microbial biotechnology

Microbial Biotechnology (2011) **4**(3), 383–402 doi:10.1111/j.1751-7915.2011.00247.x

Genome-scale diversity and niche adaptation analysis of *Lactococcus lactis* by comparative genome hybridization using multi-strain arrays

Roland J. Siezen,[1,2,3,4]* Jumamurat R. Bayjanov,[2,4] Giovanna E. Felis,[1,5] Marijke R. van der Sijde,[2,3] Marjo Starrenburg,[1] Douwe Molenaar,[1,3†] Michiel Wels,[1,2,3] Sacha A. F. T. van Hijum[1,2,3,4] and Johan E. T. van Hylckama Vlieg[1,3‡]

[1]*Kluyver Centre for Genomics of Industrial Fermentation, NIZO food research, P.O. Box 20, 6710 BA Ede, the Netherlands.*

[2]*Center for Molecular and Biomolecular Informatics, Radboud University Medical Centre, PO Box 9101, Nijmegen, the Netherlands.*

[3]*TI Food and Nutrition, P.O. Box 557, 6700 AN Wageningen, the Netherlands.*

[4]*Netherlands Bioinformatics Centre, 260 NBIC, P.O. Box 9101, 6500 HB Nijmegen, the Netherlands.*

[5]*Department of Biotechnology, University of Verona, Strada le Grazie 15 – Ca' Vignal 2, 37134 Verona, Italy.*

Summary

Lactococcus lactis **produces lactic acid and is widely used in the manufacturing of various fermented dairy products. However, the species is also frequently isolated from non-dairy niches, such as fermented plant material. Recently, these non-dairy strains have gained increasing interest, as they have been described to possess flavour-forming activities that are rarely found in dairy isolates and have diverse metabolic properties. We performed an extensive whole-genome diversity analysis on 39** ***L. lactis*** **strains, isolated from dairy and plant sources. Comparative genome hybridization analysis with multi-strain microarrays was used to assess presence or absence of genes and gene clusters in these strains, relative to all** ***L. lactis*** **sequences in public databases, whereby chromosomal and plasmid-encoded genes were computationally analysed separately. Nearly 3900 chromosomal orthologous groups (chrOGs) were defined on basis of four sequenced chromosomes of** ***L. lactis*** **strains (IL1403, KF147, SK11, MG1363). Of these, 1268 chrOGs are present in at least 35 strains and represent the presently known core genome of** ***L. lactis***, **and 72 chrOGs appear to be unique for** ***L. lactis*****. Nearly 600 and 400 chrOGs were found to be specific for either the subspecies** ***lactis*** **or subspecies** ***cremoris*** **respectively. Strain variability was found in presence or absence of gene clusters related to growth on plant substrates, such as genes involved in the consumption of arabinose, xylan, α-galactosides and galacturonate. Further niche-specific differences were found in gene clusters for exopolysaccharides biosynthesis, stress response (iron transport, osmotolerance) and bacterial defence mechanisms (nisin biosynthesis). Strain variability of functions encoded on known plasmids included proteolysis, lactose fermentation, citrate uptake, metal ion resistance and exopolysaccharides biosynthesis. The present study supports the view of** ***L. lactis*** **as a species with a very flexible genome.**

Received 15 October, 2010; accepted 28 December, 2010. *For correspondence. E-mail r.siezen@cmbi.ru.nl; Tel. +31 (0)24-36 19559; Fax +31 (0)24 36 19395.

Present addresses: †Systems Bioinformatics IBIVU, Free University of Amsterdam, 1081HV Amsterdam, the Netherlands; ‡Danone Research, Gut and Microbiology Platform, R.D. 128, 91767 Palaiseau Cedex, France.

Authors' contributions: G.F., D.M., R.S. and J.H.V. conceived the study. G.F. and M.S. performed the experimental work, while J.B., D.M., M.W., and M.v.d.S. performed the bioinformatics analyses. R.S., G.F. and J.H.V. wrote the article. S.v.H. supervised the bioinformatics analyses performed by M.v.d.S. and J.B. and corrected the article. All authors have read and approved the final manuscript.

Introduction

The Gram-positive bacterium *Lactococcus lactis* has been an important model organism for low-GC Gram-positive bacteria for many years. The primary reason for the interest in this species is the extraordinary industrial importance of *L. lactis* strains as primary components of dairy starter cultures. Genetic techniques have been widely applied in recent years to unravel the molecular basis of industrially important phenotypic traits. Complete genome sequences of three different *L. lactis* strains of dairy origin have been published, further improving our knowledge of strains used in dairy technology (Bolotin *et al.*, 2001; Makarova *et al.*, 2006; Wegmann *et al.*, 2007). The abundant occurrence of *L. lactis* strains outside the dairy environment was already known for decades (Sandine, 1972), but recently it has been rediscovered due to ecological interest and technological properties of non-dairy strains in an applied context

(van Hylckama Vlieg *et al*., 2006). The complete genome sequence of a *L. lactis* plant isolate has recently been determined and has provided a more complete view on the genomic diversity of the species *L. lactis* (Siezen *et al*., 2008; 2010). The existence of many plasmids reported for *L. lactis* further enlarges the genetic pool and thereby the number of possible phenotypic manifestations from different combinations of chromosomal and plasmid pools (Campo *et al*., 2002; Bolotin *et al*., 2004; Siezen *et al*., 2005).

