



# Large-Scale Phylogenomics of the *Lactobacillus casei* Group Highlights Taxonomic Inconsistencies and Reveals Novel Clade-Associated Features

Sander Wuyts,<sup>a</sup> Stijn Wittouck,<sup>a</sup> Ilke De Boeck,<sup>a</sup> Camille N. Allonsius,<sup>a</sup> Edoardo Pasolli,<sup>b</sup>  Nicola Segata,<sup>b</sup>  Sarah Lebeer<sup>a</sup>

Research Group Environmental Ecology and Applied Microbiology, Department of Bioscience Engineering, University of Antwerp, Antwerp, Belgium<sup>a</sup>; Centre for Integrative Biology, University of Trento, Trento, Italy<sup>b</sup>

**ABSTRACT** Although the genotypic and phenotypic properties of the *Lactobacillus casei* group have been studied extensively, the taxonomic structure has been the subject of debate for a long time. Here, we performed a large-scale comparative analysis by using 183 publicly available genomes supplemented with a *Lactobacillus* strain isolated from the human upper respiratory tract. On the basis of this analysis, we identified inconsistencies in the taxonomy and reclassified all of the genomes according to their most closely related type strains. This led to the identification of a catalase-encoding gene in all 10 *L. casei sensu stricto* strains, making it the first described catalase-positive species in the *Lactobacillus* genus. Moreover, we found that 6 of 10 *L. casei* genomes contained a SecA2/SecY2 gene cluster with two putative glycosylated surface adhesin proteins. Altogether, our results highlight current inconsistencies in the taxonomy of the *L. casei* group and reveal new clade-associated functional features.

**IMPORTANCE** The closely related species of the *Lactobacillus casei* group are extensively studied because of their applications in food fermentations and as probiotics. Our results show that many strains in this group are incorrectly classified and that reclassifying them to their most closely related species type strain improves the functional predictive power of their taxonomy. In addition, our findings may spark increased interest in the *L. casei* species. We find that after reclassification, only 10 genomes remain classified as *L. casei*. These strains show some interesting properties. First, they all appear to be catalase positive. This suggests that they have increased oxidative stress resistance. Second, we isolated an *L. casei* strain from the human upper respiratory tract and discovered that it and multiple other *L. casei* strains harbor one or even two large, glycosylated putative surface adhesins. This might inspire further exploration of this species as a potential probiotic organism.

**KEYWORDS** *Lactobacillus casei* group, accessory Sec system, catalase, comparative genomics, phylogenomics

*Lactobacillus* is the largest genus of lactic acid bacteria, comprising >200 species (1). These species are naturally present on human and animal mucosal surfaces (e.g., the gastrointestinal and vaginal tracts) and in many food-related environments, including fruits, vegetables, wine, milk, and meat, where they can become dominant if able to ferment large doses of sugar with concomitant production of lactic acid and related metabolites. These bacteria are important model microorganisms for metabolic fermentation, cell wall biosynthesis, and microbe-host interaction studies. In addition, they are currently exploited in many biotechnical applications, e.g., as starter cultures,

Received 6 June 2017 Accepted 31 July 2017 Published 22 August 2017


**Citation** Wuyts S, Wittouck S, De Boeck I, Allonsius CN, Pasolli E, Segata N, Lebeer S. 2017. Large-scale phylogenomics of the *Lactobacillus casei* group highlights taxonomic inconsistencies and reveals novel clade-associated features. *mSystems* 2:e00061-17. <https://doi.org/10.1128/mSystems.00061-17>.

**Editor** Pieter C. Dorrestein, University of California, San Diego

**Copyright** © 2017 Wuyts et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Sarah Lebeer, [Sarah.lebeer@uantwerpen.be](mailto:Sarah.lebeer@uantwerpen.be).

S. Wuyts and S. Wittouck contributed equally to this work.

 Large-scale phylogenomics of the *Lactobacillus casei* group highlights taxonomic inconsistencies and reveals novel clade-associated features

as probiotics, in the production of bioplastics, and as vaccine carriers, highlighting their high commercial value (1).

The *Lactobacillus casei* group, comprising the species *L. casei*, *Lactobacillus paracasei*, and *Lactobacillus rhamnosus*, is among the most economically interesting clades within the genus *Lactobacillus*. Commercially, these microbes are used in fermented dairy products or food supplements targeting the gastrointestinal tract (2–4) and the vaginal tract (5). Interest is also increasing in their application in other product formulations targeting different human and animal body niches. For example, an underexplored niche for topical application of probiotics is the (upper) respiratory tract and its related diseases (6).

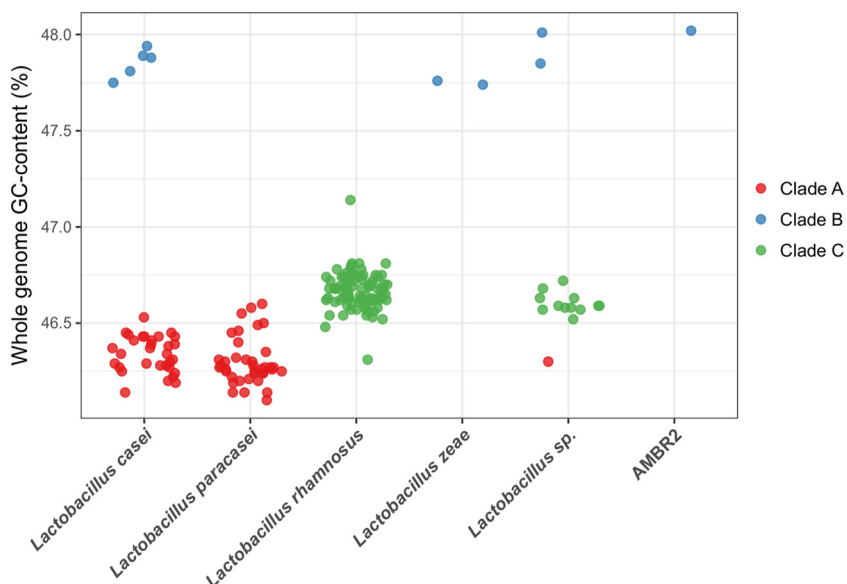
Despite broad interest in the *L. casei* group, both the nomenclature and the classification of this group are subjects of discussion. This is reflected, for example, in the introduction of the related species *Lactobacillus zae* in 1996 (7) and its subsequent rejection in 2008 (8). Furthermore, recent comparative genomic analyses showed that many strains currently classified as *L. casei* and *L. paracasei* are, in fact, members of the same species (9). In addition, many new isolates are classified as *L. casei*, while they are genetically more closely related to *L. paracasei* type strain ATCC 334 than to *L. casei* type strain ATCC 393 because of high heterogeneity in their 16S rRNA genes (10). Thus, many novel identifications are not in line with the current taxonomic classification (4, 8, 11). Multiple efforts have been made to facilitate the differentiation between *L. casei* group members on the basis of the use of PCR and/or DNA fingerprinting techniques (4, 10–13). However, with the price reduction of whole-genome sequencing and the rising availability of public genomes (210 *L. casei* group members as of February 2017), a more in-depth insight into the genetic differences and taxonomy of *L. casei* group members can be obtained by using computational comparative genomics.

In this study, we isolated a lactic acid bacterium from the respiratory tract, a rather unexpected niche for such a bacterium because it is not anaerobic or nutrient/sugar rich. On the basis of its genome sequence and a comparative genomic analysis using 183 publicly available *L. casei* group genome assemblies, we classified this strain as a member of the species *L. casei*, which—unexpectedly—turned out to be the smallest clade within the *L. casei* group. To our knowledge, this isolate, which we named *L. casei* AMBR2, is the first *L. casei* strain isolated from the upper respiratory tract. Using different comparative genomic approaches, we provide more insight into the genetic relationship of strains belonging to this group and use this information to further explore the functional potential of *L. casei* AMBR2, as well as the other *L. casei* group members.

## RESULTS

In this study, we considered the 183 publicly available genome assemblies for the *L. casei* group that passed strict quality control (N75 values of >10,000 bp and <500 undetermined bases per 100,000 bases), as reported in Table S1 in the supplemental material. Of these genomes, 92 were originally annotated as *L. rhamnosus*, 36 were *L. casei*, 38 were *L. paracasei*, and 2 were *L. zae*. In addition to the public genomes classified as belonging to the *L. casei* group, we screened all of the unclassified *Lactobacillus* genomes (categorized as *Lactobacillus* sp. in the NCBI database) for *L. casei* group members by comparing their 16S rRNA gene sequences to a filtered version of the RDP database (v11) (14). This resulted in an additional 15 genomes. Furthermore, one newly sequenced *L. casei* strain (AMBR2), which we isolated from a human upper respiratory tract, was added to the analysis. This resulted in a total of 184 *L. casei* group strains studied, thereby significantly improving the number of genomic assemblies studied since the latest comparative genomics study, where only 10 (2) or 34 (9) genomes were used.

