Uncovering carbohydrate metabolism through a genotype-phenotype association study of 56 lactic acid bacteria genomes

Gemma Buron-Moles¹ · Anna Chailyan¹ · Igor Dolejs¹ · Jochen Forster¹ · Marta Hanna Mikš¹,2,3

Received: 30 October 2018 / Revised: 14 February 2019 / Accepted: 14 February 2019 / Published online: 4 March 2019
© The Author(s) 2019

Abstract
Owing to their unique potential to ferment carbohydrates, both homo- and heterofermentative lactic acid bacteria (LAB) are widely used in the food industry. Deciphering the genetic basis that determine the LAB fermentation type, and hence carbohydrate utilization, is paramount to optimize LAB industrial processes. Deep sequencing of 24 LAB species and comparison with 32 publicly available genome sequences provided a comparative data set including five major LAB genera for further analysis. Phylogenomic reconstruction confirmed Leuconostoc and Pediococcus species as independently emerging from the Lactobacillus genus, within one of the three phylogenetic clades identified. These clades partially grouped LABs according to their fermentation types, suggesting that some metabolic capabilities were independently acquired during LAB evolution. In order to apply a genome-wide association study (GWAS) at the multigene family level, utilization of 49 carbohydrates was also profiled for these 56 LAB species. GWAS results indicated that obligately heterofermentative species lack 1-phosphofructokinase, required for D-mannose degradation in the homofermentative pathway. Heterofermentative species were found to often contain the araBAD operon, involved in L-arabinose degradation, which is important for heterofermentation. Taken together, our results provide helpful insights into the genetic determinants of LAB carbohydrate metabolism, and opens for further experimental research, aiming at validating the role of these candidate genes for industrial applications.

Keywords Carbohydrate metabolism · Genome-wide association study · Genotype-phenotype association study · Lactic acid bacteria · Functional genomics · Microbial genomics

Introduction
With the increased accessibility of next-generation sequencing (NGS) and third-generation sequencing technologies, including high-throughput sequencing (HTS) and whole-genome sequencing (WGS), the number of sequenced genomes has grown significantly, from approx. 31,000 in year 2014 to more than 160,000 prokaryotic genome sequences publicly available as for today (September 2018; http://www.ncbi.nlm.nih.gov/ GenomeBrowse/). HTS technology opened a new window to study microorganism diversity on the, so far unknown, genetic scale. WGS also provided a fresh, broad insight into bacteria functionality at the genomic level, which has not been previously possible. Indeed, these emerging technologies unlocked the potential for microbial genome-wide association studies (GWAS), aiming at dissecting the genetic basis of known phenotypic traits (e.g., carbohydrate metabolism) (Dutilh et al. 2013). Understanding certain genotype-phenotype associations brings unprecedented opportunities for better utilization/exploration of technologically useful microorganisms, such as starter cultures in food technology (Wu et al. 2017), but also for the engineering of novel microbial organisms and consortia in synthetic biology.
applications (Freddolino et al. 2014). Bacterial evolution, epidemiology, pathogens’ traceability during disease outbreaks, pathogenesis, antibiotic resistance, rapid detection, and food safety (Chen and Shapiro 2015; Bribic et al. 2016; Klemm and Dougan 2016; Deurenberg et al. 2017; Ruppé et al. 2017), exemplify areas where sequencing and GWAS revolutionized modern microbiology.

Lactic acid bacteria (LAB) is a naturally biodiverse, well-established group of microorganisms widely used in food industry and of well-documented impact on human health (Wu et al. 2016). LABs are “generally recognized as safe” (GRAS) in the USA and given the “Qualified Presumption of Safety” (QPS) status by the European Food Safety Authority (EFSA) (Zhang and Zhang 2014). They were isolated from various ecological niches, starting from the gastrointestinal tracts of humans, animals, and insects, through plant- and meat-based materials, to soil and water (Hammes and Hertel 2009). The ability acquired by LAB to metabolize several carbohydrates provided them competitive advantages to colonize numerous ecosystems. This variety of utilized substrates leverages a few metabolic pathways, which make LAB so unique and distinctive in terms of their fermentation potential. Based on the final fermentation product(s), LAB can be divided into two groups: homo- and hetero-fermentative, the latter being subdivided into facultatively and obligately fermentative species. Pediococcus, Lactococcus, Streptococcus, and selected Lactobacillus are considered obligate homofermentative, due to their ability to ferment only hexoses almost completely to lactic acid by the Embden–Meyerhof–Parnas (EMP) pathway, while pentoses are not degraded by all homofermenters (Pessione 2012). Facultatively heterofermentative LAB species, like Leuconostoc and certain Lactobacillus, degrade hexoses to lactic acid through EMP pathway, and can also metabolize pentoses and often gluconate as they possess both aldolase and phosphoketolase (Felig and Dellaglio 2007). The obligately heterofermentative LAB cannot utilize hexoses through the EMP pathway due to the lack of glycolytic enzyme fructose-1,6-bisphosphate aldolase. Instead, they degrade hexoses by the phosphogluconate pathway, producing not only lactic acid as the end product but also significant amounts of ethanol or acetic acid and carbon dioxide (Pessione 2012).

The industrial and technological relevance of LAB strains motivated extensive genomic studies, which have provided significant insight into their metabolism, physiology, and potential for new applications (Zhang and Zhang 2014; Bosma et al. 2017). Since 2001, when the first LAB genome was published (Bolotin et al. 2001), hundreds of LAB genomes have been sequenced and are publicly available at different assembly levels (Sun et al. 2015; Zheng et al. 2015; Wu et al. 2017; Salvetti et al. 2018). It has been demonstrated that combining phenotypic data and LAB strain-specific genetic information is an effective method for the assignment of unknown functions to specific genetic loci, especially those underlying important industrial traits, interaction with the host or niche adaptation (Siezen et al. 2011; Bayjanov et al. 2013; Ceapa et al. 2015). This has provided deeper understanding of LAB’s carbohydrate metabolism and other phenotypic traits, which includes the discoveries of, e.g., arabinose and melibiose utilization genes in Lactococcus lactis of plant origin (Bayjanov et al. 2013), candidate genes that correlate with l-sorbose and α-methyl-D-glucoside utilization in Lactobacillus rhamnosus (Ceapa et al. 2015), mannose-specific adhesin in Lactobacillus plantarum (Pretzer et al. 2005), or specific chromosomal orthologous groups (chrOGs) identified for Lactococcus lactis strain (separately for subspecies lactis and cremoris) (Siezen et al. 2011). However, in many LAB genomes, a number of carbohydrate-specific genes remain to be identified.

With the aim of better understanding carbohydrate utilization and fermentation, and ultimately help to optimize industrial processes, a genotype-phenotype association study was performed in a biodiverse group of LAB. This focused on Lactobacillus spp. potentially harboring industrial applications. Other four genera were added for comparative purposes, namely Lactococcus, Pediococcus, Leuconostoc, and Streptococcus. The genome of 24 LAB species was sequenced and compared to 32 already available, in order to define multigene families across the 5 LAB genera. Variation in multigene family size was then evaluated for association with 49 phenotypic traits, which characterize LAB carbohydrate metabolism. This analysis identified novel multigene families involved in carbohydrate degradation, as well as their relation with the respective LAB fermentation pathways.

