Title: Genetic and phenotypic features defining industrial relevant *Lactococcus lactis*, *L. cremoris* and *L. lactis* biovar. diacetylactis strains

Authors: Mariano Torres Manno, Federico Zuljan, Sergio Alarcón, Luis Esteban, Victor Blancato, Martín Espariz, Christian Magni

PII: S0168-1656(18)30522-4
DOI: https://doi.org/10.1016/j.jbiotec.2018.06.345
Reference: BIOTEC 8202

To appear in: *Journal of Biotechnology*

Received date: 24-2-2018
Revised date: 22-6-2018
Accepted date: 22-6-2018

Please cite this article as: Manno MT, Zuljan F, Alarcón S, Esteban L, Blancato V, Espariz M, Magni C, Genetic and phenotypic features defining industrial relevant *Lactococcus lactis*, *L. cremoris* and *L. lactis* biovar. diacetylactis strains, *Journal of Biotechnology* (2018), https://doi.org/10.1016/j.jbiotec.2018.06.345

Genetic and phenotypic features defining industrial relevant *Lactococcus lactis*, *L. cremoris* and *L. lactis* biovar. diacetylactis strains.

Mariano Torres Manno[1, 2], Federico Zuljan[1, 2], Sergio Alarcón[1, 3], Luis Esteban[2], Victor Blancato[1, 2] ,Martín Espariz[*1, 2], and Christian Magni[*1, 2].

[1] Laboratorio de Biotecnología e Inocuidad de los Alimentos. Facultad de Ciencias Bioquímicas y Farmacéuticas – Municipalidad de Granadero Baigorria (Universidad Nacional de Rosario, Suipacha 590, Rosario, Argentina.).

[2] Laboratorio de Genética y Fisiología de Bacterias Lácticas. Instituto de Biología Molecular y Celular de Rosario - IBR, CONICET – UNR.

[3] IQUIR Instituto de Química de Rosario, CONICET-UNR.

[4] Facultad de Ciencias Médicas, Universidad Nacional de Rosario. Santa Fe 3000, Rosario, Argentina.

[*] **Corresponding authors:** Mailing address: Instituto de Biología Molecular y Celular de Rosario (IBR), E-mail: magni@ibr-conicet.gov.ar or espariz@ibr-conicet.gov.ar

Highlights:

- New scheme for species and biovar. definition of *L. lactis* group is presented.
- Genomic analyses showed that *L. lactis* and *L. cremoris* are two different species.
- A rank of gene markers for MLSA was performed to classify *L. lactis* group strains.
- Industrial *L. lactis* biovar diacetylactis shows low citrate cluster diversity.

Abstract

Lactococcus lactis strains constitute one of the most important starter cultures for cheese production. In this study, a genome-wide analysis was performed including 68 available genomes of *L. lactis* group strains showing the existence of two species (*L. lactis* and *L. cremoris*) and two biovars (*L. lactis* biovar. diacetylactis and *L. cremoris* biovar. lactis). The proposed classification scheme revealed coherency among phenotypic (through *in silico* and *in vivo* bacterial function profiling), phylogenomic (through maximum likelihood trees) and genomic (using overall genome sequence-based parameters) approaches. Strain biodiversity for the industrial biovar. diacetylactis was also analyzed, finding they are formed by at least three variants with the CC1 clonal complex as the only one distributed worldwide. These findings and methodologies will help improve the selection of *L. lactis* group strains for industrial use as well as facilitate the interpretation of previous or future research studies on this diverse group of bacteria.

Keywords: *Lactococcus lactis*, *Lactococcus cremoris*, biovar. diacetylactis, citrate metabolism, cheese starters

1. Introduction

Lactococcus belongs to the Lactic Acid Bacteria (LAB) group which comprises twelve species. They can be isolated from different niches such as milk, fermented food, plants, insect gastrointestinal tract, and fish (Meucci et al., 2015; Yan Yang et al., 2016). *L. lactis* strains are widely used as starter cultures in artisanal or industrial dairy. Industrial interest on *L. lactis* strains divides this group in three phenotypes: Lactis phenotype (growth at 40ºC in 4% NaCl, and arginine breakdown), Cremoris phenotype (unable to grow at 40ºC or in 4% NaCl, and unable to break down arginine) and Diacetylactis phenotype (citrate fermentation with acetoin and diacetyl production) (Garvie and Farrow, 1982). Strains with different phenotypes provide specific features to cheese quality, for example Cremoris phenotype strains are widely used in the production of cheddar cheese reducing bitterness, while Diacetylactis phenotype strains improve the quality of cheese with buttery aroma and CO_2 production. In particular, this biovar. is commonly used as starter culture in cheese production of fresh cheese, soft ripened or unripened products such as Brie, Camembert and semi-hard cheese such as Gouda or Edam-like cheese (Curioni and Bosset, 2002).

In recent years, next-generation sequencing (NGS)-based sequencing technology has allowed to develop powerful tools and procedures for the classification of highly phylogenetically-related microorganisms. They include whole-genome multilocus sequence analysis (MLSA), and overall genome parameters such as average nucleotide identity (ANI) and *in silico* DNA-DNA hybridization (*is*-DDH) values (Espariz et al., 2016). Based on comparative analysis of ANI parameters Canavagh et al., have

suggested a revision of the taxonomy of *L. lactis* group strains (Cavanagh et al., 2015a; Cavanagh et al., 2015b). Moreover, Kelleher et al. have recently reported a functional comparison of completely sequenced *L. lactis* group strains that clearly show a phylogenetic division between the *L. lactis* and *L. cremoris* strains (Kelleher et al., 2017). While these strains were historically divided into subspecies categories based on industrially relevant phenotypic properties (Wegmann et al., 2007), genome-wide analyses, currently driving the selection of strains, require a more accurate classification. On the other hand, biovar. diacetylactis is one of the most important starters used in dairy manufacture due to its capacity to increase the amount of diacetyl in cheese (Kelly et al., 2010). This property arises from the ability of the diacetylactis group to degrade citrate present in milk (Passerini et al., 2013; Zuljan et al., 2014).