**Identification of three clades within the *L. casei* group on the basis of GC content, phylogeny, and pairwise comparisons.** The GC contents of the species *L. paracasei*, *L. rhamnosus*, and *L. zae* show low intraspecies variation, with respective average values of 46.3, 46.7, and 47.8% (Fig. 1), whereas larger interspecies diversity is

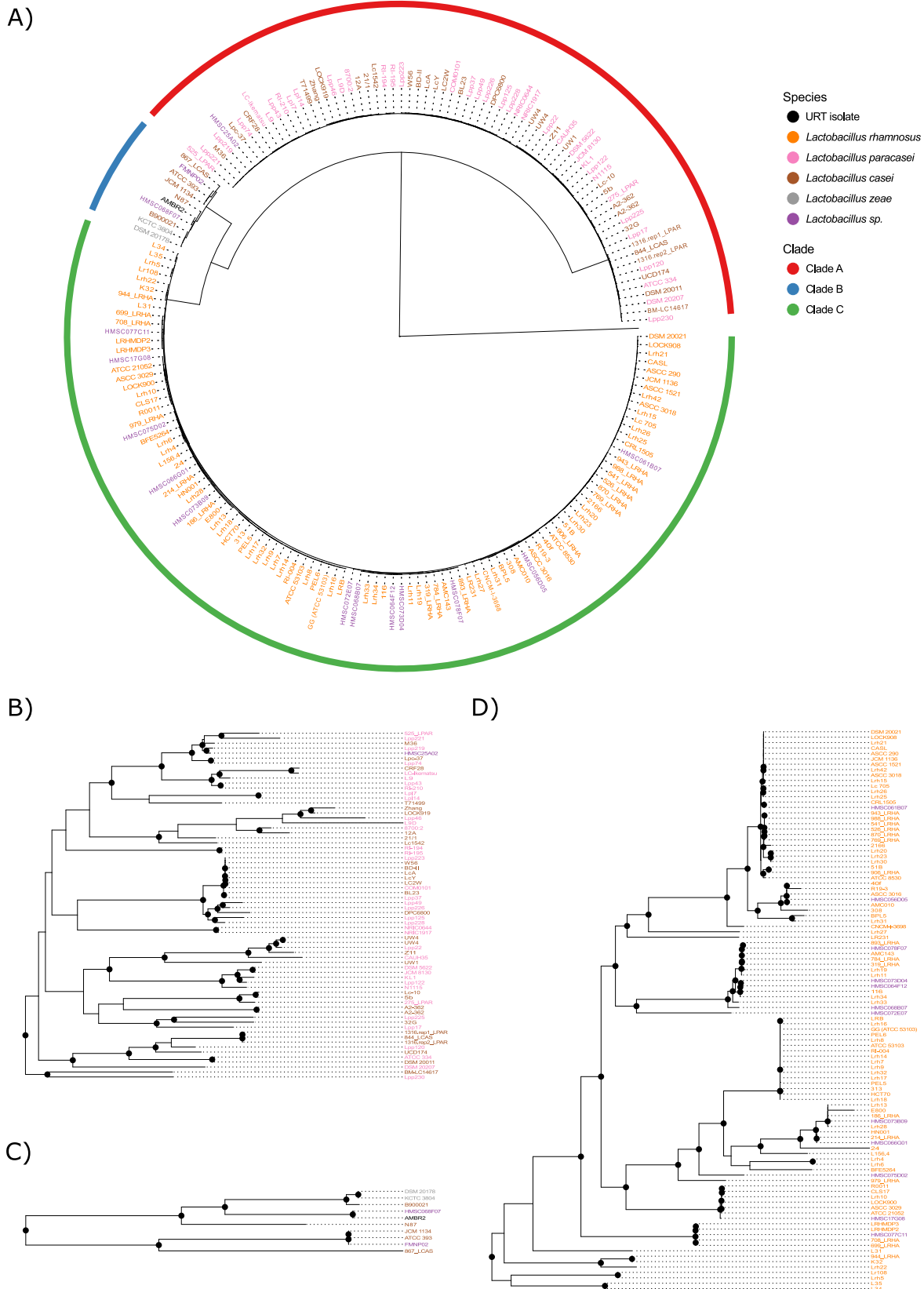


**FIG 1** GC contents of all of the genomes analyzed in this study. Genomes are grouped according to their species annotation in the NCBI database (except for upper respiratory tract isolate *L. casei* AMBR2) and colored by the phylogenetic clade they belong to, as defined in Fig. 2.

observed. In contrast, *L. casei* genomes can be divided into two groups, a large group showing a GC content in the range of that of *L. paracasei* (46.10 to 46.60%) and a small group of five genomes showing a much greater GC content, similar to that of the *L. zeae* genomes (47.74 to 47.76%). As for the unclassified assemblies (categorized as *Lactobacillus* sp.), two genomes show a GC content as great as that of the *L. zeae* genomes, while the rest of them are within the *L. rhamnosus*-*L. paracasei* range. The genome sequence of our new isolate, which we designated *L. casei* AMBR2, shows a GC content similar to that of the *L. zeae* genomes.

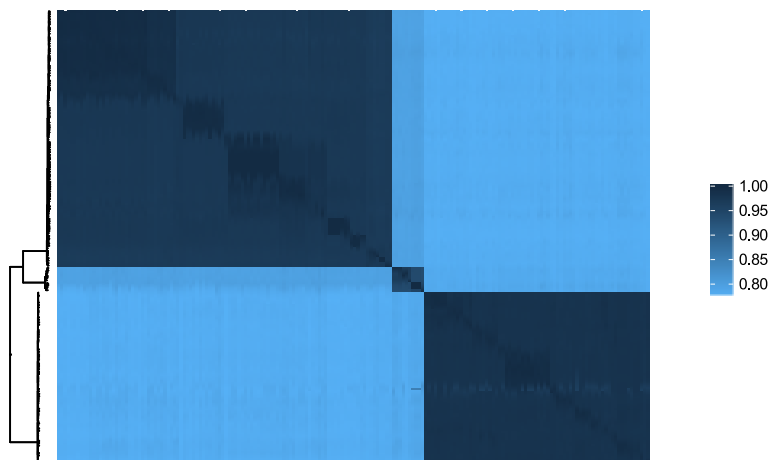
To study the genetic relatedness of the genomic assemblies, we constructed a high-quality maximum-likelihood phylogenetic tree of the *L. casei* group by using 776 conserved single-copy marker genes. These marker genes were identified with roary (15) and showed at least 70% blastp identity and were present in at least 96% of the genomes studied. In addition, the genome of *Lactobacillus nasuensis* JCM 17158 (GCA\_001434705) was added to the alignment to serve as an outgroup. This strain was chosen because it had the best-quality assembly of three strains that are closely related to the *L. casei* group (1). The resulting tree is shown in Fig. 2A. The tree structure reveals three separate clades within the *L. casei* group, with very small branch lengths within each clade in comparison to the branch lengths between the clades. Clade A contains the majority of the *L. casei* genomes and all of the *L. paracasei* genomes, as well as one unclassified *Lactobacillus* sp. assembly. The smallest clade (B), contains the two *L. zeae* genomes, five *L. casei* genomes (including *L. casei* type strain ATCC 393), two unclassified lactobacilli, and our own upper respiratory tract isolate (*L. casei* AMBR2). Interestingly, these members of clade B are also those with an elevated GC content, as shown in Fig. 1. Finally, clade C consists of all *L. rhamnosus* genomes, as well as 12 *Lactobacillus* sp. genomes retrieved from the NCBI database as described above.

The phylogeny of the clade A subtree (Fig. 2B) shows that the genomes of *L. casei* and *L. paracasei* are completely intermixed in the tree. Together with the occurrence of *L. casei* genomes in clade B, this shows that these two species constitute paraphyletic taxa when the current species annotations are used. The two *L. zeae* genomes do cluster together in the subtree of clade B (Fig. 2C), but it should be noted that these genomes represent the same strain (DSM 20178 = KCTC 3804) independently sequenced. In contrast, *L. rhamnosus* does seem to be a monophyletic taxon according to this tree (Fig. 2D). Of note is the fact that multiple genomic assemblies seem to be

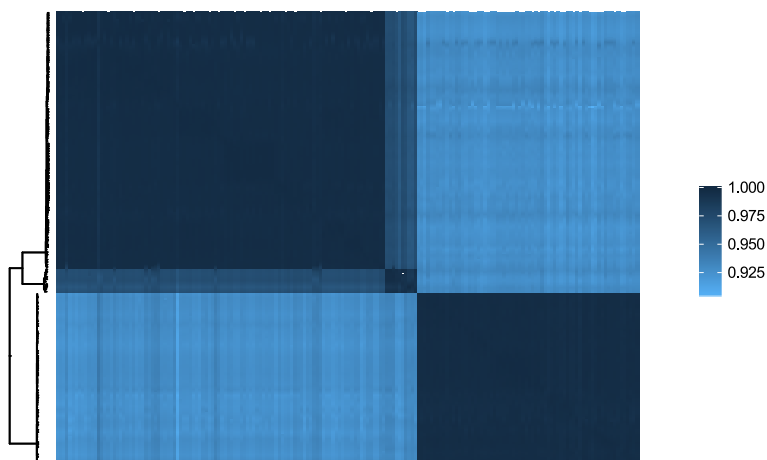


**FIG 2** Phylogenetic trees of the whole *L. casei* group and the individual clades. (A) Phylogenetic tree constructed of all 184 genome assemblies of the *L. casei* group by using 776 single-copy marker genes identified by roary (15) with *L. nasuensis* JCM 17158 as the outgroup. Colors show NCBI database classifications. URT, upper respiratory tract. (B) Subtree of clade A. (C) Subtree of clade B. (D) Subtree of clade C. Strong clades (bootstrap support of >70) are indicated by black dots.