Materials and methods

Bacterial cultures and dataset

A representative set of 50 type strains of LAB was selected to cover a bio-diverse group of microorganisms with potential industrial applications, mostly with GRAS/QPS status, different fermentation capacities, and strains’ origins (Table 1). In addition, six in-house Lactobacillus spp. isolates were also included (Carlsberg Research Laboratory, Copenhagen, Denmark). In total, 56 LAB strains have been examined comprising species from the following genera: Lactobacillus (n = 42), Lactococcus (n = 8), Leuconostoc (n = 3), Pediococcus (n = 2), and Streptococcus (n = 1). The comparative panel of LAB strains included homofermentative (n = 31), heterofermentative (n = 4), facultatively heterofermentative (n = 12), and obligately heterofermentative (n = 9) representatives. The taxonomic status of Streptococcus salivarius spp. thermophilus has been contentious. Some authors have proposed that St. salivarius spp. thermophilus is not a subspecies...
| #  | Taxon                                      | Strain code | Origin                   | Metabolism | GRAS/QPS status | NCBI accession |
|----|-------------------------------------------|-------------|--------------------------|------------|-----------------|----------------|
| 1* | Lactococcus lactis ssp. lactis 1          | CRL 0001    | Cheese                   | Ho         | +               |                |
| 2* | Lactococcus lactis ssp. lactis 2          | CRL 0002    | Cheese                   | Ho         | +               |                |
| 3* | Lactococcus lactis ssp. lactis 3          | CRL 0003    | Cheese                   | Ho         | +               |                |
| 4* | Lactococcus lactis ssp. lactis 4          | CRL 0004    | Cheese                   | Ho         | +               |                |
| 5* | Lactococcus lactis ssp. lactis 5          | CRL 0005    | Cheese                   | Ho         | +               |                |
| 6* | Lactococcus lactis ssp. lactis 6          | CRL 0006    | Cheese                   | Ho         | +               |                |
| 7* | Streptococcus salivarius ssp. thermophilus| ATCC 19258T | Unknown                 | Ho         | +               | PRJNA433425    |
| 8* | Pediococcus pentosaceus                   | ATCC33136T  | Unknown                 | Ho         | +               | PRJNA434256    |
| 9* | Lactococcus lactis ssp. lactis            | ATCC 19435T | Unknown                 | Ho         | +               | PRJNA433373    |
| 10*| Lactococcus lactis ssp. cremoris          | ATCC 19257T | Unknown                 | Ho         | +               | PRJNA433374    |
| 11*| Lactobacillus sakei ssp. sakei            | ATCC 15521T | “Moto” starter of sake   | FHe        | +               | PRJNA433475    |
| 12*| Lactobacillus amylosoticus                | DSM11664T   | Acidified beer wort      | Ho         | +               | PRJNA433476    |
| 13*| Lactobacillus delbrueckii ssp. jakobeni   | DSM 26046T  | Malted sorghum wort      | Ho         | +               | PRJNA433478    |
| 14*| Leuconostoc citreum                      | ATCC 49370T | Honey dew of rye ear     | He         | +               | PRJNA433481    |
| 15*| Leuconostoc fallax                       | ATCC        | Sauerkraut               | He         | –               | PRJNA433483    |
| 16*| Lactobacillus silagei                    | DSM 27022T  | Orchardgrass (Dactylis glomerata L.) silage | He     | –               | PRJNA433487    |
| 17*| Lactobacillus paracasei ssp. paracasei    | ATCC 25302T | Unknown                 | FHe        | +               | PRJNA43388     |
| 18*| Lactobacillus parakefiri                  | DSM 10551T  | Keif grain               | OHe        | –               | PRJNA43396     |
| 19*| Lactobacillus pentosus                   | ATCC 8041T  | Unknown                 | FHe        | +               | PRJNA43401     |
| 20*| Lactobacillus farrinimis                 | ATCC 29644T | Sausage                  | Ho         | +               | PRJNA43405     |
| 21*| Lactobacillus malfermentans              | ATCC 49373T | Beer                     | OHe        | –               | PRJNA43406     |
| 22*| Lactobacillus buchneri                   | ATCC 4005T  | Tomato pulp              | OHe        | +               | PRJNA43409     |
| 23*| Lactobacillus pasteurii                  | DSM 23907T  | Beer contaminant         | Ho         | +               | PRJNA43410     |
| 24*| Lactobacillus hilgardii                  | ATCC 8290T  | Wine                     | OHe        | +               | PRJNA43413     |
| 25*| Lactobacillus delbrueckii ssp. bulgaricus| ATCC 11482T | Yogurt                   | Ho         | +               | NC_008054      |
| 26*| Lactobacillus delbrueckii ssp. delbrueckii| ATCC 9649T  | Sour grain mash          | Ho         | +               | NZ_AZCR00000000|
| 27*| Lactobacillus delbrueckii ssp. lactis     | ATCC 12315T | Emmental cheese          | Ho         | +               | NZ_AZDE01000001|
| 28*| Lactobacillus helveticus                 | ATCC 15009T | Emmental cheese          | Ho         | +               | NZ_AZEK01000001|
| 29*| Lactobacillus dextrinicus                | ATCC 33087T | Silage                   | Ho         | +               | NZ_AYKY01000004|
| 30*| Lactobacillus mali                       | ATCC 27053T | Apple juice from cider   | Ho         | +               | NZ_AYHY01000001|
| 31*| Lactobacillus acidophilus                | ATCC 4356T  | Gastrointestinal tract   | Ho         | +               | NZ_AZCS01000001|
| 32*| Pediococcus acidilactici                 | ATCC 8042T  | Unknown                  | Ho         | +               | NZ_GL397067    |
| 33*| Lactobacillus nagelii                    | ATCC        | Partially fermented wine | Ho         | –               | NZ_AZEV01000035|
| 34*| Lactobacillus salivarius                 | ATCC 11741T | Saliva                   | Ho         | +               | NZ_GG693223    |
| 35*| Lactobacillus fermentum                  | ATCC 14931T | Fermented beets          | OHe        | +               | NZ_GG669900    |
| 36*| Lactobacillus reuteri                    | ATCC 23272T | Intestine of adult       | OHe        | +               | NC_009513.1    |
| 37*| Lactobacillus sakei                      | DSM 15831T  | Fermented meat product   | FHe        | +               | NZ_AZFG01000049|
| 38*| Lactobacillus sanfranciscensis           | ATCC 27651T | San Francisco sourdough  | OHe        | –               | NZ_AYYM01000001|
| 39*| Lactobacillus vinii                      | DSM 20605T  | Grape must, fermenting at high temperature | FHe      | –               | NZ_AYYX01000001|
| 40*| Lactobacillus amylovorus                 | ATCC 33620T | Cattle waste-corn fermentation | Ho     | +               | NZ_AZCM01000001|
| 41*| Lactobacillus composti                   | DSM 18527T  | Composting material of distilled shochu residue | FHe | –               | NZ_AZGA01000088|
| 42*| Lactobacillus farraginis                 | DSM 18382T  | Composting material of distilled shochu residue | FHe    | –               | NZ_AZY01000034|
| 43*| Leuconostoc mesenteroides ssp. cremoris  | ATCC 19254T | Hansen’s dried starter powder | He    | +               | NZ_GG693383    |
| 44*| Lactobacillus uvarum                     | DSM 19971T  | Must of Bobal grape variety | Ho   | –               | NZ_AZEG01000001|
| 45*| Lactobacillus oeni                       | DSM 19972T  | Bobal wine               | Ho         | –               | NZ_AZEH01000039|
| 46*| Lactobacillus zeae                       | ATCC 13820T | Corn steep liquor        | FHe        | +               | NZ_AZCT01000001|
of *St. salivarius*, but a separate species (Facklam 2002). Hereafter, ATCC 19258 is referred to as *St. salivarius* spp. *thermophilus*, for consistency with the name used by the provider. Type strains were purchased from American Type Culture Collection (ATCC, Manassas, VA, USA), Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSMZ, Braunschweig, Germany), Statens Serum Institut (SSI, Copenhagen, Denmark), and Microbiologics (MicroBioLogics, St. Cloud, MN, USA). All isolates were maintained as frozen stocks in 50% (w/w) glycerol tubes at −80 °C. Subsequently, the strains were activated prior to the experiments by double sub-culturing (of an aliquot of 100 μl) in 9 ml of de Man Rogosa Sharp (MRS), or M17 (Oxoid, Basingstoke, Hampshire, England) supplemented with 0.5% glucose (Merek, Darmstadt, Germany), and incubated aerobically at optimal temperature for 24–72 h.

The dataset of 32 complete genome sequences of LAB strains (at least at the scaffold level) was obtained from the NCBI (The National Center for Biotechnology Information Microbial Genomes database) (http://www.ncbi.nlm.nih.gov/genome/browse/) on 15 August 2016. The remaining 24 LAB strains, for which sequences were not publicly available or their assemblies were incomplete (i.e., deposited only at the contig level), were newly sequenced using Illumina HiSeq technology.

### Phenotypic characterization

The fermentation patterns were characterized with a standardized system, API 50 CHL (BioMérieux, Marcy l’Etoile, France) consisting of 50 biochemical tests for the study of bacterial carbohydrate metabolism. API 50 CHL was used in conjunction with API 50 CHL medium according to the manufacturer’s instructions. Briefly, active cultures of bacterial isolates or type strains were washed twice with sterile physiological saline (0.9% w/v, NaCl) and pellets were re-suspended in API 50 CHL medium (API systems, BioMérieux). Homogenized cells’ suspensions were transferred into 50-wells of the API 50 CH strips using sterile Pasteur pipettes. All wells were overlaid with sterile mineral oil (Sigma, St. Louis, MO, USA) to effect anaerobiosis. Strips were moistened and covered as recommended by the manufacturer and incubated at optimal temperature (30 or 37 °C). Changes in well color were monitored throughout the experiment, recorded after two, and verified after 10 days. For computing purposes, the results were translated to binary code, “one” for positive reaction, “zero” for negative results for all tested strains.

### Whole-genome sequencing and assembly

In this study, 24 LAB genomes have been sequenced using Illumina HiSeq 4000 technology. A paired-end library was constructed, and PE150 (HiSeq) sequencing targeting 100-fold coverage per each sample has been performed. Raw Illumina sequencing reads were trimmed using the CLC Genomics Workbench v 9.0 (https://www.qiagenbioinformatics.com/products/clc-genomics-workbench) with base calling error probability cutoff of p = 0.05 (equivalent to Phred quality score 13). Sequences shorter than 1/3 of the original read length, those containing ambiguous nucleotides, and all adapter sequences were removed. The bacterial genomes were de novo assembled using Abyss assembler (Simpson et al. 2009). For each bacterium, 12 assemblies were constructed using different *K*-mers (20, 32, 40, 50, 60, 65, 70, 75, 80, 85, 90, 96). Subsequently,
contigs assembled for each LAB strain were evaluated with in-house scripts available upon request, retaining the best assembly for each strain, according to a number of criteria calculated by QUAST (Gurevich et al. 2013). The assembly having the lowest overall rank, calculated as the sum of the individual ranks for N50, genome length, and total number of contigs, was selected as the best. The overall report on N50 of the best assembly for each bacterium is shown in Fig. 1.

**Nucleotide sequence accession numbers**

The newly sequenced genomes of 18 LAB strains, together with the corresponding gene annotations, were deposited in the GenBank database with the accession codes listed in Table 1.

**Gene prediction and functional annotation**

The genome sequences were annotated in the following steps. Firstly, ab initio gene prediction of protein-coding genes for newly sequenced LAB genomes was performed using Prodigal (ver.2.6.1) software (Hyatt et al. 2010). The gff files together with protein sequence predictions were saved for each bacterium and were added to the remaining 32 bacteria for which the protein sequences have been downloaded from NCBI. The identified proteins were searched against UniFam_Prok database (159,895 families) using HMMER (ver.3.1) software (Chai et al. 2014). Top matches of proteins passing the whole-sequence E value cutoff (< 1e-4) and coverage cutoff (aligned regions cover at least 50% of both the matched HMM and the query protein) were selected. Proteins were then annotated with the function of their best-matched UniFam families. For each genome, the proportion of gene ontology (GO) terms was provided by the CateGOrizer software (ver.3.218) (Hu et al. 2008), according to GO_slim2 annotations and assuming the “consolidated single occurrences count” option.