In this study, we present a simplified taxonomic scheme including relevant phenotypes associated with the use of lactococci in food production. Additionally, the presence of three evolutionary events that define the citrate-fermenting *L. lactis* biovar. diacetylactis is described. Remarkably, all diacetylactis strains used as starter culture belong to the same clonal complex. In sum, our results contribute to expand current knowledge regarding lactococcal strains used in the dairy industry, enabling its rational classification and future screenings.

2. Material and Methods

2.1. Comparative analysis of the *L. lactis* group genomes.

For pipeline construction, genome sequences of *L. lactis* group strains listed in Table 1 were used. Comparative genome analysis among lactococcal species was performed uploading all sequences of *L. lactis* listed in Tables 1 and S1 to the RAST server (Rapid Annotation using Subsystem Technology) (Aziz et al., 2008).

2.2. Multilocus sequence typing (MLST) and phylogenomic tree construction.

MLST was performed with MLST 1.8 online on the Center of Genomic Epidemiology server (https://cge.cbs.dtu.dk/services/MLST/). Phylogenomic tree construction was performed as described in Espariz et al. (2016) with minor modifications. Briefly, orthologous genes were assigned using all CDS of *L. lactis* 1403 as queries for bidirectional best hit BLAST searches (Boratyn et al., 2013) against the CDS of all bacterial genomes under study (Table 1) and an E-value of $1E^{-30}$. Orthologous genes present in all microorganisms including the outgroup strain *L. garvieae* TB25 (Blast-defined common ancestral genes) were individually aligned, trimmed and concatenated. The resulting alignment was used to infer the evolutionary history of strains with Maximum Likelihood algorithm and GTR Gamma distributed model using RAxML software. Support values of the branches of the tree were computed with 1000 bootstrap replicates. Substitution model parameters were optimized for each different gene segment of the alignment under study (Stamatakis, 2014).

2.3. Hierarchical clustering analysis.

Hierarchical clustering of strains was performed based on their inferred biological functions as described in Espariz et al. (2016). Consequently, the presence or absence of biological functions in the microorganisms were used as binary score and analyzed by average hierarchical clustering implemented by the R package pvclust (Suzuki and Shimodaira, 2013). Distance measurements were computed by the Manhattan distance function. Biological functions of proteins were inferred by correlation with its ortholog group assignation using the OrthoMCL software (Chen et al., 2006) and an E-value of $1E^{-5}$. In case that a particular species had more than one protein from the same group of orthologs, only the protein with the lower E-value was considered for the clustering

analysis. In case that OrthoMCL did not assign an ortholog group to a particular protein, its function was correlated from its best matching OrthoMCL-DB protein.

2.4. Random Forest approach for the generation of a new MLSA scheme.

For the construction of decision trees, a Random Forest (RF) algorithm (Breiman, 2001) as described in Espariz et al. (2016) was used. Distances of Blast-defined common ancestral genes were calculated using the R package ape (Paradis et al., 2004) and used as variables. On the other hand, the classes (or outputs) used were the suggested names of species that resulted from the overall genome parameter, phylogenomic and functional repertoire analyses. Seventeen strains were arbitrarily selected and used to train the forest. To this end, 100000 classification trees were constructed with a seed value of 12345. Variable importance was computed using internal out-of-bag estimates as described by Breiman (2001). Seventeen strains of the testing set were used to construct a confusion table and calculate its misclassification rate. Strains used as training or testing sets are indicated in Table 1. The 11 most important genes were analyzed for the design of a new MLSA as described in section 2.2, but including strains listed in Tables 1 and S1.

3. Results and discussion

3.1. Species definition and marker gene selection of *L. lactis* group strains.

In order to perform a revision of species assignations for *L. lactis* group strains a pipeline recently described in Espariz et al. (2016) was conducted with 21 strains assigned as *L. lactis* subsp. *lactis* (here called *L. lactis*) and 13 strains assigned as *L. lactis* subsp. *cremoris* (here called *L. cremoris*) (Table 1). First, an MLSA was performed since it is considered the method of choice to efficiently resolve phylogenetic relationships at the genera and species levels (Glaeser and Kampfer, 2015). Thus, the evolutionary history of the strains was inferred using the information from the aligned,

concatenated and pruned sequences of 170 common ancestral genes. These genes were selected from the 1142 core genes associated to *L. lactis* and *L. cremoris* (here called *L. lactis* group strains) since they are also present in the outgroup strain *L. garvieae* TB25 (Ricci et al., 2012). Hence, these genes could rarely be acquired after speciation through horizontally transferred events and presumably evolved following a topology similar to the organisms under study (therefore here called common ancestral genes). As shown in the phylogenetic tree depicted in Fig. 1A two different branches for *L. lactis* and *L. cremoris* strains were clearly observed.

As previously proposed, the analysis of coherence between functional properties and genotypes among strains could contribute to better define species categories (Espariz et al., 2016; Train et al., 2017). Therefore, functions encoded in the *L. lactis* group strains listed in Table 1 were annotated and compared (Table S2). Among the 2286 inferred functions, 976 were found to be present in all 33 analyzed *L. lactis* group strains constituting their core functions (strain CECT 4433 could not be annotated by RAST or NCBI and was not considered in this analysis). Interestingly, the number of core functions identified was very similar to the 904 found by Kelleher et al. (2017) using 30 completely-sequenced genomes. As expected, the dendrogram constructed according to the presence-absence of each assigned function (Fig. 1B) revealed a topology similar to the phylogenetic dendrogram (Fig. 1A). Hence, both analyses support the idea that *L. cremoris* and *L. lactis* strains should be considered different species (Cavanagh et al., 2015a; Cavanagh et al., 2015b; Laroute et al., 2017) in order to better perform comparative analyses.

