, P2, and Lrm1) and plasmids (*L. casei* str. Zhang plasmid plca36, *Lactobacillus rennini* plasmid pREN, *Lactobacillus plantarum* 16 plasmid Lp16C, plasmid pC30il, cryptic plasmid pLJ42, *Lactobacillus* helveticus H10 plasmid pH10, and *Lactobacillus salivarius* UCC118 plasmid pSF118-20) (supplementary table S2, Supplementary Material online, fig. 4). Remarkably, numerous spacers show sequence identity to *L. rhamnosus* phages (Lc-Nu, Pi2, A2) and plasmids (pLR002) present in their own as well as other clades. Overall, only 23 out of 112 of the spacer sequences targeted genetic entities from other species (the NCBI nr database) and were not found in the *L. rhamnosus* pan-genome, suggesting

that the CRISPR system in *L. rhamnosus* could play a role in protection against the entry of mobile genetic elements from the same species and thereby restrict the intra strain and intra clade genetic exchange events, which could influence clade evolution by favoring the acquisition of genetic material from more distant species rather than promote the genomic convergence of the *L. rhamnosus* clades. This concept is also supported by the close homology of many transporter encoding genes to more distant species, such as *Lactobacillus farciminis*, *Enterococcus faecalis*, and *Leuconostoc kimchii* (supplementary table S2, Supplementary Material online).

Core Genome Evolutionary Drift and Variome Distribution

The core proteome phylogenetic tree of a species relies on single amino acid polymorphisms (SAPs) of the conserved proteins, which are accumulated over time due to spontaneous mutations. Recent whole-genome sequencing has revealed that single nucleotide polymorphisms (SNPs) are the most prevalent form of genetic diversity among different strains of the same bacterial species (Harris et al. 2010). The randomness of their occurrence implies that the number of sequence variations between strains of a species (SNPs and/or SAPs) can be interpreted as a measure of evolutionary time relative to a common ancestor. However, some sequence variations (e.g., SNPs and SAPs) can be selected due to a role in niche adaptation and fitness. For example, the presence of certain SAPs in bacteria has been associated with changes in virulence (Carroll et al. 2011), antibiotic resistance (Chewapreecha et al. 2014), metabolism (Tolentino et al. 2013), and persistence in the host intestine (van Bokhorst-van de Veen et al. 2013). Patterns of SAPs within core proteome functions were used to determine the phylogenetic relationship of the *L. rhamnosus* strains (fig. 5A), by alignment of 1,008 orthologous proteins that are encoded in a single copy in the core genome of the *L. rhamnosus* strains. This analysis identified a total of 5.127 SAPs within the (single copy) core proteome of at least one of the *L. rhamnosus* strains and allowed the phylogenetic clustering of the strains on basis of evolutionary drift of the core proteome. This analysis revealed a substantial evolutionary distance between the strains in clade 7 (Lrh10 and ATCC21052; fig. 5A) and the other 38 strains of *L. rhamnosus*, whereas strains of clades 4 and 8 that are all human isolates are quite closely related. This phylogenetic grouping is in agreement with genomic BLAST analysis of public *L. rhamnosus* genomes performed by NCBI, which assigned strain ATCC21052 to a subgroup of the species with a 6% variation compared to the other genomes available. In contrast, this analysis appears to contradict a recent phylogenetic analysis of the *L. casei* group that included several *L. rhamnosus* strains (Toh et al. 2013), and concluded that strain ATCC21052 clustered closely together with strain ATCC 53103 (GG). However, this latter study employed
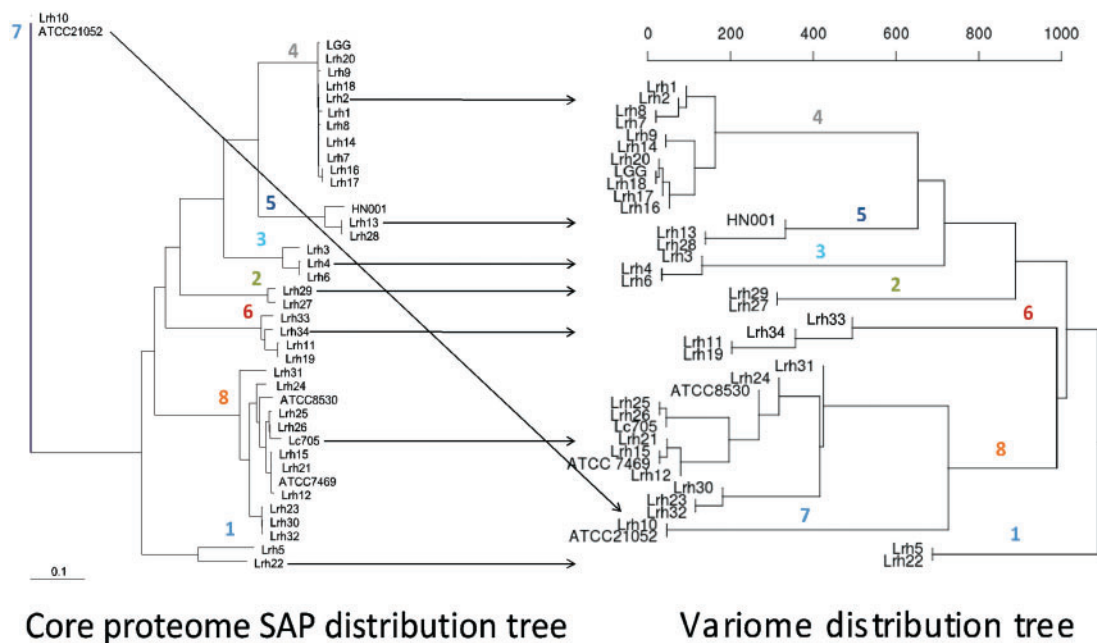
FIG. 5.—*Lactobacillus rhamnosus* genome-based phylogenetic relatedness, (Panel *A*) based on core proteome SAP distribution and (panel *B*) based on hierarchical clustering of variome gene distribution. The scale for the presence–absence tree represents the number of variable OGs the clustering is based on. Arrows connecting the same strains of both trees aims at highlighting the common groups of strains, which are also marked by the variome-based clade numbering (panel *B*).