**Inference of orthologous gene groups**

The proteins of the 56 bacteria were clustered into orthologous groups (OGs) using OrthoFinder (Emms and Kelly 2015). First, all-versus-all searches were performed using the BLASTP algorithm (Altschul et al. 1990), with an E value threshold of $10^{-3}$. This relaxed cutoff avoids discarding putative good hits for very short sequences. At the second step, all-vs-all BLAST hits were modeled for each pairwise comparison between species, revealing and removing the gene similarity dependency on gene length and phylogenetic distance. Subsequently, the information from the reciprocal best length-normalized hit (RBNH) was used to define the lowest sequence similarity that delimits putative homologous genes. On the next step, the putative homologous gene-pairs were identified and connected in the orthogroup graph with weights

![Fig. 1](image-url) Cumulative length of contigs (Nx), as reported by QUAST. In the x-axis, 50% measures the N50 across the 24 new assemblies, which are color-coded as depicted in figure legend. Assemblies with higher Nx values are more contiguous.
given by the normalized BLAST bit scores. As final OrthoFinder step, genes were clustered into orthogroups using Markov cluster algorithm (MCL).

**Phylogenetic tree**

In order to build the phylogenetic tree, 219 ortholog groups were defined by OrthoFinder (Emms and Kelly 2015) as containing a single-copy in each of the 56 LAB species (i.e., 1:1 orthologs). For each of the 219 ortholog groups, corresponding protein sequences were aligned, using MAFFT v7.310 with default parameters (Katoh and Standley 2013). The resulting 219 alignments were concatenated, yielding a multiple alignment with 74,379 positions in total. RAxML v8.2.10 (Stamatakis 2014) automatically identified LG as best-fit model for protein evolution, prior to inferring the maximum likelihood (ML) phylogenetic tree. A hundred bootstrap replicates were run to assess node support. The tree was manually rooted, assuming *St. salivarius* ssp. *thermophilus* and *Lc. lactis* group as outgroup species, in line with their basal phylogenetic placement inferred by previous phylogenetic analyses (Makarova et al. 2006; Makarova and Koonin 2007; Salvetti et al. 2013; Sun et al. 2015).

**Association of phenotypic data to genomic data**

In order to carry out the association study, two data matrices were generated. Matrix A contained the protein count for each of the 56 species, in each of the 5932 orthogroups. Matrix B were generated. Matrix A contained the protein count for each of the 56 species, in each of the 5932 orthogroups. Matrix B contained categorical phenotypes on 49 carbohydrate substrates. This amounted to a total of 290,668 statistical tests, which were adjusted for multiple testing using two different methods: Bonferroni (Bland and Altman 1995) and false discovery rate (FDR) (Benjamini and Hochberg 1995), at two different thresholds, 0.05 and 0.01.

The R packages “gplots” (Warnes et al. 2016), “devtools” (Hadley et al. 2017), “RcolorBrewer” (Neuwirth 2014), together with a customized heatmap, have been used for results visualization (www.R-project.org). The functional annotations generated with UniFam_PROK were incorporated into the final heatmap, thus facilitating the biological interpretation of the results.

**Results**

**Genome sequencing and assembly**

In this study, the 56 LAB genomes were compared, of which 32 were obtained from NCBI database on 15th August 2016. For the remaining 24 LAB strains, whole-genome sequencing (WGS) and de novo assembly were performed, including representatives of the following species: *Lactobacillus* (*Lb*; *n* = 12), *Lactococcus* (*Lc*; *n* = 8), *Leuconostoc* (*Le*; *n* = 2), *Pediococcus* (*Pd*; *n* = 1), and *Streptococcus* (*St*; *n* = 1) (Table 1). The sequences of 6 strains (out of 24) were not released owing to commercial reasons. More importantly, among these 24 newly sequenced genomes, 2 type strains, *St. salivarius* ssp. *thermophilus* (DSM 20617) and *Le. citreum* (DSM 5577), were unavailable at NCBI at the time of manuscript preparation (Table 1). Finally, these 24 sequenced genomes also included 16 strains for which the assemblies were available, however, fragmented in multiple contigs.

After processing the HiSeq 4000 sequencing reads, de novo assemblies (*n* = 24) amounted to total lengths between 1.6 Mb for *Lb. amylobolyticus* and 3.7 Mb for *Lb. pentosus*. Strains with greater genome sizes exhibited elevated GC content, a trend only broken by the genomes with most extreme GC contents, namely the two *Lc. lactis* (< 36.02%), *Lb. farcininis* (36.4%) and *Lb. delbrueckii* ssp. *jakobsenii* (50.1%) (Supplementary Table S1). The total number of contigs ranged from 10 for *Le. citreum* to 208 for *Lc. lactis* ssp. *cremoris*, showing the maximum and minimum N50 values of 522.10 kb and 20.25 kb, respectively (Fig. 1). The fraction of missing nucleotides was limited (< 0.02%).

According to these summary statistics (Supplementary Table S1), the assembly of the 16 re-sequenced LAB strains substantially improved, having now less and often longer contigs. For 6 strains (out of 16), the N50 remained unchanged, whereas it was considerably increased for the other 10 genomes, up to 3.93-fold in *Lb. pentosus*. Greater N50 was reflected in an increased contiguity in the respective assemblies. Concomitantly, the total number of contigs was reduced for all 16 strains, especially for the *Lb. parakefiri* assembly, which was fragmented in 506 contigs and here is represented by only 101 (Supplementary Table S1). Altogether, this genomic set not only constitutes a substantial extension and improvement to the LAB genomes, but has also been consistently generated with the same methodology, minimizing potential biases, and representing therefore a suitable data set for comparative genomic studies.

**Gene prediction and functional annotation**

Ab initio gene prediction was conducted for the 24 new genome assemblies, yielding a minimum and maximum of protein-coding genes between 1648 in *Le. fallax* and 3366 in...
**L. pentosus.** This range is comparable to that observed for the 32 genome assemblies downloaded from NCBI, which ranges from 1175 for *Lb. sanfranciscensis* to 3196 in *Lb. composti*. Regarding the 16 re-sequenced LAB strains for which the assembly was improved, the number of annotated genes increased, as expected (Supplementary Table S1). Indeed, genes previously disrupted in different contigs were challenging to identify, but were discovered in this study due to more contiguous assemblies. Gene mis-identification due to assembly fragmentation is expected to be particularly important in LAB genomes, since these are generally compact, with approximately one gene per kb (Supplementary Table S1).

Leveraging UniFam functional annotations, gene ontology (GO) terms were assigned to the clear majority of protein-coding genes identified. This percentage ranged from 57% in *St. salivarius* ssp. *thermophilus* to 82% in *Lb. oeni*, an averaged 70% across 56 LAB strains. Each functionally-annotated gene was characterized by a mean of 3.7 GO terms. Within each LAB strain, the distribution of GO terms was similar, as reflected by the pie charts of four species that exhibit similar GO proportions despite belonging to two different phylogenetic clades, and having different fermentation types (Supplementary Fig. S1). Among the GO terms present in all 56 LAB strains, catabolism and carbohydrate metabolism were the two functional categories showing the highest coefficient of variation (1.0258 and 1.0231, respectively). Interestingly, this indicates that the relative abundance of genes regulating carbohydrate metabolism substantially differs across the 56 LAB species.

**Ortholog groups of genes**

Across the 56 LAB genomes, 5932 multigene families have been identified, encompassing the 96.2% of the 123,255 predicted genes. Only 4700 (3.8%) of the total genes were not assigned to any multigene family, supporting consistent gene predictions across the 56 LAB genomes. *St. salivarius* ssp. *thermophilus* showed the lowest fraction of genes (82.9%) assigned to multigene families. For the remaining species, the proportion was 91.7% or above, even reaching 100% for the species *Lc. lactis* ssp. *lactis* 6. There were 315 multigene families identified in all 56 LAB species, of which 219 contained exactly a single gene copy per species. The results are presented in the histogram (Fig. 2) summarizing the size of the ortholog groups, which peaks at 56 genes. This peak is in line with some ortholog groups conserved as single copy in the 56 LAB genomes; i.e., each species has a single gene in each family, reflecting important genes that cannot be lost, known as the core genome (Supplementary Table S2). These 219 families represent, therefore, a minimal core genome, and reveal high turnover of gene gains and losses for the rest of the families. For example, in OG0000028, corresponding to a transposase multigene family, the number of genes ranges from 0 in the majority of species to 73 in *St. salivarius* ssp. *thermophilus* type strain examined in this study.

**Phylogenetic tree**

The phylogenetic relationship between the 56 LAB strains was inferred by maximum Likelihood (ML), from the concatenated alignments of the 219 proteins that constitute their core gene set (Fig. 3). Hence, branch lengths represent the number of amino acid substitutions per site. The strains of both *St. salivarius* ssp. *thermophilus* and *Lc. lactis* group were considered as outgroups, in order to root the tree. All major nodes in the tree are well-supported by high bootstrap values, demonstrating that intergroup relationships are reflected accurately.

The phylogeny clusters LAB according to their genera, separating *Leuconostoc* and *Pediococcus* within *Lactobacillus* diversity. These genera embedded in three major phylogenetic clades, namely (A), (B), and (C) (Fig. 3). Clade A corresponds to the species forced to be outgroup, which includes strains of the *Lc. lactis* group and *St. salivarius* ssp. *thermophilus*. Their basal phylogenetic placement as outgroup is robustly supported by previous studies (Makarova et al. 2006; Makarova and Koonin 2007; Salvetti et al. 2013; Sun et al. 2015). Clade B contains all possible metabolic profiles, and is comprised of three different genera, with *Pediococcus* and *Leuconostoc* emerging as two independent groups within *Lactobacillus* species. Finally, clade C only contains facultatively heterofermentative and homofermentative strains (Fig. 3).