Taxonomically, three subspecies (ssp. *lactis*, ssp. *cremoris* and ssp. *hordniae*) and one biovar (ssp. *lactis* biovar *diacetylatis*) are recognized. These are the results of reclassification of now discontinued taxa, first recognized as different species (*Streptococcus lactis, Streptococcus cremoris* and *Lactobacillus hordniae*), subsequently united under the genus *Lactococcus* and species *lactis* (the historical summary of species naming is reported in van Hylckama Vlieg *et al*. 2006). The discrimination between subspecies is formally linked to a few phenotypic tests (i.e. growth at 40°C, growth at 4% NaCl, deamination of arginine, and acid production from maltose, lactose, galactose and ribose) (Rademaker *et al*., 2007). However, phenotypic and genetic relationships do not always correlate among strains of the same subspecies, leading to considerable confusion in taxonomy (Tailliez *et al*., 1998). In fact all possible combinations of *lactis* and *cremoris* phenotypes and genotypes have been reported, although with different incidence (Kelly and Ward, 2002).

In recent years, comparative genome hybridization (CGH), sometimes referred to as genomotyping, has been increasingly applied to unravel the gene content of bacterial strains (Molenaar *et al*., 2005; Peng *et al*., 2006; Earl *et al*., 2007; Han *et al*., 2007; La *et al*., 2007; McBride *et al*., 2007; Wang *et al*., 2007; Rasmussen *et al*., 2008; Siezen *et al*., 2010). A recent CGH analysis of five *L. lactis* ssp. *cremoris* strains provided a first insight into diversity of genes and gene clusters, but was limited by the fact that the DNA microarray used for CGH specified only 1030 genes selected from the genome of a single strain *L. lactis* ssp. *cremoris* SK11, which is less than half of the genes encoded in its genome (Taibi *et al*., 2010). Therefore many of the potential genomic variations were not assessed. Chromosomal diversity of a large collection of *L. lactis* strains was recently screened on the basis of their phenotype and the macrorestriction patterns produced from pulsed-field gel electrophoresis (PFGE) analysis of SmaI digests of genomic DNA, providing insight into chromosomal size and architecture variation (Kelly *et al*., 2010).

In the current study, we performed a CGH analysis of 39 *L. lactis* strains using a multi-strain, high-resolution NimbleGen microarray, in an attempt to cover the presently known *L. lactis* pan-genome. These strains were selected from a much larger set of phenotypically and genotypically characterized *L. lactis* strains (Rademaker *et al*., 2007). The strains represent different subspecies (*cremoris, lactis, hordniae*), different phenotypic groups, and were isolated from different environmental niches. They are therefore believed to be a representative sample of diversity of the species (Table 1).

Our objectives were (i) to gain insight into the genetic diversity based on whole-genome gene content, and compare it with the results of other techniques (e.g. genome fingerprints and MLSA analysis (Rademaker *et al*., 2007), (ii) to compare chromosomal and plasmid diversity, (iii) to estimate and characterize the core genome of the species, and (iv) to analyse genes and gene clusters specific for subclades of strains. These results contribute to a more complete insight into the diversity and niche adaptation of the species.

Results

Diversity in gene distribution and population structure

A CGH analysis was performed to investigate the gene content of 39 strains of *L. lactis*. Analysis of all core genes from sequenced genomes shows that nucleotide sequence identity between strains from the same subspecies is high: sequence identity is 99% between *L. lactis* ssp. *lactis* strains IL1403 and KF147, and it is 98% between *L. lactis* ssp. *cremoris* strains SK11 and MG1363. This is in sharp contrast to the average sequence identity of only 88% that was observed between ssp. *lactis* and ssp. *cremoris* strains. Because strains from different subspecies can be quite diverse in sequence conservation and gene content (Lan and Reeves, 2000; Medini *et al*., 2005), we used a multi-strain microarray instead of a single reference strain array. This multi-strain array based on NimbleGen technology contains multiple overlapping probes targeting all known *L. lactis* genes in the NCBI database and is therefore better suited to detect the expected relatively large differences in nucleotide sequence identity. As with any CGH analysis, its limitation remains that novel genes that are not represented on the array will not be detected.

The hybridization of DNA from the query genomes to the probes on the multi-strain array was translated into absence or presence of genes in orthologous groups. The hybridization efficiency of DNA from the four reference strains shows that 96–99% of the known genes in these genomes were positively identified using our PanCGH algorithm (Table 2).

Phylogenetic relationships of strains are basically reflected in differences in chromosomal sequence and content, although adaptation to different environmental niches is also related to acquisition or loss of mobile

Table 1. Strains included in the analysis.