the sequences of "only" 34 ribosomal proteins for the phylogenetic analysis and thus provides substantially lower resolution as the current analysis that includes 1,008 core proteome sequences. Notably, the Human Microbiome Project has taken the ATCC21052 strain as a reference for the *L. rhamnosus* species, but our findings imply that this strain is not the best representative of this species. Based on the current analysis it would be recommendable to choose representative strains of the different clades as a reference gene-set for the *L. rhamnosus* species, to ensure the appropriate representation of its genetic diversity.

Remarkably, despite the different evolutionary processes that drive the phylogenetic discrimination of *L. rhamnosus* strains on basis of either the core proteome SAPs (random mutations without selective advantage) and the variome distribution (gene acquisition and loss) (fig. 1), the two phylogenetic trees present a highly congruent strain distribution. The variome-based tree allowed the recognition of eight clades among the 40 sequenced strains, and exactly the same grouping into the same eight clades can also be recognized from the core-proteome-based tree (fig. 5). The congruent evolution of core and variome relationships between strains has implications for the evolution of the *L. rhamnosus* species. This observation implies that clade determining gene acquisition and loss events have the same evolutionary age as the core-genome discriminating SAPs. This is in apparent contradiction with the concept that gene acquisition and loss are strong

drivers of niche adaptation and commonly represent more recent genome diversification events. The CRISPR-cas system may have played a prominent role in this restricted variome diversification (see above), especially within the most congruent clades. Notably, this remarkable evolutionary path is distinct from the recognizable niche-associated evolution of the taxonomically related species *L. casei*. Similar analyses in this species allowed the recognition of dairy environment associated gene loss events that is typical for genome decay in the rich environment of milk (Broadbent et al. 2012), and which appeared to have occurred independent of the evolutionary drift of the core proteome.

Conclusions

The availability of 40 genome sequences of *L. rhamnosus* has enabled us to obtain a better understanding of the functional diversity and evolutionary relatedness of this species, which encompasses many industrially relevant strains. The current analysis that focused on the variome-associated genes and their distribution among clades of strains within the *L. rhamnosus* species, complements the previously presented core genome analysis (Douillard, Ribbera, Kant, et al. 2013).