With a few exceptions, these major phylogenetic clades do not show clear differences, in terms of distinctive GC content, genome size, or number of proteins (Fig. 3). The highest GC content was observed in the obligately heterofermentative species *Lb. fermentum* (52.6%), and in the clade formed by homofermentative species *Lb. delbrueckii* ssp. *jakobsenii*, *bulgaricus*, *delbrueckii*, and *lactis* (approx. 50% each). Another group with high GC content is the one formed by the facultatively heterofermentative species *Lb. casei*, and *Lb. paracasei* ssp. *paracasei* (47.7%, 46.5%, and 46.5%, respectively). Notably, the last two groups belong to the major clade C. The group with the highest genome size is the one formed by the facultatively heterofermentative species *Lb. pentosus* (3.7 Mb) and *Lb. plantarum* ssp. *plantarum* (3.2 Mb), followed by the facultatively heterofermentative species *Lb. composti* (3.5 Mb), *Lb. casei* (3.1 Mb), and *Lb. paracasei* ssp. *paracasei* (3.0 Mb). The two strains with larger genomes *Lb. pentosus* (3.7 Mb) and *Lb. composti* (3.5 Mb), both facultatively heterofermentative species, also have the highest number of predicted protein-coding genes (3366 and 3196, respectively) (Fig. 3).

Mapping the carbohydrate fermentation patterns of the 56 LAB strains revealed that some phylogenetically-related
species showed diverse fermentative capabilities. More specifically, four pairs of closely clustered LAB species: \textit{Lb. nagelii-Lb. vini}; \textit{Lb. silagei-Lb. malefermentans}; \textit{Lb. hilgardii-Lb. farraginis}; \textit{Lb. alimentarius-Lb. farcininis}, are known to form different end products of lactic acid metabolic pathway (Felis and Dellarigio 2007). For example, even though \textit{Lb. alimentarius} and \textit{Lb. farcininis} showed relatively high sequence similarity, currently \textit{Lb. alimentarius} is classified as facultatively heterofermentative, while \textit{Lb. farcininis} as homofermentative. Similarly, the homofermentative LAB \textit{Lb. nagelii} clustered in the phylogenetic tree with the facultatively heterofermentative \textit{Lb. vini}.

\section*{Genotype-phenotype association study of LAB}

\subsection*{Phenotypic characterization of LAB}

The capability of 56 LAB strains to metabolize 49 carbohydrate substrates (using API 50 CHL biochemical assay) has been evaluated. The fermentation profiles of the carbohydrates demonstrated significant phenotypic diversity among tested LAB. For descriptive purposes, LAB species are hereafter defined through their representative type strains. In this regard, it is worth cautioning that some degree of strain-to-strain variation could exist within the same species (Siezen et al. 2010; Bayjano et al. 2013; Smokvina et al. 2013; Ceapa et al. 2015), as reflected by the phenotypic profiles of \textit{Lc. lactis} ssp. \textit{lactis} strains (Fig. 4). In addition, 49 carbohydrate substrates have been grouped into monosaccharides, disaccharides, polysaccharides, polyols, and salts (Fig. 4).

Monosaccharides are the most basic, fundamental units and can be classified according to their number of carbon atoms. Among common carbon sources classified as hexoses (C6), D-glucose is the only that can be degraded by all the 56 strains. Similarly, with a few exceptions in each phylogenetic clade, the LAB studied can metabolize D-galactose, D-fructose, and D-mannose. Interestingly, all strains in clades A and C were able to degrade D-fructose. Other hexoses such as D-tagatose, D-fucose, and L-fucose can be degraded by less LAB strains, if any. For example, D-tagatose, a stereoisomer of fructose (ketohexose), can only be degraded by a few LAB members of the clades B and C, basically \textit{Lactobacillus} and \textit{Pediooccus} strains. While only \textit{Lb. zeae} and \textit{Lb. composti} strains out of the 56 can degrade L-fucose, D-fucose cannot be degraded by any of our LAB studied (Fig. 4). Among the carbon sources classified as pentoses (C5), D-arabinose, L-arabinose, D-ribose, D-xyllose, and L-xyllose are included. Only \textit{Lb. zeae} can degrade L-arabinose, whereas different members in clades B and C, especially in clade B, were able to metabolize L-arabinose. Other pentoses such as D-ribose and D-xyllose cannot be degraded by \textit{Leuconostoc} strains. As L-xyllose, D-lyxose, a stereoisomer of ribose (aldopentose), cannot be metabolized by any of the LAB tested in this study (Fig. 4).

Calculating the percentage of monosaccharides showed that none of the strains of clade A degraded L-arabinose, L-sorbose, L-rhamnose, D-tagatose, and methyl-\textalpha-D-glucopyranoside, in contrast to clades B and C, where these monosaccharides are degraded by several strains (Fig. 4).

Disaccharides–carbohydrates composed of two monosaccharides joined by a glycosidic bond, such as D-maltose, D-lactose, and D-saccharose (sucrose), are common energy sources for bacterial cells. Almost all strains were able to ferment D-maltose, and those are distributed through the three major phylogenetic clades (Fig. 4). In general, all the strains can ferment D-lactose and D-saccharose with some exceptions. For example, \textit{Leuconostoc} cannot degrade D-lactose and \textit{Lactococcus} strains cannot utilize D-saccharose. Other disaccharides found in API strips were amygdalin, arbutin, esculin/ferric citrate, salicin, D-cellulobiose, D-melibiiose, and D-trehalose. The utilization of these disaccharides implies the presence of various hydrolytic enzymes that enable them to break down the substrate, such as \textbeta-glucosidase (EC 3.2.1.21) for salicin, cellulobiose and melibiose or \textbeta-amylase (EC 3.2.1.28) for trehalose.

The percentage of disaccharides utilized per clade showed that D-melibiiose and D-turanose are degraded by some strains of clades B and C, but by none of clade A (Fig. 4). Polysaccharides are carbohydrates composed of at least three monosaccharide units (e.g., D-melezitose, D-raffinose) or their longer linear or branched polymers (e.g., inulin or...
more strains of clade B, including *Pediococcus pentosaceus* and *Lactobacillus* strains. Raffinose can be hydrolyzed to D-galactose and sucrose by the enzyme α-galactosidase (EC 3.2.1.22), an enzyme not found in the human digestive tract. Starch is a polysaccharide formed by a large number of glucose units. It is interesting to note that species of clade B, including *Leuconostoc*, *Pediococcus*, and *Lactobacillus* strains, were not able to metabolize starch. The species that can ferment starch included *Lc. lactis* ssp. *lactis* 1–3, 5, and 6 (clade A), and *Lb. dextrinicus*, *Lb. gasseri*, *Lb. johnsonii*, *Lb. pasteurii*, *Lb. amylovorus*, and *Lb. gallinarum* (clade C). The enzyme responsible for hydrolysis of starch molecules
D-xylopyranoside
D-mannopyranoside

of fermentation (yellow). LAB were color-coded representing different monosaccharides and their derivatives, disaccharides and glycosides (GL), polysaccharides, polyols, and salts. The LAB strains clustering (clades A–C) corresponds to the phylogenetic tree in Fig. 3. Numbers on the right represent the percentage of carbohydrates utilized per strain, while numbers on the bottom the percentage of carbohydrates utilized per studied group of strains. Gradient scale 0% (red)-100% (green).

**Fig. 4** Heatmap generated based on 49 carbohydrate fermentation profiles of 56 lact acid bacteria (LAB) strains. All carbohydrate substrates (API 50 CHL) were categorized and grouped into monosaccharides and their derivatives, disaccharides and glycosides (GL), polysaccharides, polyols, and salts. LAB’s capacity to metabolize the corresponding carbohydrate is represented as positive (black) or lack of fermentation (yellow). LAB were color-coded representing different fermentation patterns (type), according to Felis and Dellaglio (2007): homofermentative (blue), heterofermentative (red), facultatively heterofermentative (green), and obligately heterofermentative (brown).
yielding glucose and maltose is α-amylase (E.C.3.2.1.1). Other two polysaccharides shown in Fig. 4 were inulin and glycogen. None of tested strains were able to utilize inulin, but \textit{Lb. amylovorus} and \textit{Lb. gallinarum} were capable of fermenting glycogen.

A polyol is an alcohol containing three or more hydroxyl groups, and examples include glycerol, erythritol, \textit{d}-adonitol, dulcitol, inositol, \textit{d}-mannitol, \textit{d}-sorbitol, xylitol, \textit{d}-arabitol, and \textit{l}-arabitol. The strains that were able to ferment glycerol belonged to \textit{Lb. uvarum}, \textit{Pd. acidilactici}, \textit{Pd. pentosaceus}, \textit{Lb. pentosus}, \textit{Lb. zeae}, \textit{Lb. composti}, \textit{Lb. dextrinus}, and \textit{Lb. delbrueckii ssp. lactis}. No strain in clade A was capable of degrading glycerol, commonly used as carbon source in biotechnology. Among all the polyols, \textit{d}-mannitol is the sugar alcohol most widely fermented by LAB members of each clade, followed by \textit{d}-sorbitol with the exception of members in clade A, which cannot degrade \textit{d}-sorbitol. On the other hand, while three strains were capable of fermenting \textit{d}-arabitol (\textit{Lb. plantarum ssp. plantarum}, \textit{Lb. silagei}, and \textit{Lb. zeae}), only \textit{Lb. zeae} was able to degrade inositol. As for the remainder of the polyols, 56 analyzed LAB strains cannot ferment erythritol, \textit{d}-adonitol, dulcitol, xylitol, and \textit{l}-arabitol (Fig. 4).

The salts characterized included potassium gluconate, potassium 2-ketogluconate, and potassium 5-ketogluconate (Fig. 4). Potassium gluconate is the potassium salt of gluconic acid, which is obtained from glucose by fermentation and subsequent neutralization with a potassium source. Of the three salts characterized, potassium gluconate was the one most widely fermented by LAB members of each clade, followed by potassium 5-ketogluconate with the exception of members in clade A, which cannot metabolize this salt. Finally, five strains were capable of metabolizing potassium 2-ketogluconate being \textit{Le. citreum} and \textit{Lb. kefiri} (clade B), \textit{Lb. zeae}, \textit{Lb. casei}, and \textit{Lb. composti} (clade C) (Fig. 4).