Strain code	Internal collection code	Isolation source	Other information
Lactococcus lactis ssp. *lactis* genotype and a *L. lactis* ssp. *lactis* phenotype			
ATCC19435^T	NIZO 29T	Milk (dairy starter)	
Li-1	NIZO 1156	Grass	
E34	NIZO 1173	Silage	
N42	NIZO 1230	Soil and grass	
DRA4	NIZO 1592	Dairy starter A	*L. lactis* ssp. *lactis* biovar *diacetylactis*
ML8	NIZO 20	Dairy starter	
LMG9446, NCFB1867	NIZO 2123	Frozen peas	
LMG9449, NCFB1868	NIZO 2124	Frozen peas	
K231	NIZO 2199	White kimchii	
K337	NIZO 2202	White kimchii	
NCDO895, NCIMB700895	NIZO 2211	Dairy starter	
KF7	NIZO 2219	Alfalfa sprouts	
KF24	NIZO 2220	Alfalfa sprouts	
KF67	NIZO 2223	Grapefruit juice	
KF134	NIZO 2226	Alfalfa and radish sprouts	
KF146	NIZO 2229	Alfalfa and radish sprouts	
KF147	NIZO 2230	Mung bean sprouts	
KF196	NIZO 2236	Japanese kaiwere shoots	
KF201	NIZO 2238	Sliced mixed vegetables	
B2244B	NIZO 3919	Mustard and cress	
KF282	NIZO 3920	Mustard and cress	
LMG14418	NIZO 2424	Bovine milk	
IL1403	NIZO 2441	Dairy starter	Plasmid-free derivative of *L. lactis* ssp. *lactis* biovar *diacetylactis* CNRZ157(IL594)
LMG8526, NCFB2091	NIZO 26	Chinese radish seeds	
UC317	NIZO 644	Dairy starter	
M20	NIZO 844	Soil	*L. lactis* ssp. *lactis* biovar *diacetylactis*
P7304	NIZO 2207	Litter on pastures	rRNA most related to isolates from prawns
P7266	NIZO 2206	Litter on pastures	rRNA most related to isolates from prawns
Lactococcus lactis ssp. *cremoris* genotype and a *L. lactis* ssp. *lactis* phenotype			
V4	NIZO 1157	Raw sheep milk	
KW10	NIZO 2249	Kaanga way	
NCDO763, ML3	NIZO 643	Dairy starter	Derivative of NCDO712
MG1363	NIZO 1492	Cheese starter	Plasmid-free derivative of NCDO712
N41	NIZO 1175	Soil and grass	
Lactococcus lactis ssp. *cremoris* genotype and a *L. lactis* ssp. *cremoris* phenotype ('true *cremoris*' strains)			
LMG6897^T	NIZO 2418T	Cheese starter	Subculture of strain HP
FG2	NIZO 2252	Dairy starter	
AM2	NIZO 33	Dairy starter	
HP	NIZO 42	Dairy starter	
SK11	NIZO 32	Dairy starter	Phage-resistant derivative of AM1
Lactococcus lactis ssp. *hordniae*			
LMG8520^T	NIZO 24T	Leaf hopper (insect)	

Table 2. Hybridization and genotyping accuracy for the four reference strains.

Genotyping	IL1403	KF147	MG1363[b]	SK11
OGs with at least one gene from reference strain	2286	2428	2406	2289
OGs with score NA[a]	132	181	274	109
OGs correctly identified as 'present' (true positives)	2101	2226	2056	2130
OGs incorrectly identified as 'absent' (false negatives)	53	21	76	50
True-positive rate	97.5%	99.1%	96.4%	97.7%
False-negative rate	2.5%	0.9%	3.6%	2.3%

a. NA means that the presence/absence of an OG could not be calculated, either because the corresponding genes were not represented on the microarray, or due to an insufficient number of probes matching to members of this OG (by default at least 10 probes must be aligned).
b. Note that strain MG1363 was not used in the CGH array design, and therefore the positive recall for this strain was slightly lower than for the other three reference strains.

Table 3. Chromosomal orthologous groups (chrOGs), derived from pan-genome CGH analysis, and their presence in *L. lactis* strains according to different criteria.

Analysed groups	Number	Other information
Total orthologous groups	3877	Based on four sequenced *L. lactis* genomes[a]
Core chrOGs for sequenced genomes	1513	Based on four sequenced *L. lactis* genomes[a]
Number of groups reliably analysable by CGH	3255	622 OGs not on array or not analysed
Core chrOGs for the species *L. lactis* (37 strains)	1121	Strains P7266 and P7304 omitted
Core chrOGs for the species *L. lactis* (35 strains)	1268	Strains KW10 and KF282 also omitted; see Table S1
Core chrOGs linked to LaCOGs	1246	Table S1
Core chrOGs only in *L. lactis*	72	Table S2; not in other LAB
Variable chrOGs in 35 strains	1987	See distribution in Fig. 2

a. *Lactococcus lactis* ssp. *cremoris* strains SK11 and MG1363, *L. lactis* ssp. *lactis* strains IL1403 and KF147.

elements (plasmids, phages, IS elements, transposons, etc.), and the interchange between mobile elements and the chromosome is well documented in lactococci. We analysed chromosomal orthologous groups (chrOGs) separately from plasmid orthologous groups (pOGs). For chrOGs, the PanCGH algorithm was used to translate hybridization signals into presence or absence of orthologous groups, rather than individual genes (Bayjanov *et al*., 2009; 2010). In total, 3877 chrOGs were defined on the basis of presence of genes in chromosomes of the four fully sequenced strains (IL1403, KF147, SK11 and MG1363). A total of 622 chrOGs were targeted by fewer than 10 probes per chrOG, and therefore excluded by the PanCGH algorithm from the analysis, reducing the total number of chrOGs investigated to 3255 (Table 3).