To select specific marker genes for species classification of the *L. lactis* group strain, a procedure previously described was performed (Espariz et al., 2016). Consequently, the genetic distances among the 170 common ancestral genes of 34 *L. lactis* group

strains listed in Table 1 were calculated generating 5780 variables for the construction of a RF classifier. We found that the species assignations of strains were predicted accurately in all cases tested. Also, each gene was ranked according to its importance in the classification by using the internal out-of-bag estimation of the RF algorithm (Breiman, 2001) (Table S3). The 11 most important genes were listed in Table 2. Finally, 34 additional *L. lactis* group strains , listed in Table S1, that became available during the preparation of this manuscript were added to the original 34 strains to generate phylogenetic trees. Ten trees were constructed using one (*pyrG*) and two to ten gene concatenated sequences where *cshA, clpE, llrC, rpsS, yjjG, ylaF, ldh, glnQ,* and *uvrA* sequences were successively added to the previous concatenated one. The *eno* gene was not included in the analysis considering that only 32% of the complete GL2 gene was available. As expected, all phylogenetic trees had similar topologies and clustered *L. lactis* and *L. cremoris* branches separately (data not shown). Moreover, the *pyrG* tree by itself could support *L. lactis/L. cremoris* bipartition with a bootstrapping measure of 99%. Remarkably, addition of *cshA* and *clpE* sequences could simultaneously resolve the *L. lactis/L. cremoris* node and the *L. cremoris* with Lactis phenotype branch (here called biovar lactis) with bootstrapping values superior to 99% (Fig. S1). This supports the idea that *L. cremoris* strains have two genetic lineages, the Lactis and Cremoris phenotype strains, as was recently suggested by Laroute et al. (2017).

3.2. *L. lactis* biovar diacetylactis group definition based on specific features.

Concerning the importance of citrate metabolism in the production of aroma compounds such as diacetyl, a thorough inspection of the *cit* genetic locus in *L. lactis* group strains was performed using BLAST (Boratyn et al., 2013). With this approach two different *cit* chromosomal cluster types were found (Fig. 2A and B). In the

archetypal strain CRL264 (Zuljan et al., 2016a; Zuljan et al., 2016b) citrate transport and metabolism is composed by the chromosomal *cit* cluster *citM(ψcitO)-I-CDEFXG* (coding for the citrate lyase complex and soluble oxaloacetate decarboxylase) located near the *als* gene (Martin et al., 2004) and by an 8.5 kbp plasmid encoding the *citP* gene (citrate transporter) (Sesma et al., 1990) (Fig. 2A). This arrangement of the *cit* locus (here called type A) was found in diacetylactis strains IL1403 (plasmid-free strain derived from IL594), CRL264, TIFN2, TIFN4, and LD61, and the recently sequenced strains M20, DRA4, and LMG 19460 (Fig. 2A). The latter is a plasmid-free strain derived from the *L. lactis* biovar. diacetylactis Bu2 (a citrate-fermenting strain) (Jahns et al., 1991). In the Bpl1 and KF67 strains an alternative *cit*$^{+}$ locus (here called type B) was found. Although this *cit* cluster has a similar genetic architecture to the type A *cit* cluster, they did not share significant nucleotide identities as discussed below (Fig. 2A and B). Moreover, the *cit* cluster was located adjacent to the lipoteichoic acid synthase *ltaS* rather than to the *als* gene in Bpl1 and KF67 (Fig. 2B). They shared approximately 98% nucleotide identity between them.

A more exhaustive analysis showed that some differences exist even in clusters of the same type. It was observed that M20, TIFN2, TIFN4 and KF67 strains encode a full-length copy of the putative CitO transporter in their respective clusters. These putative proteins display highly shared amino acid similarity when compared to the uncharacterized citrate transporter associated to the *cit* cluster present in *Oenococcus oeni* (Mills et al., 2005) and to the malate transporters associated with malolactic enzyme present in *O. oeni* and other related bacteria (Espariz et al., 2011). Distinctly, the *citO* gene has a stop codon in Bpl1 (two putative proteins of 248 and 76 amino acid residues would be produced). In addition, *citO* is described as a pseudogene in CRL264, IL1403, LD61, LMG 19460 and the Bpl1 strain.

When cluster A and B citrate lyase holoenzymes (complexes composed of CitD, CitE and CitF subunits) were compared, shared amino acid similarities of approximately 80% were found whereas accessory proteins, such as the citrate lyase ligase CitC, showed amino acid similarities of ~54% (Fig. 2A and B). Enzymes CitX and CitG responsible for the biosynthesis of the prosthetic group of citrate lyase were fused in one polipeptide CitX(G) in strains Bpl1, KF67 and M20, remaining as two gene products in strains IL1403, CRL264, TIFN2, TIFN4, LD61, and DRA4. In the case of the regulator CitI and the soluble oxaloacetate decarboxylase CitM, shared amino acid similarities < 58 % and <63% were found, respectively (Fig. 2A and B).

No citrate cluster was identified in the reported chromosomal genome sequence of *L. lactis* diacetylactis GL2, available from the NCBI site (Gabed et al., 2015). In this strain, citrate transporter CitP, the citrate lyase complex and the soluble oxaloacetate decarboxylase were encoded in a large 23 Kb plasmid (Drici et al., 2010), also found in other genus such as *Leuconostoc* and *Weissella* (Bekal-Si Ali et al., 1999; Martin et al., 1999). A different *cit* gene context as well as distribution was described for GL2, which indicates that a third type of *cit* cluster may be found in *L. lactis* strains (here called type C; Fig. 2C). Altogether, these facts suggest that at least three non-related genetic events have come together to give origin to the diacetylactis biovar strains. This also highlights the fact that acquisition of *cit* clusters in the *L. lactis* strains has arisen by horizontal transfer, which correlates with the observation that diacetylactis biovars were not clustered in a single functional or phylogenetic group (Fig. 1).