### Association of phenotypic data to genomic data

The genotype-phenotype association analysis was performed to identify significant associations between gene family expansions/contractions in the 56 LAB strains in relation to their carbohydrate metabolism, based on three different correction methods for multiple testing, namely Bonferroni \((p < 0.05)\), false discovery rate (FDR) \((p < 0.01)\), and FDR \((p < 0.05)\).

The results for Bonferroni correction, represented as a heatmap in Fig. 5, showed significant associations of 17 ortholog groups with 7 phenotypic traits related to carbohydrate metabolism. In Fig. 5, the 56 LAB strains clustered according to similarities in their carbohydrate metabolism, partially recovering the phylogenetic clades in Fig. 3. This implies that some of these metabolic traits evolved along the diversification of the 56 LAB species. Three clusters were observed, but not clearly separated based on fermentation type. While the first cluster comprises from \textit{Lb. sakei} ssp. \textit{sakei} to \textit{Lb. alimentarius}, and excludes obligately heterofermentative species, the second cluster only includes homofermentative species, revealing similarities in their carbohydrate metabolism. By contrast, in the third cluster, the heterofermentative species grouped together, including facultatively and obligately, showing similar fermentation profiles. Among the strains of all the homofermentative species, we observed a general lack of genes belonging to multigene families associated with the phenotypes \textit{l}-arabinose and \textit{d}-melezitose. Similarly, the strains of obligately heterofermentative species were depleted of genes belonging to multigene families associated with the phenotypes \textit{d}-mannose, arbutin, and salicin (Fig. 5).

The \textit{Lactococcus} genus clustered together, as these strains basically showed the same carbohydrate profile, with a few exceptions. By contrast, species of the \textit{Pediococcus} genus did not cluster according to their metabolic profiles. The main difference is that \textit{Pd. acidilactici} did not metabolize \textit{d}-saccharose, in contrast to \textit{Pd. pentosaceus}. In addition, the orthogroup OG0000939, significantly associated with the metabolism of \textit{d}-saccharose \((p\ value = 6.19e-03)\), was present only in \textit{Pd. pentosaceus}. Similarly, \textit{Leuconostoc} species did not cluster together, since \textit{Le. citreum} showed a slightly different metabolic profile compared to \textit{Le. fallax} and \textit{Le. mesenteroides ssp. cremoris}. In particular, \textit{Le. fallax} and \textit{Le. mesenteroides ssp. cremoris} have lost their ability to metabolize arbutin, salicin, \textit{d}-cellobiose, \textit{d}-mannose, and \textit{l}-arabinose. The orthogroups significantly associated with these metabolisms were OG0000070, OG0000204, OG0000133, OG0000160, OG0000409, OG0001748, and OG0001480 (Fig. 5).

Differences between the three clusters defined by the heatmap lied down mainly in the ability of the strains to metabolize arbutin, salicin, and \textit{d}-cellobiose (Fig. 5). Strains belonging to cluster 1 were often able to metabolize these three carbohydrates, in contrast to those belonging to clusters 2 and 3. Several orthogroups showed concomitant variations in the number of family members. For example, the number of genes in OG0000014 was at least 11 among strains that metabolize salicin and \textit{d}-cellobiose, while merely 1 for the remaining. In fact, \textit{St. salivarius} ssp. \textit{thermophilus}, \textit{Lb. reuteri}, \textit{Lb. sanfranciscensis}, and \textit{Lb. fermentum} did not metabolize salicin and \textit{d}-cellobiose, and have thus no genes belonging to OG0000014. Genes within this orthogroup were functionally annotated as RpiR family transcriptional regulator. Likewise, orthogroup OG0000070 was significantly associated with the metabolism of arbutin \((p\ value = 0.03)\) and salicin \((p\ value = 0.01)\), and annotated as DeoR family transcriptional regulator. \textit{Lb. zeae} showed the largest number of genes in OG0000070, with eight members. Interestingly, all strains in clusters 2 and 3 lost their ability of fermenting arbutin and salicin; however, only heterofermentative species in cluster 3 have almost no genes belonging to this orthogroup. Another orthogroup
| Orthogroup annotation | Orthogroup IDs |
|-----------------------|---------------|
| OG00000017            |               |
| OG00000009             |               |
| OG00000010             |               |
| OG00000001             |               |
| OG00000006             |               |
| OG00000007             |               |
| OG00000002             |               |
| OG00000004             |               |
| OG00000008             |               |
| OG00000009             |               |
| OG00000001             |               |
| OG00000010             |               |
| OG00000006             |               |
| OG00000011             |               |
| OG00000007             |               |
| OG00000010             |               |
| OG00000009             |               |
| OG00000001             |               |

**Carbohydrates**

| Carbohydrates | L. acidophilus | L. acidophilus | L. acidophilus | L. acidophilus | L. acidophilus | L. acidophilus |
|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| D-Mannose     | 0             | 0             | 0             | 0             | 0             | 0             |
| D-Melezitose  | 0             | 0             | 0             | 0             | 0             | 0             |
| D-Mannose     | 0             | 0             | 0             | 0             | 0             | 0             |
| D-Mellose     | 0             | 0             | 0             | 0             | 0             | 0             |
| L-Arabinose   | 0             | 0             | 0             | 0             | 0             | 0             |
| L-Ribulose-5-phosphate 4-epimerase | 0 | 0 | 0 | 0 | 0 | 0 |

**Carbohydrates associated with OG**

| Carbohydrates associated with OG | L. acidophilus | L. acidophilus | L. acidophilus | L. acidophilus | L. acidophilus | L. acidophilus |
|--------------------------------|---------------|---------------|---------------|---------------|---------------|---------------|
| D-Mannose | 0 | 0 | 0 | 0 | 0 | 0 |
| D-Melezitose | 0 | 0 | 0 | 0 | 0 | 0 |
| D-Mannose | 0 | 0 | 0 | 0 | 0 | 0 |
| D-Mellose | 0 | 0 | 0 | 0 | 0 | 0 |
| L-Arabinose | 0 | 0 | 0 | 0 | 0 | 0 |
multigene family was absent in cluster 1, with OG0000017 and OG0000026 harboring the greatest number of genes. In fact, *Lb. plantarum* ssp. *plantarum* and *Lb. alimentarius* contained 12 members in orthogroups OG00000017 and OG00000026, respectively. Regarding cluster 2, these orthogroups were mostly present in homofermentative species, in comparison to heterofermentative strains in cluster 3, where these families were nearly absent, or even absent for orthogroup OG0000160. While the orthogroup OG0000160 was specifically associated with the d-mannose metabolism, OG0000017 and OG0000026 were associated with salicin, and salicin and d-cellobiose, respectively. Genes belonging to all these orthogroups were functionally annotated as phosphotransferase (PTS) sugar transporters (Fig. 5). The capacity of utilizing salicin, d-cellobiose and d-mannose indicates that the genome of these LAB encode for proteins acting as specific transporters, such as PTS-salicin, PTS-cellobiose, and PTS-mannose.

The multigene family annotated as 6-phospho-β-glucosidase (OG0000021) was significantly associated with the ability to utilize d-mannose and salicin, as sucrose-6-phosphate hydrolase (OG0000939) was with d-saccharose, 1-phosphofructokinase (OG0000409) with d-mannose, and α-mannosidase (OG0001472) with d-cellobiose (Fig. 5). All LAB in cluster 1 and 2 have different copies of the 6-phospho-β-glucosidase and 1-phosphofructokinase genes. By contrast, heterofermentative LAB strains in cluster 3 have lost their 6-phospho-β-glucosidase and 1-phosphofructokinase genes, excepting *Lb. brevis*, *Le. fallax*, and *Lb. silagei*, which retain one gene of the 6-phospho-β-glucosidase. In cluster 1, the sucrose-6-phosphate hydrolase multigene family was only absent in *Lb. curvatus*, *Le. lactis* group, and *Pd. acidilactici* strains, whereas in cluster 2 and 3, different homofermentative and heterofermentative LAB species lost their ability to ferment D-saccharose, as well as their associated genes (OG0000939). All homofermentative and heterofermentative strains in clusters 2 and 3 lost the ability to ferment D-cellobiose, as well as the α-mannosidase genes (OG0001472), which are not found in any member of these clusters (Fig. 5).

Other statistically significant genotype-phenotype association were identified between L-arabinose and the following multigene families OG0001662, OG0001748, and OG0001480, which were annotated as L-arabinose isomerase (EC 5.3.1.4; p value = 1.41e-05), L-ribulokinase (EC 2.7.1.16; p value = 5.42e-04), and L-ribulose-5-phosphate 4-epimerase (EC 5.1.3.4, p value 4.63e-03). As for the multigene family L-arabinose isomerase (OG0001662), *Lb. silagei* that metabolize L-arabinose contained four gene copies, *Lb. buchneri* contained two, and the rest contained one gene each. In general, correlation between the presence of L-arabinose isomerase, L-ribulokinase, and L-ribulose-5-phosphate 4-epimerase genes and the fermentation type was identified, whereby heterofermentative strains showed a greater number of these genes than the homofermentative ones. Importantly, species metabolizing and non-metabolizing L-arabinose are interspersed among *Lactobacillus* and *Lactococcus*, as well as *Leuconostoc* genus. This confirms that this genotype-phenotype association is not spuriously reflecting the underlying phylogenetic relationship, and reveals that multiple species underwent independent changes in the arabinose metabolism. Remarkably, and in contrast to the majority of carbohydrates, the isomer L-arabinose is more common than D-arabinose in nature. Collectively, this suggests that only those LAB strains able to transform L- into D-arabinose, due to the expansion of these three multigene families, are also able to subsequently metabolize D-arabinose.