The complete data set of chrOGs was used to cluster the *L. lactis* strains on the basis of presence/absence of chrOGs (Fig. 1). Strains were clearly separated into two major clades corresponding to the subspecies *lactis* and *cremoris*. This confirms previous results of genotypic and phenotypic studies on these *Lactococcus* strains (Rademaker *et al*., 2007). Our whole chromosome-based tree is most similar to their tree based on a five-locus MLST cluster analysis, but our tree contains much more genomic information on strain diversity, as demonstrated below. The two major subspecies groups are further subdivided into subclades in the whole-genome tree (Fig. 1). For the ssp. *lactis* strains, dairy and plant *lactis* isolates are in separate subclades, while in the ssp. *cremoris* strains, the two subclades correspond to the two different phenotypes, i.e. the *lactis*-like and *cremoris*-like phenotypes. The type strain LMG8520^T of *L. lactis* ssp. *hordniae*, isolated from leaf hoppers, appears to have a *lactis*-like genomic content, and is grouped with plant isolates.

Core genes of L. lactis

Core genes are those that are conserved in all strains and are typically involved in the essential cellular processes of a species. Strains P7304 and P7266 were not included in this analysis, because their chromosomal sequences deviate too much from the other strains, resulting in too many false negatives in the hybridization signals (see the text in *Supporting information*). The distribution of presence shows that there are 1121 chrOGs present in the 37 *L. lactis* strains (Fig. 2A), which we coin as 'core chrOGs'.

Another 2134 chrOGs contain genes which do not appear to be present in all strains, and of these, 280 chrOGs are found in 36 strains and 79 chrOGs in 35 strains. From the genes that lack in only one strain, most are absent in KW10 (72 chrOGs) or in KF282 (70 chrOGs), possibly due to chromosomal sequence variations leading to poor hybridization signals. Since strains KW10 and KF282 show an aberrant gene presence/absence pattern compared with strains with the same genotype, the core genome would be considerably larger if these strains were also left out from the analysis (Fig. 2B). When considering only the remaining 35 strains, 1268 chrOGs constitute the core genome; a full list of these core genes in the four reference genomes and their encoded functions is presented in Table S1 in *Supporting information*. Amazingly, about 180 core chrOGs (14%) consist of proteins with as yet unknown function (hypothetical proteins), and many more encode proteins with only a general function annotated (e.g. general enzyme or transporter family predicted only). These results show that there is still much unknown about the core gene functions of lactococci.

*Linking core chrOGs to LaCOGs (*Lactobacillales-*specific Clusters of Orthologous Genes)*

The 1268 *L. lactis* core chrOGs were compared with the LaCOGs (*Lactobacillales*-specific Clusters of Orthologous Genes), which represent groups of genes present in at least two out of 12 sequenced LAB genomes (Makarova *et al*., 2006; Makarova and Koonin, 2007) and recently updated to 26 sequenced LAB genomes (Zhou *et al*., 2010). The vast majority (98%) of our core chrOGs were

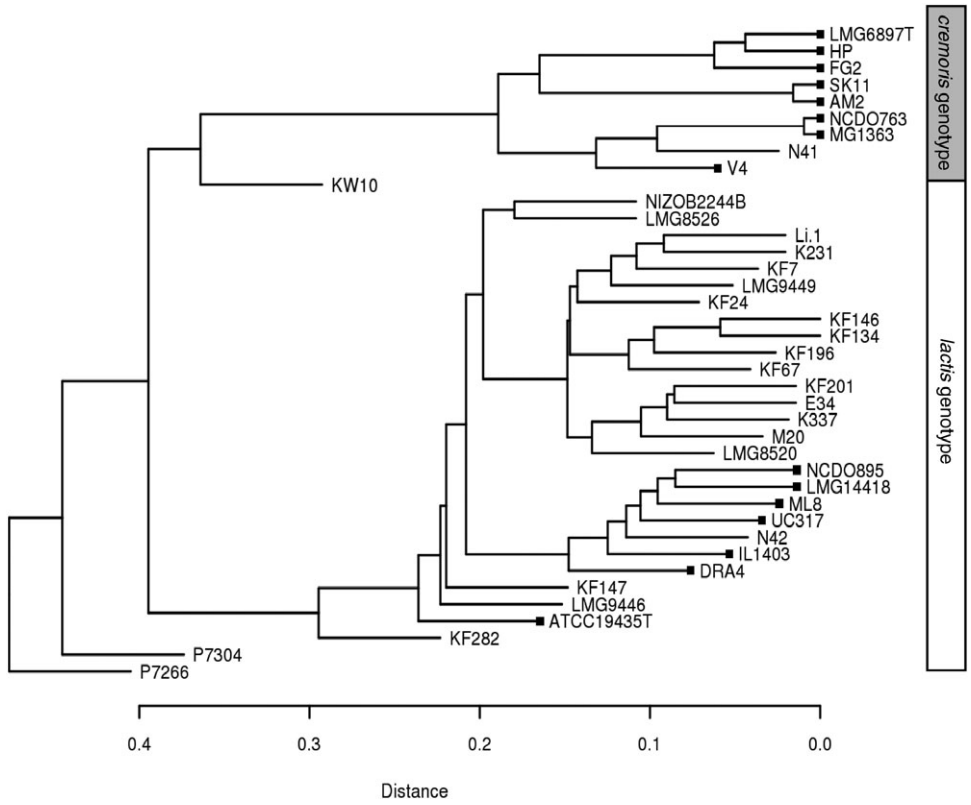

Fig. 1. Whole-genome content-based tree. Hierarchical clustering tree of *L. lactis* strains based on presence/absence of all chromosomal orthologous groups (chrOGs) in these strains. The binary distance metric was used in combination with the average linkage clustering algorithm. Solid rectangles signify dairy isolates, while the other strains signify mainly plant origin. The top clade of 10 strains corresponds to ssp. *cremoris* genotype, further divided into two subclades, corresponding to the two phenotypes, i.e. *cremoris*-like (upper subclade) and *lactis*-like phenotype (lower subclade). The lower clade of 27 strains contains only *L. lactis* ssp. *lactis* and ssp. *hordniae* type strain LMG8520^{T}. This clade grouping ssp. *lactis* strains contains subclades corresponding to isolation source (dairy versus non-dairy). Strains P7266 and P7304 are clustered far apart from the other subspecies with a *lactis* genotype.