ANiB



TETRA



**FIG 3** Pairwise ANiB and TETRA values for all genomes. The phylogenetic tree on the left is the same as that in Fig. 2A with the outgroup removed.

identical on the basis of the sequences of all 776 single-copy marker genes. For example, there appear to be no fewer than 17 assemblies that show very high similarity to *L. rhamnosus* GG, a well-known probiotic strain (16), possibly indicating that the same strain was sequenced on multiple occasions.

In the current era of whole-genome sequencing, pairwise genome comparison metrics such as average nucleotide identity (ANI) and tetranucleotide frequency (TETRA) are often used as an operational method to detect species boundaries. Figure 3 shows the ANiB (ANI calculated by using a blast implementation) and TETRA distances between all *L. casei* group genomes. Both distance metrics support the grouping of the genomes in the three clades defined by the phylogenetic tree in Fig. 2. For species delimitation, Richter et al. (17) advise the use of an ANI cutoff of 95 to 96%, although their tests show that values of 94 to 95% within a species are not uncommon. When comparing genomes belonging to different clades (e.g., a clade A genome and a clade B genome), we observed ANI values of  $\leq 85.1\%$ , indicating that each clade consists of one or more species that are distinct from the other clades. When comparing genomes that belong to the same clade (e.g., two clade A genomes), we observed ANI values of  $\geq 96.1\%$  in clade A,  $\geq 93.6\%$  in clade B, and  $\geq 96.3\%$  in clade C. These results suggest

**TABLE 1** Overview of the gene content distribution in the *L. casei* group<sup>a</sup>

| Group                 | No. of genomes | Avg no. of genes/genome ± SD | Avg no. of orthogroups/genome ± SD | No. of core orthogroups | No. of accessory orthogroups |
|-----------------------|----------------|------------------------------|------------------------------------|-------------------------|------------------------------|
| <i>L. casei</i> group | 184            | 2,827 ± 141                  | 2,654 ± 92                         | 1,814                   | 4,101                        |
| Clade A               | 70             | 2,897 ± 148                  | 2,696 ± 106                        | 1,924                   | 2,866                        |
| Clade B               | 10             | 2,847 ± 111                  | 2,615 ± 96                         | 1,924                   | 1,576                        |
| Clade C               | 104            | 2,780 ± 119                  | 2,629 ± 68                         | 2,133                   | 2,363                        |

<sup>a</sup>Gene content metrics were calculated for the *L. casei* group as a whole, as well as for the three clades defined by the phylogenetic tree. A core orthogroup is defined as an orthogroup present in >95% of the genomes.

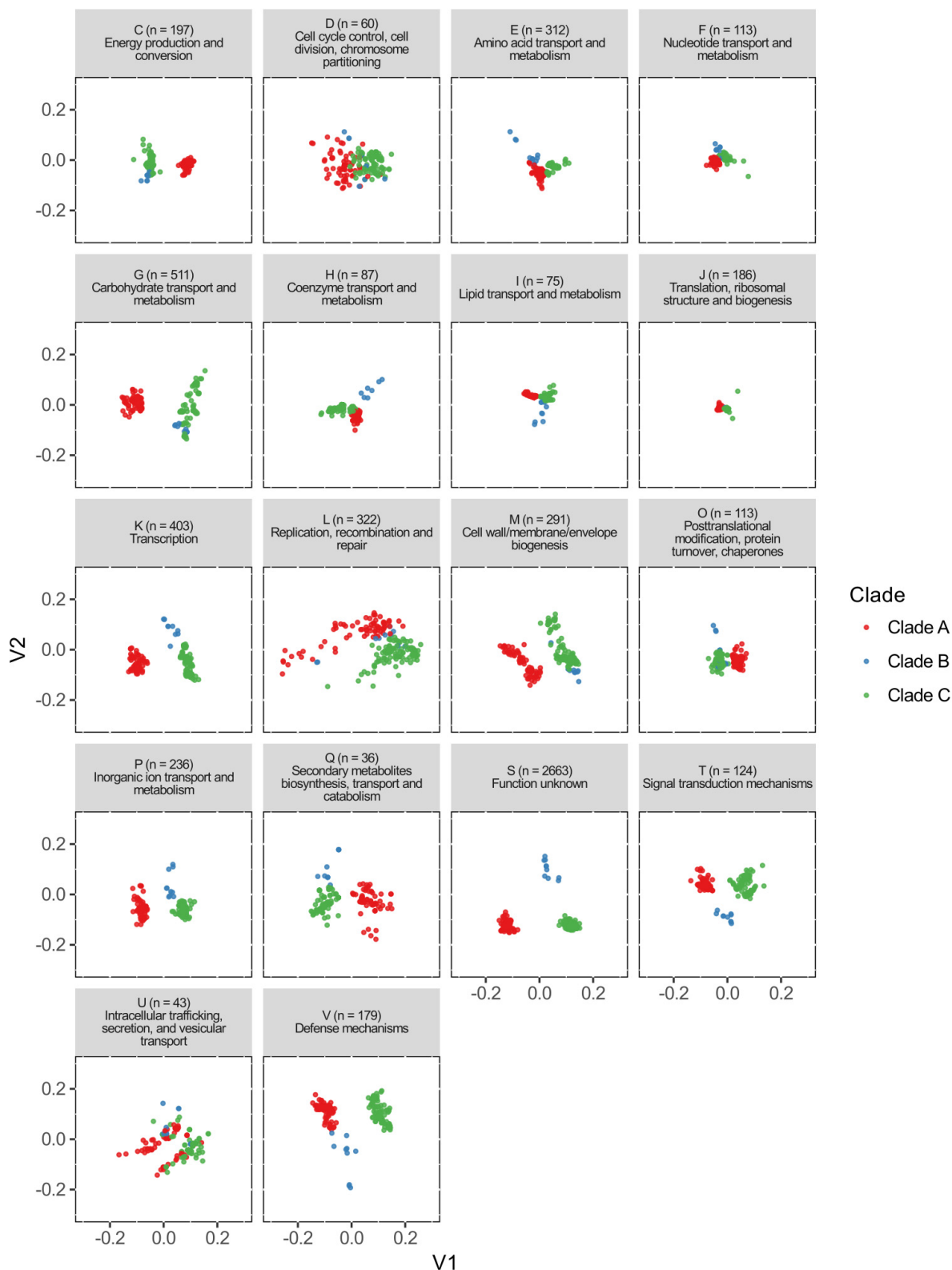
that clades A and C both consist of a single bacterial species. The minimal ANI value within clade B, however, is just below the species threshold of 94 to 95%, indicating that clade B might be considered one bacterial species with two subspecies or even two separate species. This conclusion is also supported by very high TETRA values within each clade;  $\geq 0.9941$  in clade A,  $\geq 0.9872$  in clade B, and  $\geq 0.9882$  in clade C.

**Gene content and predicted functional capacity support separation of the *L. casei* group in three clades.** In total, 521,567 genes were present in the genomes, with an average number of 2,828 ( $\pm 141$ ) genes per genome. These genes were clustered into 5,915 orthogroups by OrthoFinder (18), where an orthogroup is defined as the group of genes descended from a single gene in the most recent common ancestor of a group of species (18). Of these orthogroups, 1,814 were identified as core orthogroups and 4,101 were identified as accessory orthogroups (Table 1). When comparing clades A, B, and C identified above, there seems to be no large difference in the average number of genes per genome. Regarding the core orthogroups, clade C seems to have the largest number (2,133), while clades A and B were found to have exactly the same number of core orthogroups (1,924). Finally, clade B showed the lowest number of accessory orthogroups, probably because of the lower number of available genomic assemblies.

To compare the overall functional capacity of the strains, we visualized the differences in gene content of the genomes for different functional categories of genes. This was done by mapping the orthogroups to the eggNOG database (19). In total, between 2,332 and 2,902 genes per genome had a match within 18 functional categories (Fig. 4). However, it should be noted that a great fraction of orthogroups ( $n = 2,662$ ) was categorized as “function unknown” (S), showing the necessity for further experimental research in functional gene validation.