3.3. Clonal origin of the industrial *L. lactis* biovar diacetylactis strains.

In an attempt to better characterize the phylogenetic relationship of diacetylactis strains, a MLST analysis following the scheme of the Center of Genomic Epidemiology was performed. It was found that IL1403, LD61, TIFN2, LMG19460 and TIFN4 strains

belong to ST6 showing 100% identity in all the analyzed alleles. On the other hand, GL2 and DRA4 belong to ST1, and ST16 haplotypes, respectively, whereas M20, Bpl1, and KF67 to different and undefined ST haplotypes. Notably, all starter culture strains analyzed belong to the same clonal complex (CC1) defined by Passerini et al. (2010) and encode the type A citrate-fermenting cluster. Remarkably, CRL264 was first isolated from a Northwestern Argentinean cheese whereas the other members of the CC1 diacetylactis biovar group have European origin. Since industrial processes in cheese production were not introduced by European immigrants in South America until the late nineteenth century, we propose that dissemination of CC1 has originated recently from the central region of Europe, where starter culture strains IL594 (IL1403 parental), LD61, TIFN4, and TIFN6 were isolated (Erkus et al., 2013; Falentin et al., 2014; Gorecki et al., 2011).

On the other hand, strains isolated from soil (M20), insects (Bpl1), grape juice (KF67), and dromedary milk (GL2) harbor different *cit* clusters (Fig. 2). Therefore, while the diacetylactis biovar group is not a monophyletic group (Fig. 1) there is a correlation among industrial history of strains, their phylogenetic origin and the citrate-degrading pathway they have acquired during their evolution.

4. Conclusion

Despite the existence of guidelines and recommendations to ensure stability, reproducibility, and coherence in taxonomy, the methodology to circumscribe strains in species is still subjective and arbitrary (Gevers et al., 2005; Stackebrandt et al., 2007). *Lactococcus* classification is an example of species demarcation which is not defined by a theory-based concept but generally by a practical necessity or industrial praxis. However, an accurate species assignation extremely impacts the way industrial strains are selected due to the fact that such assignation, implicitly or not, is used to predict

strains behavior or performance (Gevers et al., 2005). Nowadays, the new DNA sequencing techniques are increasing the amounts of data that may contribute to improve the accuracy in bacteria circumscription, which is mainly based on their phylogenetic, genomic, and phenotypic coherence (Rossello-Mora, 2012). In this study, a new simplified scheme that includes a phylogenetic-based analysis of the common ancestral genes *pyrG* and *cshA* (and optionally *clpE*) was established (Fig. 3). This allows differentiation of *L. lactis* from *L. cremoris* as well as *L. cremoris* biovar. lactis among *L. cremoris* strains. This scheme resolves more accurately the approach suggested by Rademaker et al. (2007) and will improve the classification of the new isolates of the genera *Lactococcus* (Fig. 3). Considering the incorporation of this methodology into industrial routine processes we have pondered in favor of simplicity and velocity at the cost of accuracy. We are aware that some imprecisions could arise from the fact that genes could be horizontally transferred and suffer recombination. Also, from a statistical point of view, few genes represent only a fraction of the genome (Colston et al., 2014; Gevers et al., 2005). To avoid these artifacts, or at least reduce their impact, obvious modifications to the proposed scheme will be the inclusion of more common ancestral genes (listed in Table S3) that would be selected based on their importance rank. In addition, the gene combination analyzed could be modified based on their availability (i.e. in case some genes are not completely or accurately sequenced).

Citrate consumption was recognized as the major determinant of aroma production in the *L. lactis* biovar diacetylactis strains (Passerini et al., 2013). Hence, as depicted in Fig. 3 a genotypic and/or phenotypic analysis is proposed for inclusion in our scheme in order to identify citrate consumption as well as acetoin and diacetyl production capability. In this work, *cit* clusters not related to the well-known CRL264

strain were identified. Therefore, we propose that the presence of the conserved citrate lyase gene *citF* should be analyzed as a feature-encoded marker rather than the plasmidic *citP* gen (Kempler and McKay, 1980; Laroute et al., 2017). Moreover, the Diacetylactis phenotype should be validated while growing the strains in Kempler and McKay medium (Kempler and McKay, 1980; Laroute et al., 2017) and/or by the Voges-Proskauer assay using citrate as carbon source as described by Martino et al. (2016). Finally, considering that the use of diverse environmental strains in industrial processes is widely accepted (Passerini et al., 2013), further studies on citrate fermentation and aroma compound generation in M20, Bpl1, and KF67 are required in order to obtain alternative starters available for the cheese industry.

Acknowledgements

We would like to thank Agencia Nacional de Promoción Científica y Tecnológica (ANPyCT, PICT 2014-1513 and PICT 2015-2361) and CONICET (11220110100718CO and 11220150100855CO) for financial support. We would also like to thank the staff from the English Department (Facultad de Ciencias Bioquímicas y Farmacéuticas, UNR) for language correction of the manuscript. MTM and FZ are CONICET fellows. ME, SA, VB and CM are researchers of the same institution.

References

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., Zagnitko, O., (2008) The RAST Server: rapid annotations using subsystems technology. BMC genomics 9, 75.

Bekal-Si Ali, S., Divies, C., Prevost, H., (1999) Genetic organization of the *citCDEF* locus and identification of mae and *clyR* genes from *Leuconostoc mesenteroides*. Journal of bacteriology 181, 4411-4416.

Boratyn, G.M., Camacho, C., Cooper, P.S., Coulouris, G., Fong, A., Ma, N., Madden, T.L., Matten, W.T., McGinnis, S.D., Merezhuk, Y., Raytselis, Y., Sayers, E.W., Tao, T., Ye, J., Zaretskaya, I., (2013) BLAST: a more efficient report with usability improvements. Nucleic acids research 41, W29-33.