unambiguously linked to the LaCOGs (Table 3 and Table S1 in *Supporting information*). Interestingly, in the initial definition of LaCOGs (Makarova *et al*., 2006), *L. lactis* strains IL1403 and SK11 were considered as separate organisms although they belong to the same species. Therefore, LaCOGs actually include a number of OGs that are specific for the species *L. lactis* (see below). Based on our CGH analysis of 35 strains, we have now identified 72 core chrOGs/LaCOGs which are specific for the *L. lactis* species, in the sense that they are found in all *L. lactis* strains, but do not have homologues in other LAB genome sequences (Table 4; full details in Table S2).

Diversity of chromosomal genes of L. lactis

The occurrence of numerous chrOGs in only a few strains (Fig. 2) supports the hypothesis that the species *L. lactis* is genetically extremely flexible. Therefore, we investigated in more detail the genetic signatures, i.e. chrOGs, genes and gene clusters, linked to the different genomic subclades and to the different isolation niches. Based on total chromosomal gene content, the 37 strains investigated can be divided in two clusters, each including the type strains of the subspecies (Fig. 1). In the following analysis, we first focused on the chrOGs specific for each subspecies clade.

Nearly 600 and 400 chrOGs were found to be specific for either the subspecies *lactis* or subspecies *cremoris* respectively, of which nearly half specified hypothetical proteins of unknown function; full details of these subspecies-specific chrOGs and genes are listed in Table S3. Based on our CGH analysis, a small subset of these subspecies-specific chrOGs appear to be present in all tested *cremoris* (151 chrOGs) or all *lactis* strains (72 chrOGs), and hence these could be used as genotypic marker genes to distinguish the *lactis* and *cremoris* subspecies. Many of these subspecies-specific genes are organized in gene clusters in the reference genomes, and the functions specified by these gene clusters could be used in phenotypic typing. A short summary of the largest gene clusters and their predicted functions is presented in Table 5.