Figure 4 shows that all of the genomes studied have a rather overlapping orthogroup composition in categories D and J, respectively, representing cell cycle control, cell division, chromosome partitioning and translation, ribosomal structure, and biogenesis functions, two very fundamental categories in which no significant difference between closely related species is expected. Members of clades B and C cluster separately from members of clade A for the functions energy production and conservation, carbohydrate metabolism, cell envelope biogenesis, and posttranslational modification and chaperons (categories C, G, M and O). In contrast to these are categories E and V, representing amino acid transport and metabolism and defense mechanisms, where the opposite grouping was observed since clade B showed greater similarity to clade A. Interestingly, in at least 6 of 18 categories (H, I, K, P, Q, and T), a clear separation between the members of the phylogenetically different clades could be observed, indicating that these clades possess different functional capacities and highlighting the great potential for further investigation in future work. As a side note, no systematic differences between the clades in the total gene counts of the functional groups were observed (see Fig. S1), so we could exclude this as a possible explanation for why we see cladewise grouping in our principal-coordinate analysis (PCoA) plots.

**Genetic potential for oxidative stress resistance is a clear discriminative feature of the clades within the *L. casei* group.** The three phylogenetic clades differ in functional potential on the basis of the generic orthogroup analysis presented. Therefore, we further explored their functional potential by means of a more in-depth



**FIG 4** PCoA of predicted functional capacity per clade based on mapping of all orthogroups to the eggNOG database (v4.5) (19). Each letter represents a different functional category, as defined above each plot. Some orthogroups mapped to multiple functional categories. The majority of the orthogroups (2,662 of them) mapped to category S (function unknown).

interest-driven study of two industrially relevant properties of lactic acid bacteria, oxidative stress resistance and glycosylation potential.

Being able to cope with the presence of reactive oxygen species is an important feature for the industrial use, as well as the niche adaptation, of lactobacilli. Since *L. casei* AMBR2, which was isolated from the upper respiratory tract of a healthy person, a typical oxygen-rich niche, and documented oxidative-stress-resistant strain *L. casei* N87 (20) both cluster within smaller clade B of the *L. casei* group, we evaluated different oxidative stress resistance mechanisms of clade B in relation to all *L. casei* group members.

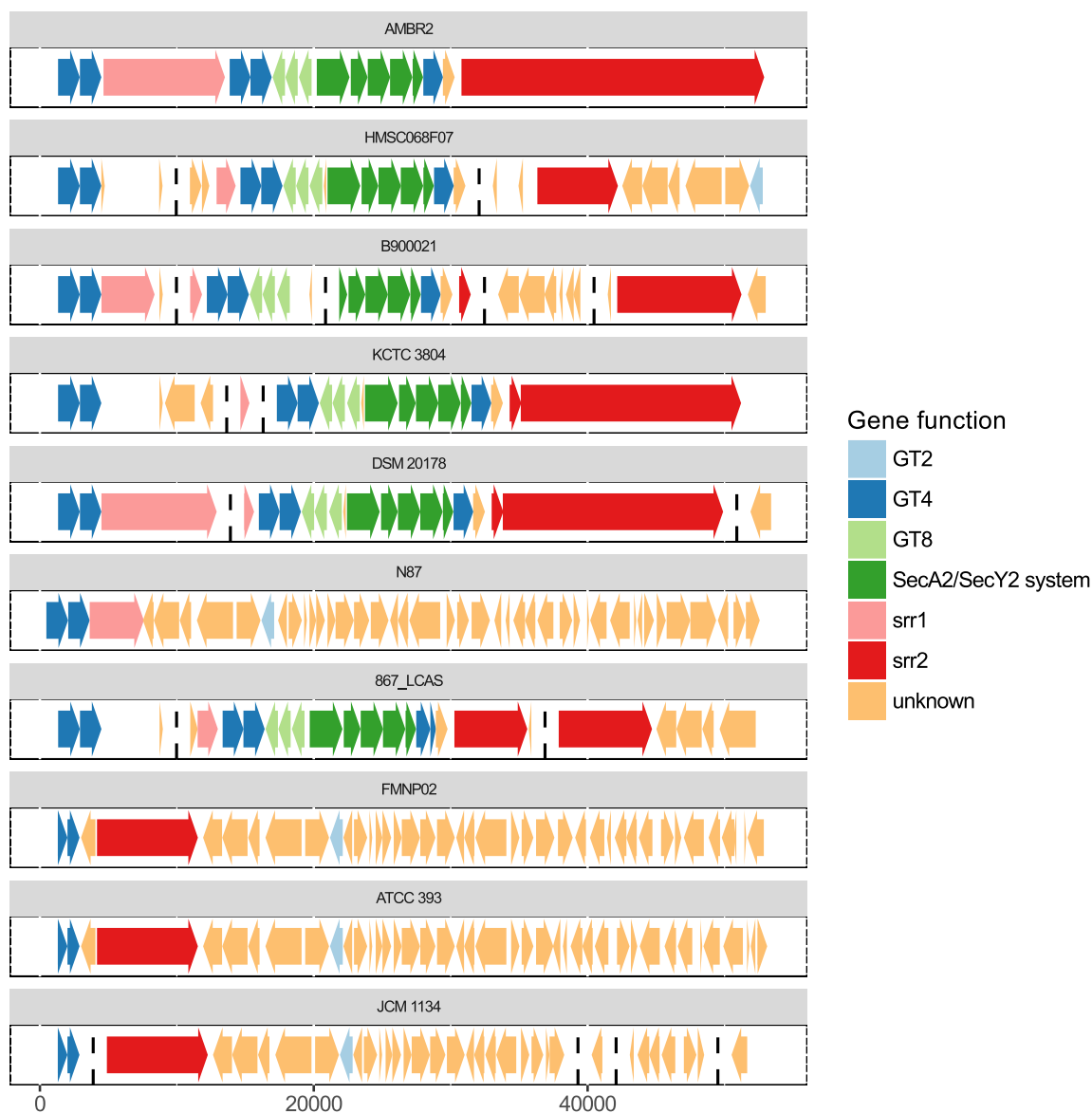
Catalase, which catalyzes the decomposition of  $\text{H}_2\text{O}_2$  to  $\text{H}_2\text{O}$  and  $\text{O}_2$ , plays an important role in protecting cells against oxidative stress. While the *Lactobacillus* genus is defined as catalase negative, recent studies have shown catalase activity in several strains, including respiration-competent strain *L. casei* N87 (21), but without linking the presence of these catalase genes with the underlying phylogeny. Therefore, the presence of catalase genes was evaluated here in relation to the newly described phylogenetic structure of the *L. casei* group. Interestingly, open reading frames annotated as catalase genes were identified only in strains belonging to clade B and in a single genome of clade C (*L. rhamnosus* CRL1505; [GCA\\_000414365](https://doi.org/10.1093/g3/kjz004)). Mapping to Pfam families PF00199 and PF05067 with HMMER (22) resulted in the identification of two different catalase types, one annotated as a heme-dependent catalase (487 amino acids) and the other as a manganese-dependent catalase (269 amino acids). The heme-dependent catalase gene was found in all 10 genome assemblies of clade B, while the manganese-dependent catalase gene was present in only 7 of 10 clade B genomes. We could experimentally confirm that only the strains of clade B tested—particularly strain AMBR2—were positive for catalase activity by a standard microbiology lab test (Table S2).

The antioxidant superoxide dismutase (SOD), which is encoded by another gene important for oxidative stress resistance and scavenges  $\text{O}_2^-$  into  $\text{O}_2$  and  $\text{H}_2\text{O}_2$ , was long believed to be absent from the *Lactobacillus* genus. However, genome analysis recently revealed the presence of SOD-encoding genes in some *L. casei* and *L. paracasei* strains (23), suggesting that two of the three *L. casei* group species could harbor SOD-encoding genes. To confirm this, we screened the whole *L. casei* group for SOD-encoding genes. Remarkably, SOD-encoding genes were present only in our newly assigned clade A strains (69 of 70 genomes) and thus in only one species instead of two. Interestingly, mapping to four different SOD Pfam families (PF00080, PF00081, PF02777, PF09055) with HMMER led to two different hits, one expected hit with PF00081 representing the gene for iron-manganese SOD (found in 69/70 clade A genomes) and one rather unexpected hit with the gene for copper SOD (found in 4/70 clade A genomes), which is the SOD most commonly used by eukaryotes. In general, these results show that because of the new phylogenetic structure of the *L. casei* group presented in this paper, the presence of SOD-encoding genes seems to be a unique property of clade A.

**Six clade B genomes encode a SecA2/SecY2 secretion system with two putative glycosylated surface adhesins as substrates.** Strain-specific glycosylation of surface molecules greatly affects interactions between *Lactobacillus* bacteria and their host cells. In addition, it confers interesting industrial properties, such as improved rheological and stress resistance properties (24, 25). Therefore, we subsequently scanned the genomes for clade-specific patterns in gene families encoding glycosyltransferases (GTs). We then designed a visualization approach that allowed us to better align these GTs and their positions in a selection of genomes (all genomes of clade B and closed genomes of clades A and C). These results are shown in Fig. S2 to S4, respectively.