The species *L. rhamnosus* is closely related to *L. casei* and *L. zeae*, and encompasses a genetically diverse group of strains, with a high frequency of discriminative polymorphisms in its core genome (SNPs and SAPs) and an impressive accessory

genome or variome distribution. The comprehensive in silico examination of the variome associated functions and their distribution in terms of metabolic and regulatory diversity further illustrates the evolutionary diversity of this versatile *Lactobacillus* species that has evolved to grow and persist in a variety of ecological niches, including the intestinal tract of humans and animals. Notably, most of the strains appear to encode an extensive protein and sugar transport and catabolism capacity, which is congruent with their environmental versatility. Nevertheless, several genes and operons seem to have been acquired by HGT in some strains or clades that encode carbohydrate transport and catabolism functions, but also several other (industrially) important phenotypic traits, like polysaccharide biosynthesis (EPS), bacteriocin production, RMS, and/or bacterial defense systems (CRISP-Cas). In addition, other aspects of genomic diversity that may reflect niche adaptations can also be recognized, including the diversity of extracellular functions putatively involved in host interactions by cell adhesion or modulation of the host immune system.

The application of comparative genome sequencing to determine core genome and variome functions provided us with an unprecedented view of genome dynamics and adaptive evolution of the L. rhamnosus species. The recognition of the highly congruent phylogenetic relatedness of the core proteome and variome distribution patterns among the different strains included in this study indicates that HGT events may have played an important role in the species' evolution. The CRISPR-Cas system spacer recognition patterns indicate a putative role for this machinery in L. rhamnosus's evolutionary path, by its ability to limit intraspecies mobility of genetic elements and prevent the evolutionary convergence of the eight clades recognized by the variome distribution patterns. Finally, complementing these in silico analyses with phenotypic profiling of the strains of this species can expand our understanding of gene-function relationships and targeted genetic engineering strategies can subsequently establish the role of specific genes and functions in the adaptation to particular niches, where the molecular mechanisms underlying the cross-talk of this bacterium with the host intestine mucosa is of particular interest in the context of its use as health-promoting diet ingredient, that is, a probiotic.

Acknowledgments

Supplementary Material

Supplementary figures S1–S4 and tables S1–S11 are available at *Genome Biology and Evolution* online http://www.gbe.oxfordjournals.org/.

Literature Cited

Abedon ST. 2009. Phage evolution and ecology. Adv Appl Microbiol. 67:1–45.

Abi Khattar Z, et al. 2009. The dlt operon of *Bacillus cereus* is required for resistance to cationic antimicrobial peptides and for virulence in insects. J Bacteriol. 191:7063–7073.

Alemayehu D, et al. 2009. Genome of a virulent bacteriophage Lb338-1 that lyses the probiotic *Lactobacillus paracasei* cheese strain. Gene 448:29–39.

Alikhan N-F, et al. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12:402.

Altermann E, et al. 1999. Primary structure and features of the genome of the *Lactobacillus gasseri* temperate bacteriophage φadh. Gene 236:333–346.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Alvarez MA, et al. 1999. Stable expression of the *Lactobacillus casei* bacteriophage A2 repressor blocks phage propagation during milk fermentation. J Appl Microbiol. 86:812–816.

Barraclough T, et al. 2012. Evolving concepts of bacterial species. Evol Biol. 148–157.

Bäuerl C, et al. 2010. Functional analysis of the p40 and p75 proteins from *Lactobacillus casei* BL23. J Mol Microbiol Biotechnol. 19:231–241.

Bergey DH. 2009. Bergey's manual of systematic bacteriology. Vol 3: The Firmicutes. New York: Springer-Verlag.

Bland C, et al. 2007. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics 8:209.

Borneman AR, et al. 2012. Comparative analysis of the *Oenococcus oeni* pan genome reveals genetic diversity in industrially-relevant pathways. BMC Genomics 13:373.