Breiman, L., (2001) Random Forests. Machine Learning 45, 5-32.

Cavanagh, D., Casey, A., Altermann, E., Cotter, P.D., Fitzgerald, G.F., McAuliffe, O., (2015a) Evaluation of *Lactococcus lactis* Isolates from Nondairy Sources with Potential Dairy Applications Reveals Extensive Phenotype-Genotype Disparity and Implications for a Revised Species. Applied and environmental microbiology 81, 3961-3972.

Cavanagh, D., Fitzgerald, G.F., McAuliffe, O., (2015b) From field to fermentation: the origins of *Lactococcus lactis* and its domestication to the dairy environment. Food microbiology 47, 45-61.

Colston, S.M., Fullmer, M.S., Beka, L., Lamy, B., Gogarten, J.P., Graf, J., (2014) Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using *Aeromonas* as a test case. mBio 5, e02136.

Curioni, P.M.G., Bosset, J.O., (2002) Key odorants in various cheese types as determined by gas chromatography-olfactometry. International Dairy Journal 12, 959-984.

Chen, F., Mackey, A.J., Stoeckert, C.J., Jr., Roos, D.S., (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic acids research 34, D363-368.

Drici, H., Gilbert, C., Kihal, M., Atlan, D., (2010) Atypical citrate-fermenting *Lactococcus lactis* strains isolated from dromedary's milk. Journal of applied microbiology 108, 647-657.

Erkus, O., de Jager, V.C., Spus, M., van Alen-Boerrigter, I.J., van Rijswijck, I.M., Hazelwood, L., Janssen, P.W., van Hijum, S.A., Kleerebezem, M., Smid, E.J., (2013) Multifactorial diversity sustains microbial community stability. The ISME journal 7, 2126-2136.

Espariz, M., Repizo, G., Blancato, V., Mortera, P., Alarcon, S., Magni, C., (2011) Identification of malic and soluble oxaloacetate decarboxylase enzymes in *Enterococcus faecalis*. The FEBS journal 278, 2140-2151.

Espariz, M., Zuljan, F.A., Esteban, L., Magni, C., (2016) Taxonomic Identity Resolution of Highly Phylogenetically Related Strains and Selection of Phylogenetic Markers by Using Genome-Scale Methods: The *Bacillus pumilus* Group Case. PloS one 11, e0163098.

Falentin, H., Naquin, D., Loux, V., Barloy-Hubler, F., Loubiere, P., Nouaille, S., Lavenier, D., Le Bourgeois, P., Francois, P., Schrenzel, J., Hernandez, D., Even, S., Le Loir, Y., (2014) Genome Sequence of *Lactococcus lactis* subsp. *lactis* bv. diacetylactis LD61. Genome announcements 2.

Gabed, N., Yang, M., Bey Baba Hamed, M., Drici, H., Gross, R., Dandekar, T., Liang, C., (2015) Draft Genome Sequence of the Moderately Heat-Tolerant *Lactococcus lactis* subsp. lactis bv. diacetylactis Strain GL2 from Algerian Dromedary Milk. Genome announcements 3.

Garvie, E.I., Farrow, J.A.E., (1982) *Streptococcus lactis* subsp. *cremoris* (Orla-Jensen) comb. nov. and *Streptococcus lactis* subsp. *diacetilactis* (Matuszewski et al.) nom. rev., comb. nov. International Journal of Systematic Bacteriology 32, 453-455.

Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F.L., Swings, J., (2005) Opinion: Re-evaluating prokaryotic species. Nature reviews. Microbiology 3, 733-739.

Glaeser, S.P., Kampfer, P., (2015) Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. Syst Appl Microbiol 38, 237-245.

Gorecki, R.K., Koryszewska-Baginska, A., Golebiewski, M., Zylinska, J., Grynberg, M., Bardowski, J.K., (2011) Adaptative potential of the *Lactococcus lactis* IL594 strain encoded in its 7 plasmids. PloS one 6, e22238.

Jahns, A., Schafer, A., Geis, A., Teuber, M., (1991) Identification, cloning and sequencing of the replication region of *Lactococcus lactis* ssp. lactis biovar. diacetylactis Bu2 citrate plasmid pSL2. FEMS microbiology letters 64, 253-258.

Kelleher, P., Bottacini, F., Mahony, J., Kilcawley, K.N., van Sinderen, D., (2017) Comparative and functional genomics of the *Lactococcus lactis* taxon; insights into evolution and niche adaptation. BMC genomics 18, 267.

Kelly, W.J., Ward, L.J., Leahy, S.C., (2010) Chromosomal diversity in *Lactococcus lactis* and the origin of dairy starter cultures. Genome biology and evolution 2, 729-744.

Kempler, G.M., McKay, L.L., (1980) Improved Medium for Detection of Citrate-Fermenting *Streptococcus lactis* subsp. diacetylactis. Applied and environmental microbiology 39, 926-927.

Laroute, V., Tormo, H., Couderc, C., Mercier-Bonin, M., Le Bourgeois, P., Cocaign-Bousquet, M., Daveran-Mingot, M.L., (2017) From Genome to Phenotype: An Integrative Approach to Evaluate the Biodiversity of *Lactococcus lactis*. Microorganisms 5.

Martin, M., Corrales, M.A., de Mendoza, D., Lopez, P., Magni, C., (1999) Cloning and molecular characterization of the citrate utilization *citMCDEFGRP* cluster of *Leuconostoc paramesenteroides*. FEMS microbiology letters 174, 231-238.

Martin, M.G., Sender, P.D., Peiru, S., de Mendoza, D., Magni, C., (2004) Acid-inducible transcription of the operon encoding the citrate lyase complex of *Lactococcus lactis* biovar diacetylactis CRL264. Journal of bacteriology 186, 5649-5660.