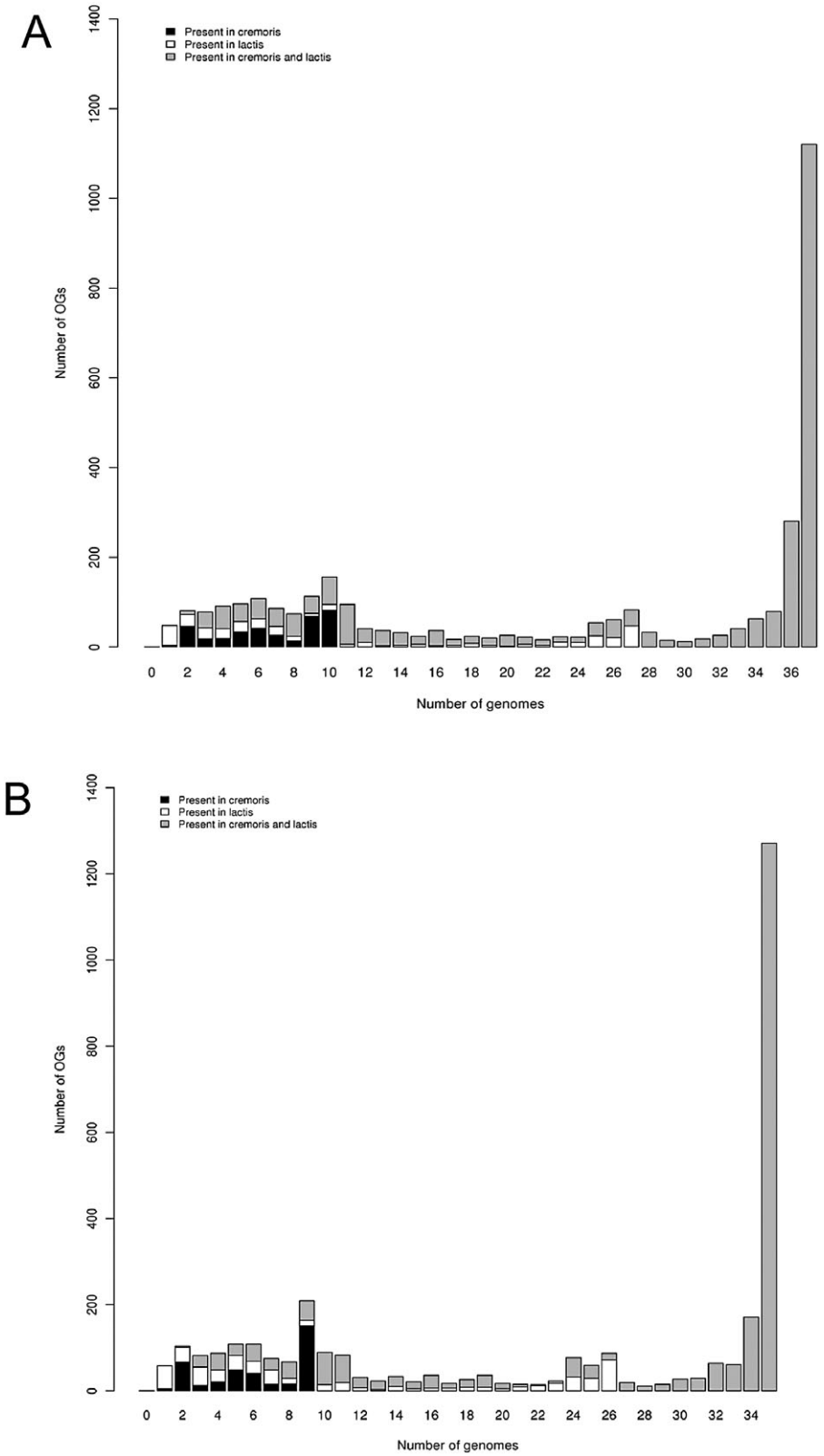

Fig. 2. Distribution of chrOGs in the strains. Distribution of chromosomal orthologous groups (chrOGs) in 37 strains (A) and in 35 strains (B). Strains P7304 and P7266 are omitted in (A) and strains KW10 and KF282 are also omitted in (B), due to ambiguities in hybridization efficiencies (see text). The bar on the outer right represents the total number of chrOGs in the core genome. Shading indicates whether the chrOGs are present only in ssp. *cremoris* strains (black), only in ssp. *lactis* strains (white) or in both subspecies (grey).

Table 4. *Lactococcus lactis* specific core genes with predicted functions[a] in 35 strains.

chrOG id	LaCOG id	Size (AA)[b]	Consensus function	Best hit in non-LAB organism
1626	LaCOG02385	162–180	Acetyltransferase, GNAT family	*Streptococcus* sp.
336	LaCOG02698	152	Acetyltransferase, GNAT family	*Bacteroides*
1134	LaCOG02731	1436	Activator of (R)-2-hydroxyglutaryl-CoA dehydratase	*Streptococcus* sp.
1884	LaCOG02722	213	Aminoglycoside phosphotransferase	*Bacillus* sp.
350	LaCOG02578	379–383	ATP/GTP-binding protein	*Enterococcus* sp.
202	LaCOG02425	784	Carbon starvation protein A	*Propionibacterium freudenreichii*
1125	LaCOG02464	134–151	Dinucleoside polyphosphate hydrolase	*Caminibacter mediatlanticus*
262	LaCOG02721	251–261	Metallophosphoesterase	*Enterococcus* sp.
463	LaCOG02619	462–465	MF superfamily multidrug resistance protein	*Listeria grayi*
174	LaCOG02554	443	NAD(FAD)-utilizing dehydrogenase	*Turicibacter* sp.
2067	LaCOG02712	535	NADH dehydrogenase	*Paenibacillus* sp.
1192	LaCOG02661	101	O6-methylguanine-DNA methyltransferase	*Bacillus* sp.
339	LaCOG02380	145	Osmotically inducible protein C	*Pseudomonas* sp.
1256	LaCOG02428	1190–1223	Pyruvate-flavodoxin oxidoreductase	*Enterococcus* sp.
1483	LaCOG02566	276–296	Rgg/GadR/MutR family transcriptional regulator	*Streptococcus* sp.
2408	LaCOG02658	160–163	SUF system FeS assembly protein	*Nakamurella multipartita*
1011	LaCOG02734	151	Transporter	None
2258	LaCOG02509	143	Universal stress protein	*Enterococcus* sp.
636	LaCOG02670	141	Universal stress protein A	*Enterococcus* sp.
1370	LaCOG02404	269–303	Zinc-binding dehydrogenase	*Streptomyces* sp.

a. For a full list of the 72 *L. lactis*-specific chrOGs see Table S2.
b. Size (in AA) of protein in four reference *L. lactis* genomes.

Gene clusters unique for all ssp. *lactis* strains (and not present in any ssp. *cremoris* strain) include a large cluster of 17 genes for glycan (xylan, mannan or glucan) and xylose metabolism (Table 5), which is typical for plant-derived *lactis* strains as they can use these plant cell-wall components for growth, but apparently this cluster is also maintained in dairy *lactis* strains. In some *lactis* strains, the arabinose-utilization genes are also part this gene cluster (see below). The *thgA–lacZ* genes for galactose metabolism appear to be unique for all *lactis* strains, but are absent in all ssp. *cremoris* strains. Another *lactis*-unique cluster is predicted to be involved in nitrogen metabolism of agmatine and putrescine, both breakdown products of arginine. Several other *lactis*-specific genes are predicted to be involved in stress response (Table 5).

Gene clusters unique for (almost) all ssp. *cremoris* strains (and not present in any ssp. *lactis* strain) include antibiotic resistance, sugar metabolism (α-glucosides, β-glucosides, ribose), but also many hypothetical proteins (Table 5). Many of the *cremoris*-specific gene clusters are identified as pseudogenes in the reference *cremoris* genomes, which could indicate ongoing degeneration of genes and encoded functions.

Subclade-specific clusters

Next, each branch in the tree was investigated separately for gain and loss of chrOGs to determine the degree of relatedness between strains and subclades, and to obtain insight into possible insertions and deletions of genes and gene clusters during diversification. Per split in the tree, the genes in these chrOGs were used to find clusters of adjacent genes in the corresponding reference genomes. Several large gene clusters were identified of which a selection is described below and summarized in Table 6 (others can be found in the text in *Supporting information*). Tree splits, annotation of the gene clusters and their best BLAST hits are presented in detail in Table S4.