Moreover, the genotype-phenotype association analysis revealed that two orthogroups predicted as unknown hypothetical proteins (functionally unannotated families; OG0001154 and OG0002201) were involved in a series of carbohydrate metabolic functions (Fig. 5). The orthogroup OG0001154 was significantly associated with the utilization of arbutin, salicin, and d-cellobiose. In cluster 1, only a few strains lost the genes of this orthogroup, whereas it was absent among homofermentative and heterofermentative species of clusters 2 and 3, with the exception of *Lb. oeni*, which contained one gene only. The other orthogroup of unidentified function was OG0002201, and it was significantly associated with D-melezitose metabolism. With the exception of *Lb. oeni* (homofermentative), only heterofermentative strains contained OG0002201 genes.

The results above proved that this method can be used to identify in which metabolic pathways the unannotated multigene families were involved, and consequently to gain deeper insights into their function.
Discussion

LABs are used worldwide industrially in the manufacture of fermented foods and beverages (Holzapfel and Wood 2014), because their metabolic products improve nutritional value, organoleptic properties, as well as the microbiological safety of food products (Thierry et al. 2015). Since LAB use preferably carbohydrates as the primary carbon and energy sources, understanding the genetic basis of their utilization in metabolic pathways is essential to optimize the fermentative processes.

In this study, the comparison of LAB genomes revealed 219 single-copy genes shared among the 56 strains, which is known as the core genome. In another study comparing 20 Lactobacillus genomes, the core gene set was estimated to include 383 genes (Kant et al. 2011). Increasing the number of genomes to 213 reduced the core gene set to 73 genes (Sun et al. 2015). The core gene set is not only reduced with the number of compared genomes but also with the phylogenetic distances between them and the incompleteness of genome assemblies. For instance, Lukjancenko and colleagues (Lukjancenko et al. 2012) compared 81 LAB genomes from 6 different genera, and found that the core genome contains 63 genes. A greater core gene set of 172 single-copy genes was estimated in another study including 174 type strains from only 2 genera, Lactobacillus and Pediococcus (Zheng et al. 2015). In this study, the core gene set of 219 orthologous groups was slightly greater than those estimated in previous studies, probably reflecting that all genomes investigated were assembled at least at the scaffold level, precisely to improve gene completeness.

Functional annotation revealed that these 219 genes are mostly involved in nitrogen metabolism (data not shown). This suggests that these core genes are essential. Single-copy genes were unlikely duplicated or lost during evolution, implying they are true orthologs. Therefore, they reflect the divergence and phylogenetic relationship between the LAB species. Leveraging these 219 core genes, the phylogenetic tree was constructed, rooted assuming St. salivarius ssp. thermophilus and Lc. lactis as outgroup species (Fig. 3). Early after the first genomic studies, Lactobacillus species were found to belong to two separate clades (Makarova et al. 2006; Makarova and Koonin 2007; O’Sullivan et al. 2009; Salvetti et al. 2018). Our maximum-likelihood (ML) phylogenetic reconstruction recovered these two major Lactobacillus clades, referred as clade B and C, with high bootstrap support (Fig. 3). Nevertheless, the species that comprised these two clades have been traditionally controversial. The maximum phylogenetic uncertainty is associated with the species of the Lb. casei group, which harbor a large diversity in gene content across strains, including horizontal gene transfer (HGT) from Lactobacilli species of clade B (Broadbent et al. 2012). Indeed, multiple studies placed Lb. casei species within clade C (Makarova et al. 2006; Makarova and Koonin 2007; Claesson et al. 2008; Zhang et al. 2011; Salvetti et al. 2013; Sun et al. 2015), whereas a few others within clade B (Liu et al. 2008; O’Sullivan et al. 2009; Lukjancenko et al. 2012; Zheng et al. 2015). Our reconstruction based on a greater number of core genes supports the former hypothesis, embedding the Lb. casei group within group C (Fig. 3). Our phylogenomic tree also support Leuconostoc and Pediococcus as emerging within the Lactobacillus species of clade B, confirming previous results (Makarova et al. 2006; Makarova and Koonin 2007; Claesson et al. 2008; Liu et al. 2008; Zhang et al. 2011; Salvetti et al. 2013; Sun et al. 2015).

Mapping the fermentative types of 56 LAB revealed that some phylogenetically related species have different fermentation capabilities, a phenomenon that occurred within both, clades B and C, and that has been previously reported (Hammes and Vogel 1995; Felis and Dellaglio 2007; Pot et al. 2014), but disregarded by Zheng and colleagues (Zheng et al. 2015) in a more recent study. This incongruence could be explained by misclassifying the fermentation capabilities of a very few Lactobacillus strains. For example, Lb. vini has been assigned as a homofermentative species (Zheng et al. 2015), despite it produces small amounts of ethanol from arabinose and ribose (Felis and Dellaglio 2007; Endo and Dicks 2014), which were indeed degraded by Lb. vini in this study (Fig. 4).

A similar approach has been recently adopted to map lifestyle transitions (from free-living to host-adapted) onto certain phylogenetic nodes (Duar et al. 2017). Interestingly, the phylogenetic clades in this study mirror the grouping of these lifestyles. More specifically, most obligately heterofermentative strains of phylogenetic clade B, such as Lb. brevis and Lb. parakefiri, were classified as free-living (Duar et al. 2017). In contrast, all homofermentative strains of clade C, and for which lifestyle information is available, were classified as adapted to vertebrates (Fig. 3). Taken together, this suggests that fermentation types and lifestyles strongly shaped LAB genomic evolution. For example, and with the most notable exception of the Lb. delbrueckii group, GC content tends to be lower in homofermentative species (Fig. 3), most of which are host-adapted bacteria (Duar et al. 2017). Drastic variation in GC content is usually associated with gene acquisition through HGT (Garcia-Vallvé et al. 2000; Guindon and Perrière 2001), suggesting that heterofermentative capabilities could have been acquired from phylogenetically distant species. Regardless of the mechanism of gene acquisition, the gene repertoire of Lactobacillales species is known to be highly dynamic (Zheng et al. 2015; Papizadeh et al. 2017), as reflected by the limited number of core genes found in this and other studies (O’Sullivan et al. 2009; Kant et al. 2011; Lukjancenko et al. 2012). Therefore, most of the LAB phenotypic variation is expected to result from gene gain and loss processes. In Lactobacillus, some genes related to carbohydrate utilization, for example, have
been inferred to be acquired by HGT (Barrangou et al. 2003; Klaenhammer et al. 2005). Likewise, Zheng and colleagues (Zheng et al. 2015) found that genes lost in heterofermentative species are mostly associated with carbohydrate metabolism. Collectively, this supports gene variation as a major source of functional innovation in carbohydrate metabolism, and justifies genotype-phenotype association studies at the orthogroup level. Phenotyping different features of the carbohydrate metabolism revealed that LAB are separated in three metabolic clusters (Fig. 4). These clusters do not fully recapitulate the exact phylogenetic clades. This implies that some of these metabolic traits evolved along the diversification of the 56 LAB species, but others might have been acquired through HGT or independently lost in multiple lineages. Interestingly, the presence of certain genes does not always imply evidence for the degradation of the associated carbohydrate. For example, St. thermophilus was not able to metabolize salicin despite harboring multiple genes that in the other species were associated with salicin degradation (e.g., OG0000017, in Figs. 4 and 5). This is in line with St. thermophilus undergoing a massive pseudogenization process (gene inactivation), which is also observed in Lb. helveticus and Lb. delbrueckii (Bolotin et al. 2004; O’Sullivan et al. 2009). Conversely, Lb. sanfranciscensis was able to utilize diverse sources of carbohydrates, such as D-mannose (Fig. 4), albeit it lacked all genes found to be significantly associated with its degradation (Fig. 5). These findings emphasize that phenotypes cannot be always inferred from the presence/absence of specific genes, but require explicitly integrating phenotypic information. Yet, this study revealed another interesting pattern. Metabolic cluster 3 contained all obligately heterofermentative strains (Fig. 5), and the association results indicated that all members within this cluster lacked 1-phosphofructokinase (PFK), in contrast to the other two metabolic clusters, where this gene was present at least as single-copy (Fig. 5). PFK was found to be a key gene distinguishing hetero- and homofermentative species in previous studies (Morita et al. 2008; Zheng et al. 2015), and more recently confirmed based on phylogenetic framework (Salvetti et al. 2018). The association study implemented in this work provides further insights, revealing that PFK is indispensable not only for mediating fructose-6-P phosphorylation but also for D-mannose degradation during homofermentation. Since PFK acts on the homofermentative pathway, species lacking this enzyme produce CO₂, ethanol, and lactate, through the heterofermentative pathway. Besides the PFK absence, obligately but also facultative heterofermentative species are partially characterized by the presence of three genes, represented by OG0001662, OG0001748, and OG0001480 (Fig. 5). These genes are associated with the degradation of L-arabinose, and annotated as L-arabinose isomerase, L-ribulose kinase, and ribulose phosphate epimerase, respectively. BLAST homology searches suggested that these genes correspond to araA, araB, and araD, conforming the araBAD operon. In Escherichia coli, this operon has been described to degrade L-arabinose into xylulose-5P, an essential compound in the heterofermentative pathway (Schleif 2010). Therefore, obligately heterofermentative strains are characterized by the absence of 1-phosphofructokinase, and to a lesser extent by the presence of the three enzymes forming the araBAD operon (Fig. 5).