Martino, G.P., Quintana, I.M., Espariz, M., Blancato, V.S., Magni, C., (2016) Aroma compounds generation in citrate metabolism of *Enterococcus faecium*: Genetic characterization of type I citrate gene cluster. International journal of food microbiology 218, 27-37.

Meucci, A., Zago, M., Rossetti, L., Fornasari, M.E., Bonvini, B., Tidona, F., Povolo, M., Contarini, G., Carminati, D., Giraffa, G., (2015) *Lactococcus hircilactis* sp. nov. and *Lactococcus laudensis* sp. nov., isolated from milk. International journal of systematic and evolutionary microbiology 65, 2091-2096.

Mills, D.A., Rawsthorne, H., Parker, C., Tamir, D., Makarova, K., (2005) Genomic analysis of *Oenococcus oeni* PSU-1 and its relevance to winemaking. FEMS microbiology reviews 29, 465-475.

Paradis, E., Claude, J., Strimmer, K., (2004) APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289-290.

Passerini, D., Beltramo, C., Coddeville, M., Quentin, Y., Ritzenthaler, P., Daveran-Mingot, M.L., Le Bourgeois, P., (2010) Genes but not genomes reveal bacterial domestication of *Lactococcus lactis*. PloS one 5, e15306.

Passerini, D., Laroute, V., Coddeville, M., Le Bourgeois, P., Loubiere, P., Ritzenthaler, P., Cocaign-Bousquet, M., Daveran-Mingot, M.L., (2013) New insights into *Lactococcus lactis* diacetyl- and acetoin-producing strains isolated from diverse origins. International journal of food microbiology 160, 329-336.

Rademaker, J.L., Herbet, H., Starrenburg, M.J., Naser, S.M., Gevers, D., Kelly, W.J., Hugenholtz, J., Swings, J., van Hylckama Vlieg, J.E., (2007) Diversity analysis of dairy and nondairy *Lactococcus lactis* isolates, using a novel multilocus sequence analysis scheme and (GTG)5-PCR fingerprinting. Applied and environmental microbiology 73, 7128-7137.

Ricci, G., Ferrario, C., Borgo, F., Rollando, A., Fortina, M.G., (2012) Genome sequences of *Lactococcus garvieae* TB25, isolated from Italian cheese, and *Lactococcus garvieae* LG9, isolated from Italian rainbow trout. Journal of bacteriology 194, 1249-1250.

Rossello-Mora, R., (2012) Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. Environmental microbiology 14, 318-334.

Sesma, F., Gardiol, D., de Ruiz Holgado, A.P., de Mendoza, D., (1990) Cloning of the citrate permease gene of *Lactococcus lactis* subsp. *lactis* biovar diacetylactis and expression in *Escherichia coli*. Applied and environmental microbiology 56, 2099-2103.

Stackebrandt, E., Pauker, O., Steiner, U., Schumann, P., Straubler, B., Heibei, S., Lang, E., (2007) Taxonomic characterization of members of the genus *Corallococcus*: molecular divergence versus phenotypic coherency. Syst Appl Microbiol 30, 109-118.

Stamatakis, A., (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312-1313.

Suzuki, R., Shimodaira, H., (2013) pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling. R package version 1.2.2. Available at: http://www.sigmath.es.osaka-u.ac.jp/shimo-lab/prog/pvclust/.

Train, C.M., Glover, N.M., Gonnet, G.H., Altenhoff, A.M., Dessimoz, C., (2017) Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. Bioinformatics 33, i75-i82.

Wegmann, U., O'Connell-Motherway, M., Zomer, A., Buist, G., Shearman, C., Canchaya, C., Ventura, M., Goesmann, A., Gasson, M.J., Kuipers, O.P., van Sinderen, D., Kok, J., (2007) Complete genome sequence of the prototype lactic acid bacterium *Lactococcus lactis* subsp. *cremoris* MG1363. Journal of bacteriology 189, 3256-3270.

Yan Yang, S., Zheng, Y., Huang, Z., Min Wang, X., Yang, H., (2016) *Lactococcus nasutitermitis* sp. nov. isolated from a termite gut. International journal of systematic and evolutionary microbiology 66, 518-522.

Zuljan, F., Espariz, M., Blancato, V.S., Esteban, L., Alarcon, S., Magni, C., (2016a) Draft Genome Sequence of *Lactococcus lactis* subsp. lactis bv. diacetylactis CRL264, a Citrate-Fermenting Strain. Genome announcements 4.

Zuljan, F.A., Mortera, P., Alarcón, S.H., Blancato, V.S., Espariz, M., Magni, C., (2016b) Lactic acid bacteria decarboxylation reactions in cheese. International Dairy Journal 62, 53-62.

Zuljan, F.A., Repizo, G.D., Alarcon, S.H., Magni, C., (2014) alpha-Acetolactate synthase of *Lactococcus lactis* contributes to pH homeostasis in acid stress conditions. International journal of food microbiology 188, 99-107.