Bove CG, et al. 2011. Evaluation of genetic polymorphism among *Lactobacillus rhamnosus* non-starter Parmigiano Reggiano cheese strains. Int. J. Food Microbiol. 144:569–572.

Bove CG, et al. 2012. Metabolic and proteomic adaptation of *Lactobacillus rhamnosus* strains during growth under cheese-like environmental conditions compared to de Man, Rogosa, and sharpe medium. Proteomics 12:3206–3218.

Broadbent JR, et al. 2012. Analysis of the *Lactobacillus casei* supragenome and its influence in species evolution and lifestyle adaptation. BMC Genomics 13:533.

Brouns SJJ, et al. 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321:960–964.

Canchaya C, et al. 2006. Diversity of the genus Lactobacillus revealed by comparative genomics of five species. Microbiology 152:3185–3196.

Carroll RK, et al. 2011. Naturally occurring single amino acid replacements in a regulatory protein alter streptococcal gene expression and virulence in mice. J Clin Invest. 121:1956–1968.

Ceapa C, et al. 2015. Phenotype profiling links genotype to carbohydrate utilization signatures in the species *Lactobacillus rhamnosus*. Appl Environ Microbiol. 81(16):5458–5470.

Chewapreecha C, et al. 2014. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic Genes. PLoS Genet. 10:e1004547.

Courtemanche N, Barrick D. 2008. Folding thermodynamics and kinetics of the leucine-rich repeat domain of the virulence factor Internalin B. Protein Sci. 17:43–53.

De Crécy-Lagard V, et al. 2007. Comparative genomics of bacterial and plant folate synthesis and salvage: predictions and validations. BMC Genomics 8:245.

De Keersmaecker SCJ, et al. 2006. Strong antimicrobial activity of *Lactobacillus rhamnosus* GG against *Salmonella typhimurium* is due to accumulation of lactic acid. FEMS Microbiol Lett. 259:89–96.

Deutscher J. 2008. The mechanisms of carbon catabolite repression in bacteria. Curr Opin Microbiol. 11:87–93.

Douillard FP, Ribbera A, Järvinen HM, et al. 2013. Comparative genomic and functional analysis of *Lactobacillus casei* and *Lactobacillus rhamnosus* strains marketed as probiotics. Appl Environ Microbiol. 79:1923–1933.

Douillard FP, Ribbera A, Kant R, et al. 2013. Comparative genomic and functional analysis of 100 *Lactobacillus rhamnosus* strains and their comparison with strain GG. PLoS Genet. 9:e1003683.

Durmaz E, et al. 2008. Genome sequence and characteristics of Lrm1, a prophage from industrial *Lactobacillus rhamnosus* strain M1. Appl Environ Microbiol. 74:4601–4609.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

Fanning S, et al. 2012. Bifidobacterial surface-exopolysaccharide facilitates commensal-host interaction through immune modulation and pathogen protection. Proc Natl Acad Sci U S A. 109:2108–2113.

Flieger M, et al. 2003. Biodegradable plastics from renewable sources. Folia Microbiol (Praha) 48:27–44.

Francius G, et al. 2008. Detection, localization, and conformational analysis of single polysaccharide molecules on live bacteria. ACS Nano 2:1921–1929.

Frost LS, et al. 2005. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol. 3:722–732.

Gagic D, et al. 2013. Unique secreted-surface protein complex of *Lactobacillus rhamnosus*, identified by phage display. Microbiologyopen 2:1–17.

Golbach JL, et al. 2007. Growth response of lactobacillus rhamnosus to microaerophilic incubation and addition of reductants in a riboflavin microtiter plate assay. J Rapid Methods Autom Microbiol. 15:320–331.

Guindon S, et al. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology 59(3):307–321.

Habimana O, et al. 2007. Positive role of cell wall anchored proteinase PrtP in adhesion of lactococci. BMC Microbiol. 7:36.

Haldenwang WG. 1995. The sigma factors of *Bacillus subtilis*. Microbiol Rev. 59:1–30.

Harris SR, et al. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. Science 327:469–474.

Heilig HG, et al. 2002. Molecular diversity of Lactobacillus spp. and other lactic acid bacteria in the human intestine as determined by specific amplification of 16S ribosomal DNA. Appl Environ Microbiol. 68:114–123.