Metabolic clusters 1 and 2 excluded obligately heterofermentative strains, and only contained homofermentative and facultative heterofermentative LAB (Fig. 5). Particularly, cluster 1 showed an increased number of genes annotated as (i) carbohydrate phosphotransferase system (PTS); (ii) transcriptional regulators, such as the RpiR and DeoR families, and a transcription antiterminator; and (iii) glycosyl hydrolyases, namely 6-phospho-β-glucosidase and sucrose-6-phosphate hydrolase. Interestingly, all these genes participate in the uptake of different carbohydrates, and their subsequent hydrolysis. PTS proteins are known to first import and phosphorylate carbohydrate substrates, such as salicin, D-cellobiose, and D-mannose. According to the results of this study, their degradation is significantly associated with the number of PTS transporters (Fig. 5). This is in line with other studies that have also described variation in PTS abundance across species, depending on their fermentation capabilities. More specifically, PTS components have been found to be numerous in homofermentative strains, while rarely in heterofermentative strains (Reizer et al. 1988; Bolotin et al. 2001; Michlmayr and Kneifl 2014). Despite being rare, PTS components had been described in only two obligately heterofermentative species, namely Lb. brevis (Saier et al. 1996) and Lb. fermentum (Zhang et al. 2013). This study corroborates PTS presence in Lb. brevis and Lb. fermentum, but also expands this observation to the type strains of the following obligately heterofermentative species: Lb. buchneri, Lb. parakefiri, Lb. kefiri, Lb. hilgardii, Lb. malefermentans, and Lb. reuteri (Fig. 5).

Following carbohydrate uptake by PTS proteins, transcriptional regulators modulate the degradation of β-glucosides like D-cellobiose, as well as aromatic β-glucosides, such as salicin and arbutin (Bardowski et al. 1994; Sonowal et al. 2013). All three carbohydrates, D-cellobiose, salicin, and arbutin, were shown in this study to be significantly associated with transcriptional regulators, supporting their role in the degradation of these β-glucosides (Fig. 5). Finally, glycosyl hydrolyases are known to be also involved in the breakdown of β-glucosides in the cytoplasm. This association was also reflected in our results, as two glycosyl hydrolyases were associated with the metabolism of D-mannose, salicin, and D-saccharose (Fig. 5). Interestingly, obligately heterofermentative strains within metabolic cluster 3 are often devoid of these two glycosyl hydrolyases (Fig. 5), in line with previous studies finding a lower amount of glycosyl hydrolyases in obligately heterofermentative Lactobacillus species (Michlmayr et al. 2013). More interestingly, PTSs, transcriptional regulators, and glycosyl hydrolyases are often arranged in gene clusters,
suggesting they orchestrate the co-regulation of carbohydrate metabolism (Morita et al. 2009), as potentially reflected by the results of the present study found for cluster 1 (Fig. 5).

In conclusion, genotype-phenotype association enabled the identification of genes involved and responsible for the transport and metabolism of specific carbon sources. Functional annotation of those orthogroups shed light into the metabolic pathways underlying different fermentative capabilities of selected LAB strains. As an example, PFK absence in obligately heterofermentative strains mirrored its essential function in the homolactic fermentation pathway. Complementarily, the results of this study also revealed novel associations between unannotated orthogroups (i.e., hypothetical proteins) with essential traits of the LAB carbohydrate metabolism, including OG0002201 (involved in D-melezitose degradation) and OG0001154 (with arbutin, salicin, and D-cellobiose). This highlights how genotype-phenotype associations can provide deeper insights into the function of hypothetical proteins. Further validation is required to confirm these findings, as well as to elucidate the potential roles of these genes in carbohydrate metabolism, which is of paramount importance for industrial applications.

Acknowledgements We thank Dr. Kasper Nielsen (Carlsberg Research Laboratory, Copenhagen, Denmark) for his helpful comments on the manuscript.

Author contribution statements G.B.M and M. M designed the project. G.B.M and I. D performed the phenotypic characterization. A. C and G.B.M performed the computational analyses. G.B.M and M. M wrote the paper with input from all authors. M.M. and J. F supervised the project.

Funding This study was supported by Carlsberg Foundation (CF14-0450) under the Project “Tapping the biodiversity of Lactobacilli for novel fermented beverages.”

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Alsulch SF, Gish W, Pennsylvania T, Park U (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Bardowski J, Ehrlich SD, Chopin A (1994) BglR protein, which belongs to the Bagl family of transcriptional antiterminators, is involved in β-glucoside utilization in Lactococcus lactis. J Bacteriol 176(18): 5681–5685

Barrangou R, Altermann E, Hutkins R, Cano R, Klæenhammer TR (2003) Functional and comparative genomic analyses of an operon involved in fructooligosaccharide utilization by Lactobacillus acidophilus. Proc Natl Acad Sci 100(15):8957–8962. https://doi.org/10.1073/pnas.1332765100

Bayjanov JR, Starrenburg MJC, van der Sijde MR, Siezen RJ, van Hijum SAFT (2013) Genotype-phenotype matching analysis of 38 Lactococcus lactis strains using random forest methods. BMC Microbiol 13:68. https://doi.org/10.1186/1471-2180-13-68

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc 57(1):289–300

Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. BMJ 310(6973):170. https://doi.org/10.1136/bmj.310.6973.170

Bolotin A, Quinquis B, Renaud P, Sorokin A, Ehrlich SD, Kulakauskas S, Lapidus A, Goltsman E, Mazur M, Pusch GD, Fonstein M, Overbeek R, Kyrpides N, Purnelle B, Prozzi D, Ngui K, Masuy D, Hancy F, Burreau S, Boutry M, Delcour J, Goffeau A, Hols P (2004) Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus. Nat Biotechnol 22(12): 1554–1558. https://doi.org/10.1038/nbt1034

Bolotin A, Wincker P, Mauger S, Jaillon O, Malarme K, Weissenbach J, Ehrlich SD, Sorokin A (2001) The complete genome sequence of the lactic acid bacterium Lactococcus lactis ssp lactis IL1403. Genome Res 11:731–753. https://doi.org/10.1101/gr.169701

Bosma EF, Forster J, Nielsen AT (2017) Lactobacilli and pediococci as versatile cell factories—evaluation of strain properties and genetic tools. Biotechnol Adv 35(4):419–442. https://doi.org/10.1016/j.biotechadv.2017.04.002

Brbic M, Piskorec M, Vidulin V, Krisko A, Smuc T, Supek F (2016) The landscape of microbial phenotypic traits and associated genes. Nucleic Acids Res 44(21):10074–10090. https://doi.org/10.1093/nar/gkw964

Broadbent JR, Neeno-Eckwall EC, Stahl B, Tandee K, Cai H, Morovic W, Horvath P, Heidenreich J, Perna NT, Barrangou R, Steele JL (2012) Analysis of the Lactobacillus casei supragenome and its influence in species evolution and lifestyle adaptation. BMC Genomics 13:533. https://doi.org/10.1186/1471-2164-13-533

Ceapa C, Lambert J, van Limpt K, Wels M, Smokvina T, Knol J, Kleerebezem M (2015) Correlation of Lactobacillus rhamnosus genotypes and carbohydrate utilization signatures determined by phenotype profiling. Appl Environ Microbiol 81(16):5458–5470. https://doi.org/10.1128/AEM.00851-15

Chai J, Kora G, Ahn TH, Hyatt D, Pan C (2014) Functional phylogenomics analysis of bacteria and archaea using consistent genome annotation with UniFam. BMC Evol Biol 14(1):1–13. https://doi.org/10.1186/s12862-014-0207-y

Chen PE, Shapiro BJ (2015) The advent of genome-wide association studies for bacteria. Curr Opin Microbiol 25:17–24. https://doi.org/10.1016/j.mib.2015.03.002

Claesson MJ, van Sinderen D, O’Toole PW (2008) Lactobacillus phylogenomics—towards a reclassification of the genus. Int J Syst Evol Microbiol 58(12):2945–2954. https://doi.org/10.1099/ijs.0.65848-0
Deurenberg RH, Bathoom E, Chlebowicz MA, Couto N, Ferdous M, Garcia-Cobos S, Koosstra-Smid AMD, Raangs EC, Rosema S, Veloo ACM, Zhou K, Friedrich AW, Rossen JW (2017) Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol 243:16–24. https://doi.org/10.1016/j.jbiotec.2016.12.022

Duar RM, Lin XB, Zheng J, Martino ME, Grenier T, Pérez-Muñoz ME, Leulier F, Gänzle M, Walter J (2017) Trends in evolution: history and natural history of the genus *Lactobacillus*. FEMS Microbiol Rev 41:527–548. https://doi.org/10.1093/femsre/fux030

Dutilh BE, Backus L, Edwards RA, Wels M, Bayjanov JR, Sacha AF (2013) Explaining microbial phenotypes on a genomic scale: GWAS for microbes. Brief Funct Genomics 12(4):366–380. https://doi.org/10.1093/bfgp/elct008

Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16(1):1–14. https://doi.org/10.1186/s13059-015-0721-2

Endo A, Dicks LMT (2014) Physiology of the LAB. In: Holzapfel WH, John Wiley & Sons, Hoboken, NJ, pp 13–20

Facklam R (2002) What happened to the Streptococci: overview of taxonomic and nomenclature changes. Clin Microbiol Rev 15(4):613–630. https://doi.org/10.1128/CMR.15.4.613-630.2002

Felis GE, Dellaglio F (2007) Taxonomy of lactobacilli and bifidobacteria. Curr Issues Intest Microbiol 8:44–61

Freddolino PL, Goodarzi H, Tavazoiea S (2014) Revealing the genetic and functional underpinnings of species within the *Lactobacillus* genus. PLoS Pathog 10(8):e1004295. https://doi.org/10.1371/journal.ppat.1004295

Guindon S, Perrière G (2001) Intragenomic base content variation is a basis for natural bacterial phenotypic divergence. J Bacteriol 190(4):1754–1758. https://doi.org/10.1128/JB.00135-01

Hammes WP, Vogel RF (1995) The genus *Lactobacillus*. In: Holzapfel WH, Food Microbiol. Academic & Professional, London, pp 249–353

Hammes WP, Hertel C (2009) Genus I. *Lactobacillus*. In: Wood BJB, Bergey's Manual of Systematic Bacteriology Vol 2. John Wiley & Sons, Hoboken, NJ, pp 13–511

Hu Z, Bao J, Reecy JM (2008) CateGOrizer: a web-based program to batch analyze gene ontology classification categories. Online J Bioinforma 9(2):108–112

Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119

Kant R, Blom J, Palva A, Siezen RJ, de Vos WM (2011) Comparative genomics of *Lactobacillus*. Microb Biotechnol 4(3):323–332. https://doi.org/10.1111/j.1751-7915.2010.00215.x

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010

Klaenhammer TR, Barrangou R, Buck BL, Azcarate-Peril MA, Altermann E (2005) Genomic features of lactic acid bacteria effecting bioprocessing and health. FEMS Microbiol Rev 29:393–409. https://doi.org/10.1016/j.femsre.2005.04.007

Klemm E, Dougan G (2016) Advances in understanding bacterial pathogenesis gained from whole-genome sequencing and phylogenetics. Cell Host Microbe 19(5):599–610. https://doi.org/10.1016/j.chom.2016.04.015

Koivisto T, Palva A, Siezen RJ, de Vos WM (2011) Comparative genomics of *Lactobacillus*. Microb Ecol 63:651–673. https://doi.org/10.1007/s00248-011-9948-y

Kumar KS, Koonin EV (2007) Evolutionary genomics of lactic acid bacteria. J Bacteriol 189(4):1199–1208. https://doi.org/10.1128/JB.01351-06

Kumar KS, Slesarev A, Wolf Y, Sornin A, Mirkin B, Koonin EV, Pavlov A, Pavlova N, Karamychev V, Polouchine N, Shakhova V, Grigoriev I, Lou Y, Rohksar D, Lucas S, Huang K, Goodstein DM, Hawkins T, Plengvidhya V, Welker DL, Hughes JE, Goh Y, Benson A, Baldwin KA, Lee J-H, Diaz-Muñiz I, Dosti B, Smevaniv V, Wechter W, Barabote R, Lorea GL, Altermann E, Barrangou R, Ganesan B, Xie Y, Rawsthorne H, Tamir D, Parker C, Breidt F, Broadbent JR, Hutkins R, O’Sullivan D, Steele JL, Unlu G, Saier MH, Klaenhammer TR, Richardson P, Kozyavin S, Weimer BC, Mills DA (2006) Comparative genomics of the lactic acid bacteria. Proc Natl Acad Sci U S A 103(42):15611–15616. https://doi.org/10.1073/pnas.0607117103

Michlmayr H, Hall J, Lorenz C, Böhm dorfer S, Rosenau T, Kneifel W (2013) Arabinoxylan oligosaccharide hydrolysis by family 43 and 51 glycosidases from *Lactobacillus brevis* DSM 20054. Appl Environ Microbiol 79(21):6747–6754. https://doi.org/10.1128/AEM.02130-13

Michlmayr H, Kneifel W (2014) β-Glucosidase activities of lactic acid bacteria: mechanisms, impact on fermented food and human health. FEMS Microbiol Lett 352(1):1–10. https://doi.org/10.1111/1574-6968.12348

Morita H, Toh H, Osumi K, Taylor TD, Igimi S, Hattori M, Morita H, Toh H, Oshima K, Murakami M, Taylor TD, Igimi S, Hattori M (2008) Comparative genome analysis of *Bifidobacterium* ATCC 25925 and related probiotic genera. Microb Biotechnol 4(3):323–332.

O’Sullivon O, O’Callaghan J, Sangrador-Vegas A, McAuliffe O, Slattery L, Kaela P, Callanan M, Fitzgerald GF, Ross RP, Beresford T (2009) Comparative genomics of lactic acid bacteria reveals a niche-specific gene set. BMC Microbiol 9(1):50. https://doi.org/10.1186/1471-2164-9-50

Pot B, Felis G, De Bruyne K, Tsakalidou E, Papadimitriou K, Leisner J, Vandamme P (2014) The genus *Lactobacillus*. In: Holzapfel WH, Wood BB (eds) Lactic acid bacteria: biodiversity and taxonomy. John Wiley & Sons, Hoboken, NJ, pp 19–54
Reizer J, Peterkofsky A, Romano AH (1988) Evidence for the presence of heat-stable protein (HPr) and ATP-dependent HPr kinase in heterofermentative lactobacilli lacking phosphoenolpyruvate: glycolic phosphotransferase activity. Proc Natl Acad Sci U S A 85(7):2041–2045. https://doi.org/10.1073/pnas.85.7.2041

Ruppé E, Cherkouaoui A, Lazarevic V, Emonet S, Schrenzel J (2017) Establishing genotype-to-phenotype relationships in bacteria causing hospital-acquired pneumonia: a prelude to the application of clinical metagenomics. Antibiotics 6(4):30. https://doi.org/10.3390/antibiotics6040030

Sauer MH, Ye JJ, Klinke S, Nino E (1996) Identification of an anaerobically induced phosphoenolpyruvate-dependent fructose-specific phosphotransferase system and evidence for the Embden-Meyerhof glycolytic pathway in the heterofermentative bacterium Lactobacillus brevis. J Bacteriol 178(1):314–316. https://doi.org/10.1128/jb.178.1.314-316.1996

Salvetti E, Fondi M, Fani R, Torriani S, Felis GE (2013) Evolution of lactic acid bacteria in the order Lactobacillales as depicted by analysis of glycolysis and pentose phosphate pathways. Syst Appl Microbiol 36(5):291–305. https://doi.org/10.1016/j.syapm.2013.03.009

Salvetti E, Harris HMB, Felis GE, O’Toole PW (2018) Comparative genomics reveals robust phylogroups in the genus Lactobacillus as the basis for reclassification. Appl Environ Microbiol 84. https://doi.org/10.1128/AEM.00913-18

Schleif R (2010) AraC protein, regulation of the L-arabinose operon in Escherichia coli, and the light switch mechanism of AraC action. FEMS Microbiol Rev 34(5):779–796. https://doi.org/10.1111/j.1574-6976.2010.00226.x

Siezen RJ, Bajyanov JR, Felis GE, van der Sijde MR, Starrenburg M, Molenaar D, Wels M, van Hijum SAFT, van Hylckama Vlieg JET (2011) Genome-scale diversity and niche adaptation analysis of Lactococcus lactis by comparative genome hybridization using multi-strain arrays. Microb Biotechnol 4(3):383–402. https://doi.org/10.1111/j.1751-7915.2011.00247.x

Siezen RJ, Tzeneva VA, Castioni A, Wels M, Phan HTK, Rademaker JLW, Starrenburg MJC, Kleerebezem M, van Hylckama Vlieg JET (2010) Phenotypic and genomic diversity of Lactobacillus plantarum strains isolated from various environmental niches. Environ Microbiol 12(3):758–773. https://doi.org/10.1111/j.1462-2900.2009.02119.x

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19(6):1117–1123. https://doi.org/10.1101/gr.089322.108

Smokvina T, Wels M, Polka J, Cheruva C, Brisse S, Boekhorst J, van Hylckama Vlieg JET, Siezen RJ (2013) Lactobacillus paracasei comparative genomics: towards species pan-genome definition and exploitation of diversity. PLoS One 8(7):e68731. https://doi.org/10.1371/journal.pone.0068731

Sonowal R, Nandimuthu K, Kulshahi SS, Koushika SP, Nanjundiah V, Mahadevan S (2013) Hydrolysis of aromatic β-glucosides by non-pathogenic bacteria confers a chemical weapon against predators. Proc R Soc B 280:201307. https://doi.org/10.1098/rspb.2013.0721

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Sun Z, Harris HMB, McCann A, Guo C, Argümon S, Zhang W, Yang X, Jeffery IB, Cooney JC, Kagawa TF, Liu W, Song Y, Salvetti E, Wroobel A, Rasinkangas P, Parkhill J, Rea MC, O’Sullivan O, Riti J, Douillard FP, Paul Ross R, Yang R, Briner AE, Felis GE, De Vos WM, Barrangou R, Klaenhammer TR, Caufield PW, Cui Y, Zhang H, O’Toole PW (2015) Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. Nat Commun 6:8322. https://doi.org/10.1038/ncomms9322

Thierry A, Pogačić T, Weber M, Lortal S (2015) Production of flavor compounds by lactic acid bacteria in fermented foods. In: Mozzi F, Raya RR, Vignolo GM (eds) Biotechnology of lactic acid bacteria: novel applications, 2nd edn. Wiley-Blackwell, West Sussex, UK, pp 314–340

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B (2016) Various R programming tools for plotting data: gplots. R package version 3.0.1. https://CRAN.R-project.org/package=gplots

Whitley E, Ball J (2002) Statistics review 6: nonparametric methods. Crit Care 6:509–513. https://doi.org/10.1186/cc1820

Wu C, Huang J, Zhou R (2017) Genomics of lactic acid bacteria: current status and potential applications. Crit Rev Microbiol 43(4):393–404. https://doi.org/10.1080/1040841X.2016.1179623

Zhang W, Sun Z, Wu R, Meng H, Zhang H (2013) Comparative genome analysis of probiotic Lactobacillus casei Zhang. In: iconet Press (ed) Genomics II—bacteria, viruses and metabolic pathways. Brisbane, pp 276–296

Zhang W, Zhang H (2014) Genomics of lactic acid bacteria. In: Zhang H, Cui Y (eds) Lactic acid bacteria: fundamentals and practice. Springer Netherlands, Dordrecht, pp 205–247

Zhang ZZ, Ye ZZ, Yu L, Shi P (2011) Phylogenomic reconstruction of lactic acid bacteria: an update. BMC Evol Biol 11(1):1–12. https://doi.org/10.1186/1471-2148-11-1

Zhao S, Guo Y, Sheng Q, Shyr Y (2015) An improved heatmap package: heatmap3. R package version 3.0.1. https://CRAN.R-project.org/package=heatmap3

Zheng J, Ruan L, Sun M, Günzle M (2015) A genomic view of lactobacilli and pediococci demonstrates that phylogeny matches ecology and physiology. Appl Environ Microbiol 81(20):7233–7243. https://doi.org/10.1128/AEM.02116-15