Figure captions:

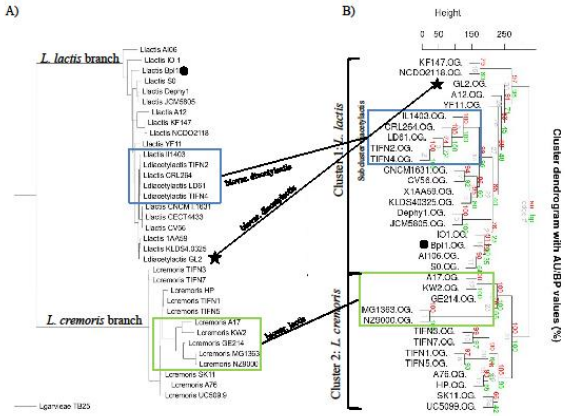
Fig. 1. Comparison of phylogenomic and functional dendrograms of *L. lactis* group strains. A) Phylogenomic dendrogram. 170 Blast-defined common ancestral genes were individually aligned, concatenated and trimmed. The resulting final alignment was used to infer the evolutionary history of the indicated strains using a maximum-likelihood approach in MEGA 5. B) Functional dendrogram. Biological functions of proteins encoded in the genome of the indicated strain were inferred using the OrthoMCL software and then used as binary score for hierarchical clustering implemented with the R pvcluster package. The diacetylactis ST6 strains are indicated with blue boxes. Strain GL2 (ST1) is highlighted with black-filled stars while the diacetylactis strain Bpl1, found through *cit* genes inspection, with black-filled circles. *L. cremoris* biovar lactis strains are indicated with green boxes.

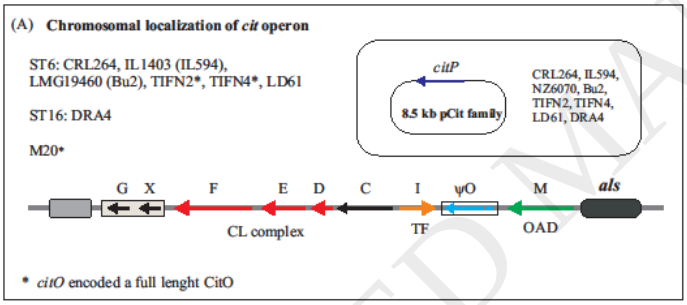
Fig. 2. Citrate utilization gene clusters found in *L. lactis* biovar. diacetylactis strains. A) *cit* cluster A architecture found in the ST6, M20 and DRA4 strains. The presence of the *citO* pseudogene (*ψcitO*) is complemented with an 8.5 kb plasmid carrying *citP*. B) *cit* cluster B architecture found in Bpl1 and KF67. C) Reported plasmidic organization of the *cit* cluster C in GL2.

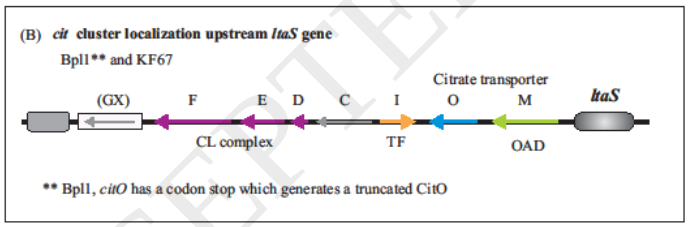
Fig. 3. Newly simplified scheme for *L. lactis* group strains classification. First, a phylogenic approach based on a *pyrG* gene sequence is proposed in order to differentiate *L. lactis* and *L. cremoris* strains. Further classification could be performed with the inclusion of the *cshA* gene that allows the identification of *L. cremoris* biovar lactis strains. Addition of the *clpE* (optional) gene increases the robustness of the phylogenic analysis (see Fig. S1). Finally, a featured-based approach should be included to identify *L. lactis* biovar. diacetylactis strains. Citrate consumption and C4 production capability could be inferred by the identification of *citF* by PCR, or more directly by

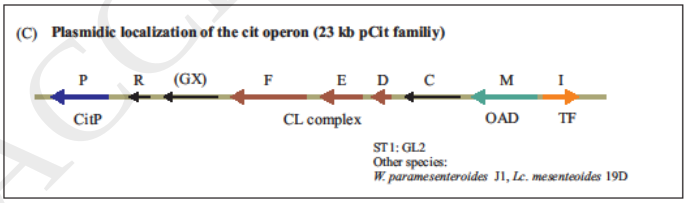
growing strains in Kempler and McKay (K-M) medium and/or via the Voges-Proskauer (V-P) assay in M17 supplemented with citrate.

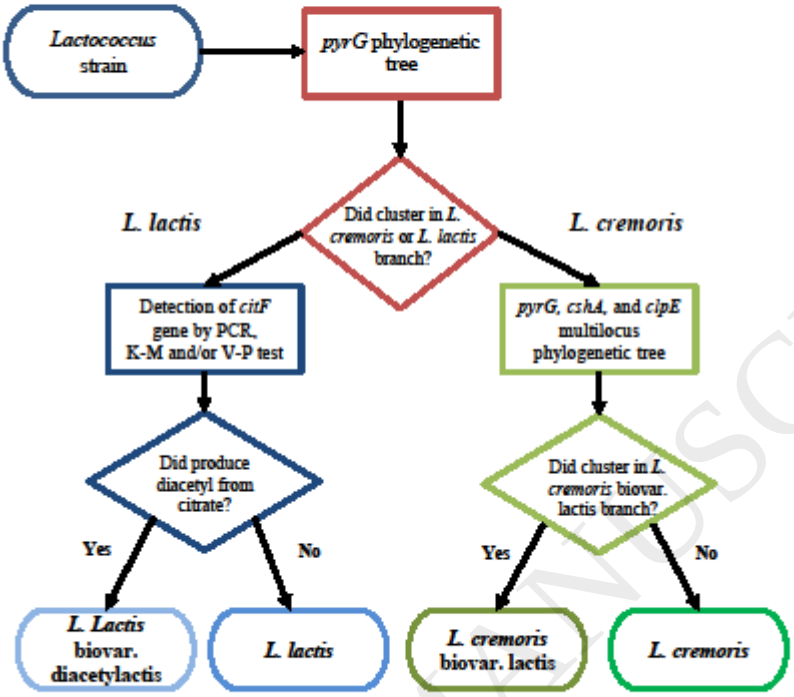
Figures:











Tables:

Table1 *Lactococcus* strains used in this study.