Simple sugar metabolism

• *Arabinose metabolism.* Arabinose is a monosaccharide commonly found in plants as a component of biopolymers such as hemicellulose and pectin. Plant *L. lactis* strains KF147 and KF282 have previously been shown to grow on L-arabinose, in contrast to IL1403 and SK11 (Siezen *et al*., 2008). The arabinose operon (Fig. 3A) was indeed found to be specific for plant strains. Only strains N41, KF147, KF282, LMG8526 and B2244B were predicted to contain the complete arabinose gene cluster *araADBTFPR*. Eight other plant *lactis* strains contain an arabinose operon lacking the genes *araFP*, encoding an alpha-*N*-arabinofuranosidase and a disaccharide permease, suggesting that they cannot utilize arabinose polymers/oligomers, but can still use arabinose itself.

• *Sucrose metabolism.* Sucrose is the major stable product of photosynthesis in plants and it is also the form in which most carbon is transported. It has been described that genes for the biosynthesis of nisin and the fermentation of sucrose are located on a 70 kb conjugative transposon in *L. lactis* ssp. *lactis* (Kelly *et al*.,

Table 5. Main subspecies-specific conserved gene clusters and functions.

Locus[a]	Gene	Function	Comment
(A) Subspecies *lactis*-specific			
LLKF_0567	*umuC*	ImpB/MucB/SamB family protein	
LLKF_0568	*yfiC*	Acetyltransferase, GNAT family	
LLKF_0569	*rmaJ*	Transcriptional regulator, MarR family	
LLKF_0570	*yfiE*	Organic hydroperoxide resistance family protein	
LLKF_1314	*nhaP*	NhaP-type Na+/H+ and K+/H+ antiporter	Cluster not in UC317, LMG8520
LLKF_1315	*ymhC*	Hypothetical protein	
LLKF_1316	*amyL*	Alpha-amylase	
LLKF_1317	*lctO*	L-lactate oxidase	
LLKF_1605	*ypcCD*	Endo-beta-*N*-acetylglucosaminidase (EC 3.2.1.96)	Arabinose gene cluster is inserted between *ptk–xylT* in some strains
LLKF_1606	*dexB*	Glucan 1,6-alpha-glucosidase (EC 3.2.1.70)	
LLKF_1607	*lnbA*	Lacto-*N*-biosidase (EC 3.2.1.140)	
LLKF_1608	*ypcG*	Sugar ABC transporter, substrate-binding protein	
LLKF_1609	*ypcH*	Sugar ABC transporter, permease protein	
LLKF_1610	*ypdA*	Sugar ABC transporter, permease protein	
LLKF_1611	*ypdB*	Alpha-mannosidase (EC 3.2.1.24)	
LLKF_1612	*ypdC*	Hypothetical protein	
LLKF_1613	*rliB*	Transcriptional regulator, GntR family	
LLKF_1614	*ypdD*	Alpha-1,2-mannosidase (EC 3.2.1.24)	
LLKF_1615	*ptk*	Phosphoketolase (EC 4.1.2.9)	
LLKF_1623	*xylT*	D-xylose-proton symporter	
LLKF_1624	*xylX*	Acetyltransferase (EC 2.3.1.-)	
LLKF_1625	*xynB*	Beta-1,4-xylosidase	
LLKF_1626	*xynT*	Xyloside transporter	
LLKF_1627	*xylM*	Aldose-1-epimerase (EC 5.1.3.3)	
LLKF_1628	*xylB*	Xylulose kinase (EC 2.7.1.17)	
LLKF_1859	*arcC*	Carbamate kinase (EC 2.7.2.2)	Cluster partially absent in LMG9449; there are other copies of carbamate kinase
LLKF_1860	*aguA*	Agmatine deiminase (EC 3.5.3.12)	
LLKF_1861	*yrfD*	Agmatine/putrescine antiporter	
LLKF_1862	*pctA*	Putrescine carbamoyltransferase (EC 2.1.3.6)	
LLKF_1863		Transcriptional regulator, LuxR family	
LLKF_2026	*corC*	Magnesium and cobalt efflux protein	
LLKF_2027	*pacB*	Penicillin acylase (EC 3.5.1.11)	
LLKF_2028	*ytaD*	Protein-tyrosine phosphatase (EC 3.1.3.48)	
LLKF_2164	*lacZ*	Beta-galactosidase (EC 3.2.1.23)	
LLKF_2165	*thgA*	Galactoside *O*-acetyltransferase (EC 2.3.1.18)	
(B) Subspecies *cremoris*-specific			
LACR_0451		Antibiotic export permease protein	Inserted relative to IL1403, KF147
LACR_0452		Antibiotic export ATP-binding protein	
LACR_0453		Transcriptional regulator, MarR family	
LACR_0498		Hypothetical protein	Cluster unique for *L. lactis*
LACR_0501		Hypothetical protein	Gene absent in FG2, HP
LACR_0502		Hypothetical protein	
LACR_0505		Hypothetical protein	
LACR_0506		Hypothetical protein	
LACR_0507		Hypothetical protein	
LACR_0508		Hypothetical protein	
LACR_0509		Hypothetical protein	
LACR_0754		Hypothetical protein	
LACR_0755		Cold-shock DNA-binding protein family protein	
LACR_0756		Cold-shock DNA-binding protein family protein	
LACR_0761		Sugar ABC transporter permease	In IL1403 a transposase at this position
LACR_0762		Sugar ABC transporter permease	
LACR_0763		Oligosaccharide-binding protein	
LACR_0764		Integral membrane protein	
LACR_0765		Alpha-glucosidase (EC 3.2.1.30)	
LACR_1288		Transcriptional regulator, AraC family	Glycan degradation; similar clusters in *Leuconostoc mesenteroides, Clostridium difficile, Bifidobacteria, Ruminococcus obeum*
LACR_1289		Major facilitator superfamily permease	Gene absent in FG2, HP
LACR_1290		Glucan 1,3-beta-glucosidase (EC 3.2.1.58)	Gene absent in FG2, HP, LMG6897T
LACR_1291		Beta-xylosidase (EC 3.2.1.37)	
LACR_1632		PTS system cellobiose-specific, IIC component	Whole gene cluster absent in V4, KW10
LACR_1633		Transcriptional regulator, AraC family	Gene absent in FG2, HP, LMG6897T
LACR_1636		Glucokinase (EC 2.7.1.2)/transcription regulator	Gene absent in FG2, HP, LMG6897T
LACR_1637		6-Phospho-beta-glucosidase (EC 3.2.1.86)	
LACR_1638	*rpiB*	Ribose-5-phosphate isomerase B (EC 5.3.1.6)	
LACR_1639	*rpe*	Ribulose-5-phosphate 3-epimerase (EC 5.1.3.1)	
LACR_1640		Transcription regulator, LacI family	
LACR_2591		Hypothetical protein	
LACR_2592		Hypothetical protein	
LACR_2593		Hypothetical protein	
LACR_2594		Hypothetical protein	