We found five different gene clusters that were enriched in GTs, where a gene cluster was considered “enriched” if it contained three or more GTs with  $\leq 5$  kb between two successive GTs in at least one genome. Two of these five clusters contained a homolog of a known priming GT and thus are probably responsible for the biosynthesis of heteropolymetric exopolysaccharides (EPS) or capsular polysaccharides





**FIG 5** Gene content and order of the clade B-specific GT-rich gene cluster. The gene cluster is shown in all 10 clade B strains. Contigs were mapped to AMBR2; contig boundaries are indicated by broken vertical lines. Functional annotation was performed on the orthogroup level; multiple orthogroups can have the same function assigned to them. For example, the SecA2/SecY2 system consists of five orthogroups, SecA2, SecY2, Asp1, Asp2, and Asp3.

(CPS). These two clusters were present in all of the genomes but varied strongly between strains in the quantity and type of associated GTs found; the minimum was one GT (always the priming GT homolog), and the maximum was five GTs in one cluster. The third cluster contained one GT2 and one GT83 (on the basis of CAZY family numbers) in its minimal form (clade B), while more copies of these two GT types were found in clades A and C. A fourth cluster was detected in some clade C genomes (one to three GTs) and in all clade B genomes (four GTs). The fifth cluster could be found in 6 of 10 clade B genomes and was absent from clades A and C. It was located directly upstream of the fourth cluster.

Because this fifth GT-rich gene cluster appeared to be a unique characteristic of clade B and was present in our isolate, *L. casei* AMBR2, we explored this cluster in more depth (Fig. 5). Of the six strains where the cluster was found, AMBR2 was the only one with an unfragmented genome sequence in this region. Therefore, we used this strain as a reference to describe the cluster. We defined the beginning and end of the cluster

by considering the stretch of genes whose orthogroups were absent from clades A and C or at least less abundant than the orthogroups of the surrounding genes (see Table S3). In AMBR2, the cluster consisted of the following genes. The first is a long gene (8,868 bp) enriched in serine residues (15%), which we annotated as *srr1* (serine-rich repeat [Srr] gene 1). Directly downstream from that gene, we found two tandem GTs of the GT4 family, followed by three tandem GTs of the GT8 family located on the opposite strand. Furthermore, a SecA2/SecY2 secretion system, another GT4 gene, and a gene with an unknown function were found. The last is a second very long gene (22,113 bp; *srr2*) even more strongly enriched in serine residues than *srr1* (38%). The complete gene cluster, including the order of the genes, appeared to be fully conserved among the six strains that contained it. The cluster was absent from strain *L. casei* N87, except for the first part of *srr1*. In *Lactobacillus* sp. strain FMNP02, *L. casei* ATCC 393 (the *L. casei* type strain), and *L. casei* JCM 1134, only the last part of *srr2* was present.

The clade B-specific gene cluster described above contains two large genes that encode serine-rich proteins. These Srr proteins are often described as substrates for the accessory Sec (or SecA2/SecY2) system, which is responsible for the secretion of a single substrate (26). Therefore, we further investigated whether *srr1* and *srr2* could encode SecA2/SecY2 substrates. Unfortunately, we did not have the complete sequence of *srr1*. In AMBR2, the protein lacked its first 57 residues because of a dinucleotide deletion that caused a frameshift. In the other strains where the gene was present, it was truncated (N87) or its sequence was interrupted by a contig boundary (other strains). However, we could reconstruct the probable full sequence of the gene by looking at the alignment of the partial sequences (Text S1). From this reconstruction, it was clear that the protein contained both the KxYKxGKxW and the LPxTG motifs that are characteristic of a SecA2/SecY2 substrate (26), marking it as a highly probable accessory Sec system substrate and thus as a potential glycosylated surface adhesin.

The *srr2* gene was present in its full form in AMBR2 (Text S2). In at least two strains, the first ~280 amino acids of the protein, including the N-terminal signal sequence, were split off as an extra gene, also because of a frameshift mutation (a stretch of five A bases gained an additional A). In the other three strains, the sequence was split up because of contig boundaries. Interestingly, this protein also contained the KxYKxGKxW and LPxTG motifs, although the latter was a rare variant (LPQTS). This would mean that *srr2* codes for a second, different glycosylated surface adhesin secreted by the same accessory Sec system.

In an attempt to identify the ligands of these two putative glycosylated surface adhesins, we performed a sequence homology search of the UniProt database. The best-scoring hit for Srr protein 1 was an uncharacterized protein from *Mycobacterium kansasii* that showed only 47% identity and 39% query coverage. The best-scoring hit for Srr2 was a “cell surface anchor protein” from *Streptococcus pneumoniae*, which also showed low identity and query coverage values (54 and 64%, respectively). Thus, we conclude that the two putative adhesins are as-yet-uncharacterized proteins that require further functional validation.

## DISCUSSION

In this study, the genome of *L. casei* AMBR2, an isolate from the upper respiratory tract, was sequenced and compared to all currently available high-quality genomic assemblies of the *L. casei* group, which comprises the closely related species *L. casei*, *L. paracasei*, *L. rhamnosus*, and *L. zeae*.

Here we show that the *L. casei* group consists of three different taxonomic clades on the basis of (i) differences in whole-genome GC content (Fig. 1), (ii) a phylogenetic tree constructed on the alignment of 776 single-copy marker genes (Fig. 2A), and (iii) pairwise ANI and TETRA values (Fig. 3). The branches of the phylogenetic tree that separate the three clades are well supported, but interestingly, a nonnegligible number of branches within the clades show low bootstrap support, indicating extensive horizontal gene transfer within the clades. We found that clade C represents the species *L. rhamnosus*, as it is uniquely made up of *L. rhamnosus* isolates, including the type

strain *L. rhamnosus* DSM 20021. In contrast, both *L. casei* and *L. paracasei* isolates are found in clade A, which contains *L. paracasei* type strain ATCC 334. Because of the presence of this type strain, we suggest that all of the strains within clade A should be reclassified as *L. paracasei*, which is in line with the findings of Smokvina et al. (9). The third group, clade B, is much smaller than the other two clades (10 genomic assemblies) and consists of *L. casei*, *L. zae*, and upper respiratory tract isolate *L. casei* AMBR2. Following the reclassification of *L. zae* as *L. casei* (27), this group contains only *L. casei* strains, including *L. casei* type strain ATCC 393. Therefore, we propose that clade B represents the species *L. casei*. According to these results, only 5 of 36 genomes annotated as *L. casei* in the NCBI database are, in fact, genuine members of the species *L. casei*. The rest should be classified as *L. paracasei* instead, making *L. casei* the least sequenced species within the *L. casei* group. Of note is also the observation that *L. casei* is more closely related to *L. rhamnosus* than to *L. paracasei*, rendering the naming slightly confusing.

To our knowledge, this study is the first to perform an in-depth comparative genomics analysis of the *L. casei* group as a whole. We found the core genome of the whole group to be around 1,814 orthogroups, while the core genomes of reclassified clades A (*L. casei*), B (*L. paracasei*), and C (*L. rhamnosus*) contained around 1,924, 1,924, and 2,133 orthogroups, respectively (Table 1). The number of core orthogroups we found in the *L. casei* group is similar to that found by Smokvina et al. (9), who described a core genome of around 1,800 orthologous genes for 34 *L. casei* and *L. paracasei* genomes together. In clade C, we identified a slightly lower number of orthogroups than the 2,419 core genes described by Douillard et al. (28). Mapping of these orthogroups to the eggNOG database (v4.5) (19) revealed a difference in the predicted functional capacity of all three clades in at least 6 of 18 categories (Fig. 4). These results suggest that each phylogenetic clade could show unique functional properties previously overlooked because of taxonomic misclassifications. As an example, we worked out a comparative genomic analysis of two such functional properties, oxidative stress resistance and surface glycosylation potential. We show that both properties exhibit clade-specific gene distributions.

Of the three clades, B shows the largest intraclade variation. The lowest ANI value between two genomes in this clade was 93.6%, which is slightly lower than the cutoff normally used for bacterial species delimitation (17). In addition, clade B shows larger variation than the other clades in terms of their orthogroup content (see Fig. 4). Therefore, future work (including the isolation and sequencing of more clade B genomes) will indicate whether this clade should be split into more separate species.