NCBI taxonomic label	Strain name	Origin	Assembly[1]	Size(Mb)	Scaffolds	Genes	Proteins	RF[2]
L. lactis	IL1403	Dairy	GCA_000006865.1	2.36559	1	2406	2277	Train
L. cremoris	MG1363	Dairy	GCA_000009425.1	2.52948	1	2583	2400	Test
L. cremoris	SK11	Dairy	GCA_000014545.1	2.59835	6	2682	2412	Test
L. lactis	KF147	Plant	GCA_000025045.1	2.63565	2	2595	2445	Test
L. cremoris	NZ9000	Dairy	GCA_000143205.1	2.53029	1	2583	2404	Train
L. lactis	CV56	Human	GCA_000192705.1	2.51874	6	2533	2378	Train
L. cremoris	A76	Dairy	GCA_000236475.1	2.5771	5	2679	2382	Test
L. cremoris	UC509.9	Dairy	GCA_000312685.1	2.45735	9	2489	2188	Train
L. lactis	IO-1	Drain water	GCA_000344575.1	2.42147	1	2342	2230	Test
L. cremoris	KW2	Dairy	GCA_000468955.1	2.42705	1	2345	2223	Train

L. lactis	NCDO 2118	Plant	GCA_000478255.2	2.59226	2	2545	2382	Test
L. lactis	KLDS 4.0325	Dairy	GCA_000479375.2	2.59549	4	2648	2448	Train
L. lactis	AI06	Amazonian plant	GCA_000761115.1	2.39809	1	2320	2178	Test
L. lactis	S0	Dairy	GCA_000807375.1	2.4887	1	2482	2311	Train
L. lactis	Dephy 1	Dairy	GCA_000493355.1	2.60355	56	2634	2459	Test
L. lactis	CNCM I-1631	Dairy	GCA_000284735.1	2.51133	131	2546	2403	Train
L. lactis	YF11	Fermented corn	GCA_000348965.1	2.52731	71	2452	2328	Test
L. lactis	A12	Sourdough bread	GCA_000442845.1	2.7	42	2663	2425	Train
L. cremoris	TIFN1	Dairy	GCA_000447885.1	2.67978	291	2828	2285	Train
L. lactis biovar. diacetylactis	TIFN2	Dairy	GCA_000447905.1	2.50507	143	2603	2296	Test
L. cremoris	TIFN3	Dairy	GCA_000447925.1	2.72521	412	2869	2291	Test
L. lactis biovar.	TIFN4	Dairy	GCA_00044798	2.55039	182	2635	2349	Trai

diacetylactis			5.1					n
L. lactis biovar. diacetylactis	LD61	Dairy	GCA_000488975.1	2.59924	132	2687	2490	Train
L. cremoris	HP	Dairy	GCA_000534815.1	2.26951	213	2325	2042	Test
L. cremoris	GE214	Dairy	GCA_000731635.1	2.80103	243	2835	2603	Train
L. lactis	Bpl1	Insect	GCA_000759595.1	2.3057	64	2200	2092	Train
L. lactis	CECT 4433[3]	Dairy	GCA_000761565.1	2.57915	111	2629	2290	Test
L. lactis	1AA59	Dairy	GCA_000786755.1	2.57654	218	2570	2406	Test
L. cremoris	A17	Dairy	GCA_000805435.1	2.67994	16	2551	2372	Train
L. lactis	JCM 5805	Dairy	GCA_000835975.1	2.54579	88	2553	2359	Train
L. lactis biovar. diacetylactis	CRL264	Dairy	GCA_001455265.1	2.57372	83	2650	2446	Test
L. lactis biovar.	GL2	Dromedary	GCA_000721865.2	2.2454	48	2179	2022	Test

diacetylactis								
L. cremoris	TIFN5	Dairy	GCA_000447825.1	2.54151	646	2555	2232	Train
L. cremoris	TIFN7	Dairy	GCA_000447965.1	2.63409	370	2853	2505	Test
L. garvieae	TB25	Cheese	GCA_000236515.3	2.00888	91			-

[1] GeneBank assembly accession number.

[2] The sequences were used in Random Forest algorithm as Train or Test. See text for details

[3] This sequence was not used in ANI and *is*-DDH calculations.

Table 2. Eleven most important genes for *L. lactis* group strains classification.

Genes	Function[1]	mean[2]	variance[2]	maximun[3]	Importance[3]
pyrG	GATase1 CTP Synthase	0.04540152	8.97E-08	0.29253211	0.00166292
cshA	recombination factor protein RarA	0.10504793	6.20E-07	0.43475447	0.00165721
clpE	ATP-dependent protease ATP-binding subunit	0.05519423	4.73E-08	0.30766005	0.00165573
llrC	two-component systems; involved in acid stress resistance development	0.0545341	4.67E-08	0.28488155	0.00165217
rpsS	30S ribosomal protein S19	0.01751105	4.36E-10	0.10784057	0.0016476
eno	phosphopyruvate hydratase	0.15664178	4.35E-06	0.89668675	0.00164746
yjjG	hypothetycal protein	0.02865096	8.34E-10	0.2267373	0.00164598
ylaF	nicotinate phosphoribosyltransferase	0.08017139	2.24E-07	0.37720716	0.00164551
ldh	lactate dehydrogenase	0.03380669	7.25E-09	0.2833626	0.00164437
glnQ	glutamine ABC transporter ATP-binding protein	0.08334978	3.70E-07	0.33834446	0.00164403
uvrA	excinuclease ABC subunit A	0.08805436	7.63E-07	0.3352154	0.00163455

[1] Function and locus names for each gene were obtained for the reference sequence of *L. lactis* IL1403.
[2] Distances of orthologs genes were calculated using the R package ape and then used to compute their means, variances and maximums.
[3] Importance of each gene was computed using internal out-of-bag estimates as described by Breiman, L. (2001) with a forest composed by 100000 classification trees, trained by the 17 strains mentioned in Table 1 and the input data of all 170 core genes.