Henderson JT, et al. 1992. Purification and primary structure of pediocin PA-1 produced by *Pediococcus acidilactici* PAC-1.0. Arch Biochem Biophys. 295:5–12.

Hess A, et al. 1979. Lipoquinones of some spore-forming rods, lactic-acid bacteria and actinomycetes. J Gen Microbiol. 115:247–252.

Hüfner, et al. 2008. Global transcriptional response of Lactobacillus reuteri to the sourdough environment. Syst Appl Microbiol. 31(5):323–338.

Kanehisa M, et al. 2009. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38:

Kankainen M, et al. 2009. Comparative genomic analysis of *Lactobacillus rhamnosus* GG reveals pili containing a human- mucus binding protein. Proc Natl Acad Sci U S A. 106:17193–17198.

Kant R, et al. 2011. Comparative genomics of *Lactobacillus*. Microb Biotechnol. 4:323–332.

Kleerebezem M, et al. 2003. Complete genome sequence of *Lactobacillus plantarum* WCFS1. Proc Natl Acad Sci U S A. 100:1990–1995.

Lebeer S, et al. 2009. 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. 75:3554–3563.

Lee I-C, et al. 2013. The quest for probiotic effector molecules–unraveling strain specificity at the molecular level. Pharmacol Res. 69:61–74.

Levin BR, Cornejo OE. 2009. The population and evolutionary dynamics of homologous gene recombination in bacteria. PLoS Genet. 5(8): e1000601.

Li Y. 2011. Low-coverage sequencing: Implications for design of complex trait association studies. Genome Res. 21:940–951.

Lin LF, et al. 2001. Comparative genomics of the restriction-modification systems in *Helicobacter pylori*. Proc Natl Acad Sci U S A. 98:2740–2745.

Liu M, et al. 2009. In silico prediction of horizontal gene transfer events in *Lactobacillus bulgaricus* and *Streptococcus thermophilus* reveals protocooperation in yogurt manufacturing. Appl Environ Microbiol. 75:4120–4129.

Lo TC, et al. 2005. Complete genomic sequence of the temperate bacteriophage?AT3 isolated from *Lactobacillus casei* ATCC 393. Virology 339:42–55.

Loenen WAM, Dryden DT, Raleigh EA, Wilson GG. 2014. Type I restriction enzymes and their relatives. Nucleic Acids Res. 42:20–44.

Loenen WAM, Dryden DT, Raleigh EA, Wilson GG, Murray NE. 2014. Highlights of the DNA cutters: a short history of the restriction enzymes. Nucleic Acids Res. 42:3–19.

Lukjancenko O, et al. 2012. Comparative genomics of Bifidobacterium, *Lactobacillus* and related probiotic genera. Microb Ecol. 63:651–673.

Mahony J, et al. 2012. Phages of lactic acid bacteria: the role of genetics in understanding phage-host interactions and their co-evolutionary processes. Virology 434:143–150.

Michael B, et al. 2001. SdiA of *Salmonella enterica* is a LuxR homolog that detects mixed microbial communities. J Bacteriol. 183:5733–5742.

Molenaar D, et al. 2005. Exploring *Lactobacillus plantarum* genome diversity by using microarrays. J Bacteriol. 187(17):6119–6127.

Morita H, et al. 2009. Complete genome sequence of the probiotic *Lactobacillus rhamnosus* ATCC 53103. J Bacteriol. 191:7630–7631.

Moriya Y. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35:W182–W185.

Motlagh AM, et al. 1992. Nucleotide and amino acid sequence of pap-gene (pediocin AcH production) in *Pediococcus acidilactici* H. Lett Appl Microbiol. 15:45–48.

Nermes M, et al. 2011. Interaction of orally administered *Lactobacillus rhamnosus* GG with skin and gut microbiota and humoral immunity in infants with atopic dermatitis. Clin Exp Allergy 41:370–377.

Okonechnikov K, et al. 2012. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28:1166–1167.

Overbeek R, et al. 2014. The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res. 42:D206–D214.