In the past, several studies have successfully induced a more oxidative-stress-resistant phenotype in members of the *L. casei* group by means of heterologous expression of SOD and/or catalase (29–32), emphasizing the industrial importance of this phenotype. In this study, we evaluated the presence of oxidative-stress-related genes in the whole *L. casei* group. Surprisingly, we identified a SOD-encoding gene in all of the genomic assemblies of clade A. Apart from Chaillou et al. (33) and Liu et al. (34), who found the presence of a SOD-encoding gene in *Lactobacillus sakei* 23K and *L. casei* Lc18, respectively, to our knowledge, no other SOD-encoding gene has been described in the *Lactobacillus* genus. The translated SOD-encoding gene identified shows high similarity (99% identity, blastp) to the translated SOD-encoding gene from *L. casei* Lc18, as described by Liu et al. (34). This gene has been used for expression in *Escherichia coli*, which resulted in a small increase in scavenging activity, but the results were not convincing. We believe that further work in characterizing the activity of this SOD-encoding gene is necessary. Nevertheless, the presence of this gene is an interesting unique genomic property of clade A, representing the species *L. paracasei*, that was previously overlooked, possibly because of misclassification of the *L. casei* group members.

Catalase is another important driver in establishing an oxidative-stress-resistant phenotype. Although lactic acid bacteria are generally defined as catalase negative, catalase activity has been found in members of the genera *Lactobacillus*, *Pediococcus*,

and *Leuconostoc* (35). For the genus *Lactobacillus* in particular, catalase genes and activity were reported and studied in *L. sakei* (29, 36), *Lactobacillus plantarum* (37–39), *L. casei* N87, and *L. zeae* (40). Here we show that all of the members of clade B in the *L. casei* group carry a gene for a heme-dependent catalase, making it the first described *Lactobacillus* species that is catalase positive. In addition, we identified a gene encoding a manganese-dependent catalase in 7 of 10 clade B members. In other words, the majority of the assemblies in this clade carry two different catalase-encoding genes. One member of this clade is *L. casei* N87, an oxygen-tolerant and respiratorily competent strain that has been extensively studied in the last few years (20, 21, 23, 41). These studies show that this strain possesses remarkably high oxidative resistance without the need for DNA cloning and thus support the fact that the predicted genes do function. This property makes the members of clade B, especially those carrying two catalase genes, of outstanding interest for industrial applications. Of interest is also the fact that strain AMBR2, which we isolated from the human nasopharyngeal niche, clusters within clade B and thus possibly harbors a very oxidant-resistant phenotype, as suggested here by the experimental observation of its catalase activity (Table S2).

The SecA2/SecY2 secretion system (or accessory Sec system) occurs in some Gram-positive bacteria in addition to the general SecA secretion system, which is universally present in bacteria. In each strain where it has been described, the accessory Sec system appears to be responsible for the secretion of a single substrate (26). This substrate differs between strains, but it is always a glycosylated surface adhesin belonging to a family of proteins (called Srr proteins) that can bind to a variety of ligands. The accessory Sec system and its substrates have been investigated primarily in Gram-positive pathogens such as *Streptococcus pneumoniae* and *Staphylococcus aureus* (26). In these species, virulence could be abolished by inactivating the secretion of the Srr protein. The accessory Sec system has also been found in some commensal Gram-positive bacteria, where it proved to be vital for successful host colonization. For example, Frese et al. (42) identified the SecA2/SecY2 system as the primary factor responsible for host-specific biofilm formation in *Lactobacillus reuteri*. Inactivation of the putative surface adhesin secreted by the system fully prevented host-specific biofilm formation. Since strain AMBR2 was isolated from the human upper respiratory tract, we hypothesize that one or both of its putative glycosylated surface adhesins are responsible for binding to a host ligand molecule in this niche, but this requires further experimental validation.

To our knowledge, this is the first time that two putative substrates for the same accessory Sec system have been identified within the same strain. However, even though both Srr proteins were present in some form in all six strains containing the accessory Sec system, we never observed them both in their fully intact form (i.e., including the N-terminal signal motif) within the same genome. In three strains, they might both be fully intact, but this was impossible to assess because of the quality of the assemblies. On three occasions, a frameshift mutation seemed to have occurred in one of the *srr* genes that cut off the N-terminal signal motif from the rest of the gene; once in *srr1* of AMBR2 and once in *srr2* of KCTC 3804 and DSM 20178 (or in their common ancestor). This observation might not be coincidental. It could be that the two surface adhesins are in competition with each other for secretion by the SecA2/SecY2 pathway. In the presence of a surface-attached ligand for only one of the adhesins, a short-term advantage might be gained by deactivating the other adhesin to adhere faster and better to the surface. This might constitute an interesting example of phase variation, i.e., the process of switching the expression of a gene on and off every couple of generations, leading to phenotype heterogeneity within a clonal bacterial population (43).

**Conclusions.** In this study, we sequenced a novel upper respiratory tract isolate, *L. casei* AMBR2, and studied its genome in relation to all of the currently available genomic assemblies of the members of the *L. casei* group. We found that the *L. casei* group harbors three different taxonomic clades by using a core genome phylogenetic

tree, GC content analysis, and pairwise genome distances (ANIb and TETRA). On the basis of the presence of a different type strain in each of these clades, we propose that clade A represents the species *L. paracasei*, clade B represents the species *L. casei*, and clade C represents the species *L. rhamnosus*. Our study clearly shows that many *L. casei* strains are wrongly annotated in the NCBI database and should be reclassified as *L. paracasei*.

Reclassification of the *L. casei* group members led to the discovery of at least one catalase gene in all of the members of clade B, representing *L. casei*, making it the first described catalase-positive species in the whole *Lactobacillus* genus. In addition, we found that all *L. casei* group strains contain two putative EPS/CPS clusters and that six strains of clade B, among which is our isolate AMBR2, contain an accessory secretion system with two putative glycosylated surface adhesins as secreted substrates.

Finally, we propose the use of whole-genome ANI with respect to *L. casei* group type strains as an easy, computationally inexpensive metric to differentiate between the species *L. casei* and *L. paracasei*, on the condition that the genome has been sequenced. Alternatively, if sequencing of the 16S rRNA gene leads to the identification of a member of the species *L. casei* or *L. paracasei*, then we propose the use of the heme-dependent catalase gene or the SOD-encoding gene as a marker gene for the correct identification of these species.

## MATERIALS AND METHODS

**Sequencing and downloading of publicly available assemblies.** Whole-genome sequencing of *L. casei* AMBR2 was performed with the Nextera XT DNA Sample Preparation kit (Illumina, San Diego, CA), and sequencing with the Illumina MiSeq platform with  $2 \times 300$  cycles at the Center of Medical Genetics Antwerp (University of Antwerp). Assembly was performed with SPAdes 3.8.0 (44).

All genomic assemblies classified as *L. casei*, *L. paracasei*, *L. rhamnosus*, and *L. zeae* (210 in total) were downloaded from the NCBI database on 19 February 2017 with in-house scripts. In addition, all unclassified *Lactobacillus* assemblies (annotated as *Lactobacillus* sp.; 28 in total) were screened for *L. casei* group members by blast searching (45) them against a filtered RDP database (v11) (14) containing only good-quality *Lactobacillus* 16S rRNA gene sequences longer than 1,200 nucleotides from cultured isolates. This resulted in 15 additional assemblies that were subjected to quality control.

**Quality control and whole-genome GC content.** The quality of the genomic assemblies was evaluated by using the output generated by Quast 4.3 (46). After visualization of different quality control parameters, genomes with N75 values of <10,000 bp and >500 undetermined bases per 100,000 bases were discarded. Subsequently, one genomic assembly (GCA\_001063295) was removed, as it had a genome size of 5.8 Mbp and was identified as a hybrid assembly. The whole-genome GC content was calculated with Quast 4.3 (46) and inforeq from EMBOSS 6.6.0.0 (47), respectively. Visualization was done in R with ggplot2 (48).

**Gene prediction and annotation.** A custom genus-specific BLAST database was created by using all of the complete *Lactobacillus* genomes found in the NCBI database. This database was used in Prokka (49) with the `-usegenus` option to predict genes and annotate all genomic assemblies.

**ANIb and TETRA.** All pairwise ANIb and TETRA values were calculated with the Python pyani package (50) and visualized with the R package ggtree (51).

**Phylogenetic tree inference.** The generation of an alignment of the DNA sequences of a set of single-copy core genes was performed with roary (15) by using a minimum blastp identity of 70% and a threshold of 96% as the percentage of isolates a gene must be in to be defined as a conserved marker gene. Marker genes were translated and compared with a BLAST database of the outgroup GCA\_001434705 (*L. nasuensis* JCM 17158) genome proteins. The DNA sequences of all hits with a coverage of >75% and an identity of >50% were added to the alignment by using in-house scripts. This alignment was used in RAxML 8.2.9 (52) to build a maximum-likelihood phylogenetic tree with the `-a` option, which combines a rapid bootstrap algorithm with an extensive search of the tree space starting from multiple different starting trees. The tree and subtrees were plotted with the R package ggtree (51).

**Orthogroup inference and gene content analysis.** Orthogroups were inferred with OrthoFinder (18) and analyzed in R by using in-house scripts. A core orthogroup is defined as an orthogroup present in >95% of the genomes. All representative sequences of all clade-specific core genes were scanned against a hidden Markov model (HMM) database of all *Bacillus* eggNOG (v4.5) (19) profile HMMs by using HMMER version 3.1b1 (22). Distance matrices based on orthogroup abundance profiles were calculated with the R package vegan (53) and visualized with ggplot2 (48).