Péant B, et al. 2005. Comparative analysis of the exopolysaccharide biosynthesis gene clusters from four strains of *Lactobacillus rhamnosus*. Microbiology 151:1839–1851.

Pithva S, et al. 2014. Probiotic attributes of autochthonous *Lactobacillus rhamnosus* strains of human origin. Appl Biochem Biotechnol. 173:259–277.

Pruitt KD, et al. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35:D61–D65.

R Development Core Team RFFSC 2008. R: a language and environment for statistical computing. R Found Stat Comput. 1:2673 [cited 2014 Jul 30]. Available from: http://www.r-project.org.

Reunanen J, et al. 2012. Characterization of the SpaCBA pilus fibers in the probiotic *Lactobacillus rhamnosus* GG. Appl Environ Microbiol. 78:2337–2344.

Richards VP, et al. 2011. Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted *Streptococcus agalactiae*. Infect Genet Evol. 11:1263–1275.

Riley MA Wertz JE., 2002 Bacteriocin diversity: ecological and evolutionary perspectives. Biochimie. 84:357–364.

Sánchez B, et al. 2010. Adhesive properties, extracellular protein production, and metabolism in the *Lactobacillus rhamnosus* GG strain when grown in the presence of mucin. J Microbiol Biotechnol 20:978–984.

Siezen RJ, et al. 2012. Complete resequencing and reannotation of the *Lactobacillus plantarum* WCFS1 genome. J Bacteriol. 194:195–196.

Sims, et al. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics 15(2):121–132.

Smokvina T, et al. 2013. *Lactobacillus paracasei* comparative genomics: towards species pan-genome definition and exploitation of diversity. PLoS One 8:e68731.

Toh H, et al. 2013. Genomic adaptation of the *Lactobacillus casei* group. PLoS One 8:e75073.

Tolentino R, et al. 2013. Identification of the amino acid involved in the regulation of bacterial pyruvate, orthophosphate dikinase and phosphoenolpyruvate synthetase. Adv Biol Chem. 03:12–21.

Tuohimaa A, et al. 2006. The genome of the virulent phage Lc-Nu of probiotic *Lactobacillus rhamnosus*, and comparative genomics with *Lactobacillus casei* phages. Arch Virol 151:947–965.

van Bergenhenegouwen J, et al. 2014. Extracellular vesicles modulate host-microbe responses by altering TLR2 activity and phagocytosis. PLoS One 9:e89121. doi: 10.1371/journal.pone.0089121.

Van Bokhorst-van de Veen H, et al. 2013. Genotypic adaptations associated with prolonged persistence of *Lactobacillus plantarum* in the murine digestive tract. Biotechnol J. 8:895–904.

Van Heel AJ, et al. 2013. BAGEL3: automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. Nucleic Acids Res. 41:W448–W453.

Vélez MP, et al. 2010. Characterization of MabA, a modulator of *Lactobacillus rhamnosus* GG adhesion and biofilm formation. FEMS Immunol Med Microbiol. 59:386–398.

Von Ossowski I, et al. 2011. Functional characterization of a mucus-specific LPXTG surface adhesin from probiotic *Lactobacillus rhamnosus* GG. Appl Environ Microbiol. 77:4465–4472.

Vong L, et al. 2014. Probiotic *Lactobacillus rhamnosus* inhibits the formation of neutrophil extracellular traps. J Immunol. 192:1870–1877.

Waite-Rees PA, et al. 1991. Characterization and expression of the *Escherichia coli* Mrr restriction system. J Bacteriol. 173:5207–5219.

Warnes GR. 2012. Gplots: Various R programming tools for plotting data. J Phycol. 3:569–575.

Xu Q, et al. 2000. Identification of type II restriction and modification systems in *Helicobacter pylori* reveals their substantial diversity among strains. Proc Natl Acad Sci U S A. 97:9671–9676.

Zdobnov EM, Apweiler R. 2001. InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17:847–848.

Zhang, et al. 2010. Complete Genome Sequence of Lactobacillus casei Zhang, a New Probiotic Strain Isolated from Traditional Homemade Koumiss in Inner Mongolia, China. 192:5268–5269.

Associate editor: Daniel Sloan