The visualization of gene content per functional category was constructed in two steps. First, we mapped all orthogroups to gene families in the eggNOG database. All of the families in this database have functional categories assigned to them, which allowed us to split the orthogroups into 25 functional categories. In the second step, we calculated distance matrices between the genomes on the basis of their orthogroup count profiles constructed by OrthoFinder as described above. We did this separately for each functional category and then visualized the distance matrix of each functional category by performing a PCoA of each of them.

**Interest-driven approaches.** Screening for catalase- and SOD-encoding genes was done by three different methods. First, the presence of the gene of interest was evaluated on the basis of the Prokka annotation. Second, known variants of genes of interest were manually identified from the literature (NCBI database gene accession numbers 1062016, 29637976, and 4413348 for catalase; NCBI database nucleotide accession number HM070825 for SOD) and then blast searched against the pangenome. Third, the Pfam database was used to download HMMs of the protein families of the genes of interest (PF00199 and PF05067 for catalase; PF00080, PF00081, PF02777, and PF09055 for SOD). HMMER (22) was then used to scan the pangenome against these HMMs. GTs were detected by screening a database of orthogroup representative sequences against profile HMMs downloaded from dbCAN (54) supplemented with three Pfam profile HMMs (PF02397, PF02109, and PF04756) with HMMER with an E value cutoff of  $1e-18$ . Further processing and visualization were done with custom R scripts.

**Data availability.** Raw sequence data and assembled contigs of AMBR2 are available at the European Nucleotide Archive under accession number [PRJEB21025](https://doi.org/10.1128/PRJEB21025). All of the scripts used in this study are available at [https://github.com/LebeerLab/caseiGroup\\_mSystems\\_pipeline](https://github.com/LebeerLab/caseiGroup_mSystems_pipeline). The genome of *L. casei* AMBR2 has been deposited in the Belgian Coordinated Collections of Microorganisms under accession number LMG P-30039.

## SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at <https://doi.org/10.1128/mSystems.00061-17>.

**TEXT S1**, TXT file, 0.03 MB.

**TEXT S2**, TXT file, 0.1 MB.

**FIG S1**, EPS file, 0.1 MB.

**FIG S2**, EPS file, 0.04 MB.

**FIG S3**, EPS file, 0.04 MB.

**FIG S4**, EPS file, 0.03 MB.

**TABLE S1**, CSV file, 0.01 MB.

**TABLE S2**, DOCX file, 0.01 MB.

**TABLE S3**, CSV file, 0 MB.

## ACKNOWLEDGMENTS

We thank Eline Oerlemans, Dieter Vandenheuvell, and all other members of the ENdEMIC group for their assistance and/or fruitful discussions. We thank Bart Cuypers (ADReM, University of Antwerp, Antwerp, Belgium), Charlotte Claes and Arvid Suls (Centre of Medical Genetics, University of Antwerp, Antwerp, Belgium), the Lab of Computational Metagenomics (University of Trento, Trento, Italy), and CalcUA (University of Antwerp, Antwerp, Belgium) for their valuable input regarding whole-genome sequencing and data analysis.

Some computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation—Flanders (FWO) and the Flemish Government—department EWI. We acknowledge financial support from the Agency for Innovation by Science and Technology in Flanders (IWT-SB 141198 and IWT-SBO ProCure project [IWT/50052]) and the University of Antwerp (DOCPRO FFB150344) and Ph.D. grants from the Research Foundation Flanders (FWO 1S03516N and FWO 1S17916N).

## REFERENCES

- Sun Z, Harris HMB, McCann A, Guo C, Argimón S, Zhang W, Yang X, Jeffery IB, Cooney JC, Kagawa TF, Liu W, Song Y, Salvetti E, Wrobel A, Rasinkangas P, Parkhill J, Rea MC, O'Sullivan O, Ritari 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>.
- Toh H, Oshima K, Nakano A, Takahata M, Murakami M, Takaki T, Nishiyama H, Igimi S, Hattori M, Morita H. 2013. Genomic adaptation of the *Lactobacillus casei* group. *PLoS One* 8:e75073. <https://doi.org/10.1371/journal.pone.0075073>.
- Stefanovic E, Fitzgerald G, McAuliffe O. 2017. Advances in the genomics and metabolomics of dairy lactobacilli: a review. *Food Microbiol* 61: 33–49. <https://doi.org/10.1016/j.fm.2016.08.009>.
- Iacumin L, Ginaldi F, Manzano M, Anastasi V, Reale A, Zotta T, Rossi F, Coppola R, Comi G. 2015. High resolution melting analysis (HRM) as a new tool for the identification of species belonging to the *Lactobacillus casei* group and comparison with species-specific PCRs and multiplex PCR. *Food Microbiol* 46:357–367. <https://doi.org/10.1016/j.fm.2014.08.007>.
- Reid G. 2017. The development of probiotics for women's health. *Can J Microbiol* 63:269–277. <https://doi.org/10.1139/cjm-2016-0733>.
- Cope EK, Lynch SV. 2015. Novel microbiome-based therapeutics for chronic rhinosinusitis. *Curr Allergy Asthma Rep* 15:504. <https://doi.org/10.1007/s11882-014-0504-y>.
- Dicks LM, Du Plessis EM, Dellaglio F, Lauer E. 1996. Reclassification of *Lactobacillus casei* subsp. *casei* ATCC 393 and *Lactobacillus rhamnosus* ATCC 15820 as *Lactobacillus zeae* nom. rev., designation of ATCC 334 as the neotype of *L. casei* subsp. *casei*, and rejection of the name *Lactoba-*

- cellus paracasei*. Int J Syst Bacteriol 46:337–340. <https://doi.org/10.1099/00207713-46-1-337>.
8. Judicial Commission of the International Committee on Systematics of Bacteria. 2008. The type strain of *Lactobacillus casei* is ATCC 393, ATCC 334 cannot serve as the type because it represents a different taxon, the name *Lactobacillus paracasei* and its subspecies names are not rejected and the revival of the name “*Lactobacillus zeae*” contravenes rules 51b (1) and (2) of the International Code of Nomenclature of Bacteria. Opinion 82. Int J Syst Evol Microbiol 58(Pt 7):1764–1765. <https://doi.org/10.1099/ijs.0.2008/005330-0>.
  9. Smokvina T, Wels M, Polka J, Chervaux C, Brisse S, Boekhorst J, van Hylckama Vlieg JETH, Siezen RJ. 2013. *Lactobacillus paracasei* comparative genomics: towards species pan-genome definition and exploitation of diversity. PLoS One 8:e68731. <https://doi.org/10.1371/journal.pone.0068731>.
  10. Vásquez A, Molin G, Pettersson B, Antonsson M, Ahrné S. 2005. DNA-based classification and sequence heterogeneities in the 16S rRNA genes of *Lactobacillus casei/paracasei* and related species. Syst Appl Microbiol 28:430–441. <https://doi.org/10.1016/j.syapm.2005.02.011>.
  11. Stefanovic E, Kilcawley KN, Rea MC, Fitzgerald GF, McAuliffe O. 2017. Genetic, enzymatic and metabolite profiling of the *Lactobacillus casei* group reveals strain biodiversity and potential applications for flavour diversification. J Appl Microbiol 122:1245–1261. <https://doi.org/10.1111/jam.13420>.
  12. Ward LJ, Timmins MJ. 1999. Differentiation of *Lactobacillus casei*, *Lactobacillus paracasei* and *Lactobacillus rhamnosus* by polymerase chain reaction. Lett Appl Microbiol 29:90–92. <https://doi.org/10.1046/j.1365-2672.1999.00586.x>.
  13. Vásquez A, Ahrné S, Pettersson B, Molin G. 2001. Temporal temperature gradient gel electrophoresis (TTGE) as a tool for identification of *Lactobacillus casei*, *Lactobacillus paracasei*, *Lactobacillus zeae* and *Lactobacillus rhamnosus*. Lett Appl Microbiol 32:215–219. <https://doi.org/10.1046/j.1472-765X.2001.00901.x>.
  14. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42:D633–D642. <https://doi.org/10.1093/nar/gkt1244>.
  15. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693. <https://doi.org/10.1093/bioinformatics/btv421>.
  16. Segers ME, Lebeer S. 2014. Towards a better understanding of *Lactobacillus rhamnosus* GG-host interactions. Microb Cell Fact 13:57. <https://doi.org/10.1186/1475-2859-13-51-57>.
  17. Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 106:19126–19131. <https://doi.org/10.1073/pnas.0906412106>.
  18. Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157. <https://doi.org/10.1186/s13059-015-0721-2>.
  19. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, Von Mering C, Bork P. 2016. EGGNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286–D293. <https://doi.org/10.1093/nar/gkv1248>.
  20. Zotta T, Ricciardi A, Parente E, Reale A, Ianniello RG, Bassi D. 2016. Draft genome sequence of the respiration-competent strain *Lactobacillus casei* N87. Genome Announc 4:e00348-16. <https://doi.org/10.1128/genomeA.00348-16>.
  21. Ianniello RG, Zotta T, Matera A, Genovese F, Parente E, Ricciardi A. 2016. Investigation of factors affecting aerobic and respiratory growth in the oxygen-tolerant strain *Lactobacillus casei* N87. PLoS One 11:e0164065. <https://doi.org/10.1371/journal.pone.0164065>.
  22. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29–W37. <https://doi.org/10.1093/nar/gkr367>.
  23. Zotta T, Ricciardi A, Ianniello RG, Parente E, Reale A, Rossi F, Iacumin L, Comi G, Coppola R. 2014. Assessment of aerobic and respiratory growth in the *Lactobacillus casei* group. PLoS One 9:e99189. <https://doi.org/10.1371/journal.pone.0099189>.
  24. Bron PA, Tomita S, Mercenier A, Kleerebezem M. 2013. Cell surface-associated compounds of probiotic lactobacilli sustain the strain-specificity dogma. Curr Opin Microbiol 16:262–269. <https://doi.org/10.1016/j.mib.2013.06.001>.
  25. Lebeer S, Vanderleyden J, De Keersmaecker SCJ. 2010. Host interactions of probiotic bacterial surface molecules: comparison with commensals and pathogens. Nat Rev Microbiol 8:171–184. <https://doi.org/10.1038/nrmicro2297>.
  26. Bensing BA, Seepersaud R, Yen YT, Sullam PM. 2014. Selective transport by SecA2: an expanding family of customized motor proteins. Biochim Biophys Acta 1843:1674–1686. <https://doi.org/10.1016/j.bbamcr.2013.10.019>.
  27. Salvetti E, Torriani S, Felis GE. 2012. The genus *Lactobacillus*: a taxonomic update. Probiotics Antimicrob Proteins 4:217–226. <https://doi.org/10.1007/s12602-012-9117-8>.
  28. Douillard FP, Ribbera A, Kant R, Pietilä TE, Järvinen HM, Messing M, Randazzo CL, Paulin L, Laine P, Ritari J, Caggia C, Lähtinen T, Brouns SJJ, Satokari R, von Ossowski I, Reunanen J, Palva A, de Vos WM. 2013. Comparative genomic and functional analysis of 100 *Lactobacillus rhamnosus* strains and their comparison with strain GG. PLoS Genet 9:e1003683. <https://doi.org/10.1371/journal.pgen.1003683>.
  29. Rochat T, Gratadoux JJ, Gruss A, Corthier G, Maguin E, Langella P, Van De Guchte M. 2006. Production of a heterologous nonheme catalase by *Lactobacillus casei*: an efficient tool for removal of H<sub>2</sub>O<sub>2</sub> and protection of *Lactobacillus bulgaricus* from oxidative stress in milk. Appl Environ Microbiol 72:5143–5149. <https://doi.org/10.1128/AEM.00482-06>.
  30. Watterlot L, Rochat T, Sokol H, Cherbuy C, Bouloufa I, Lefèvre F, Gratadoux JJ, Honvo-Hueto E, Chilmonczyk S, Blugeon S, Corthier G, Langella P, Bermúdez-Humarán LG. 2010. Intra-gastric administration of a superoxide dismutase-producing recombinant *Lactobacillus casei* BL23 strain attenuates DSS colitis in mice. Int J Food Microbiol 144:35–41. <https://doi.org/10.1016/j.jfoodmicro.2010.03.037>.
  31. Wang G, Yin S, An H, Chen S, Hao Y. 2011. Coexpression of bile salt hydrolase gene and catalase gene remarkably improves oxidative stress and bile salt resistance in *Lactobacillus casei*. J Ind Microbiol Biotechnol 38:985–990. <https://doi.org/10.1007/s10295-010-0871-x>.
  32. Lin J, Zou Y, Cao K, Ma C, Chen Z. 2016. The impact of heterologous catalase expression and superoxide dismutase overexpression on enhancing the oxidative resistance in *Lactobacillus casei*. J Ind Microbiol Biotechnol 43:703–711. <https://doi.org/10.1007/s10295-016-1752-8>.
  33. Chaillou S, Champomier-Vergès MC, Cornet M, Crutz-Le Coq AM, Dudez AM, Martin V, Beauflis S, Darbon-Rongère E, Bossy R, Loux V, Zagorec M. 2005. The complete genome sequence of the meat-borne lactic acid bacterium *Lactobacillus sakei* 23K. Nat Biotechnol 23:1527–1533. <https://doi.org/10.1038/nbt1160>.
  34. Liu Q, Hang X, Liu X, Tan J, Li D, Yang H. 2012. Cloning and heterologous expression of the manganese superoxide dismutase gene from *Lactobacillus casei* Lc18. Ann Microbiol 62:129–137. <https://doi.org/10.1007/s13213-011-0237-2>.
  35. Watanabe M, van der Veen S, Nakajima H, Abee T. 2012. Effect of respiration and manganese on oxidative stress resistance of *Lactobacillus plantarum* WCFS1. Microbiology 158:293–300. <https://doi.org/10.1099/mic.0.051250-0>.
  36. An H, Zhou H, Huang Y, Wang G, Luan C, Mou J, Luo Y, Hao Y. 2010. High-level expression of heme-dependent catalase gene *katA* from *Lactobacillus sakei* protects *Lactobacillus rhamnosus* from oxidative stress. Mol Biotechnol 45:155–160. <https://doi.org/10.1007/s12033-010-9254-9>.
  37. Igarashi T, Kono Y, Tanaka K. 1996. Molecular cloning of manganese catalase from *Lactobacillus plantarum*. J Biol Chem 271:29521–29524. <https://doi.org/10.1074/jbc.271.47.29521>.
  38. Kleerebezem M, Boekhorst J, van Kranenburg R, Molenaar D, Kuipers OP, Leer R, Turchini R, Peters SA, Sandbrink HM, Fiers MWEJ, Stiekema W, Lankhorst RMK, Bron PA, Hoffer SM, Groot MNN, Kerkhoven R, de Vries M, Ursing B, de Vos WM, Siezen RJ. 2003. Complete genome sequence of *Lactobacillus plantarum* WCFS1. Proc Natl Acad Sci U S A 100:1990–1995. <https://doi.org/10.1073/pnas.0337704100>.
  39. Brooijmans RJW, De Vos WM, Hugenholtz J. 2009. *Lactobacillus plantarum* WCFS1 electron transport chains. Appl Environ Microbiol 75:3580–3585. <https://doi.org/10.1128/AEM.00147-09>.
  40. Zotta T, Parente E, Ricciardi A. 2017. Aerobic metabolism in the genus *Lactobacillus*: impact on stress response and potential applications in the food industry. J Appl Microbiol 122:857–869. <https://doi.org/10.1111/jam.13399>.
  41. Ianniello RG, Ricciardi A, Parente E, Tramutola A, Reale A, Zotta T. 2015. Aeration and supplementation with heme and menaquinone affect survival to stresses and antioxidant capability of *Lactobacillus casei*

- strains. *LWT Food Sci Technol* 60:817–824. <https://doi.org/10.1016/j.lwt.2014.10.020>.
42. Frese SA, MacKenzie DA, Peterson DA, Schmaltz R, Fangman T, Zhou Y, Zhang C, Benson AK, Cody LA, Mulholland F, Juge N, Walter J. 2013. Molecular characterization of host-specific biofilm formation in a vertebrate gut symbiont. *PLoS Genet* 9:e1004057. <https://doi.org/10.1371/journal.pgen.1004057>.
  43. van der Woude MW. 2011. Phase variation: how to create and coordinate population diversity. *Curr Opin Microbiol* 14:205–211. <https://doi.org/10.1016/j.mib.2011.01.002>.
  44. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. <https://doi.org/10.1089/cmb.2012.0021>.
  45. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. <https://doi.org/10.1186/1471-2105-10-421>.
  46. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29:1072–1075. <https://doi.org/10.1093/bioinformatics/btt086>.
  47. Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular biology open software suite. *Trends Genet* 16:276–277. [https://doi.org/10.1016/S0168-9525\(00\)02024-2](https://doi.org/10.1016/S0168-9525(00)02024-2).
  48. Wickham H. 2009. ggplot2: elegant graphics for data analysis. Springer Verlag, New York, NY.
  49. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>.
  50. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. 2016. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods* 8:12–24. <https://doi.org/10.1039/C5AY02550H>.
  51. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 8:28–36. <https://doi.org/10.1111/2041-210X.12628>.
  52. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>.
  53. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H. 2016. Vegan: community ecology package. R Foundation, Vienna, Austria.
  54. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* 40:W445–W451. <https://doi.org/10.1093/nar/gks479>.