a. For the conserved OGs, members from a reference genome are listed, i.e. LLKF = *L. lactis* ssp. *lactis* KF147; LACR = *L. lactis* ssp. *cremoris* SK11. Numbering indicates genomic position relative to other chromosomal genes, where consecutively numbered genes are generally in an operon.
These genes are predicted to be present in all strains of a subspecies, either *lactis* or *cremoris*, and absent in all strains of the other subspecies. Exceptions are indicated.

Table 6. Diversity of chromosomally encoded gene clusters and functions.

Strain	Subspecies	Dairy	Arabinose metabolism	Sucrose metabolism	Galacturonate metabolism	α-Galactoside metabolism	Xylan breakdown	Starch/ maltose breakdown	Trp metabolism	Leu/Ile/Val metabolism	Citrate metabolism	High-affinity K+ transport	Nisin production/ immunity	EPS biosynthesis (*epsX–epsL*)	Teichoic acid biosynthesis
LMG6897T	C	D								+					S+
HP	C	D								+					S+
FG2	C	D								+					S+
SK11	C	D						+	+	+					S+
AM2	C	D						+	+	+				+/–	S+
NCD0763	C	D						+	+	+					M+
MG1363	C	D*						+	+	+					M+
N41	C		+					+	+	+/–				+/–	M+
V4	C	D		+				+	+	+			+		M+
KW10	C							+	+	+/–			+		?
B2244B	L		+	+		+	+	+/–	+	+		+	+		
LMG8526	L		+	+	+		+	+	+	+		+	+	+/–	K+/–
Li-1	L		+/–	+			+	+/–	+	+		+	+		I+/–
K231	L		+/–	+			+/–	+/–	+	+		+	+		I+/–
KF7	L		+/–	+				+	+	+		+	+	+/–	K+/–
LMG9449	L		+/–	+		+	+	+/–	+	+		+			I+/–
KF24	L		+/–	+				+/–	+	+		+/–	+		
KF146	L			+	+	+		+	+	+		+	+	+	
KF134	L			+				+	+	+		+	+	+/–	
KF196	L			+	+			+	+	+		+	+	+/–	
KF67	L			+	+			+	+	+		+	+	+/–	
KF201	L			+			+	+	+	+		+			I+/–
E34	L						+/–	+	+	+		+			
K337	L		+/–	+			+/–	+/–	+	+		+	+		I+/–
M20	L		+/–				+	+/–	+	+	+	+			K+/–
LMG8520	H			+				+	+	+/–					
UC317	L	D						+	+	+		+			I+
NCD0895	L	D		+				+	+	+		+	+		I+
ML8	L	D						+	+			+			I+
LMG14418	L	D		+			+	+/–	+	+		+	+/–		I+
N42	L						+	+	+	+		+			I+
IL1403	L	D*						+	+	+	+				I+
DRA4	L	D						+	+	+	+				I+
LMG9446	L			+	+/–			+	+	+		+	+/–		
KF147	L		+	+	+	+	+	+	+	+		+	+	+	K+/–
ATCC19435^T	L	D	+/–				+/–	+/–	+	+		+			K+/–
KF282	L		+	+			+/–	+/–	+	+		+	+		K+/–

Predicted presence of chromosomally encoded gene clusters and their functions in the *L. lactis* strains. L: ssp. *lactis*; C: ssp. *cremoris*; D: dairy; * denotes plasmid-cured strain; + denotes presence of all of the required genes; +/– denotes presence of some of the required genes. Teichoic acid biosynthesis: I = IL1403 type, M = MG1363 type, S = SK11 type, K = KF147 type. Strains P7266 and P7304 were omitted from this analysis.

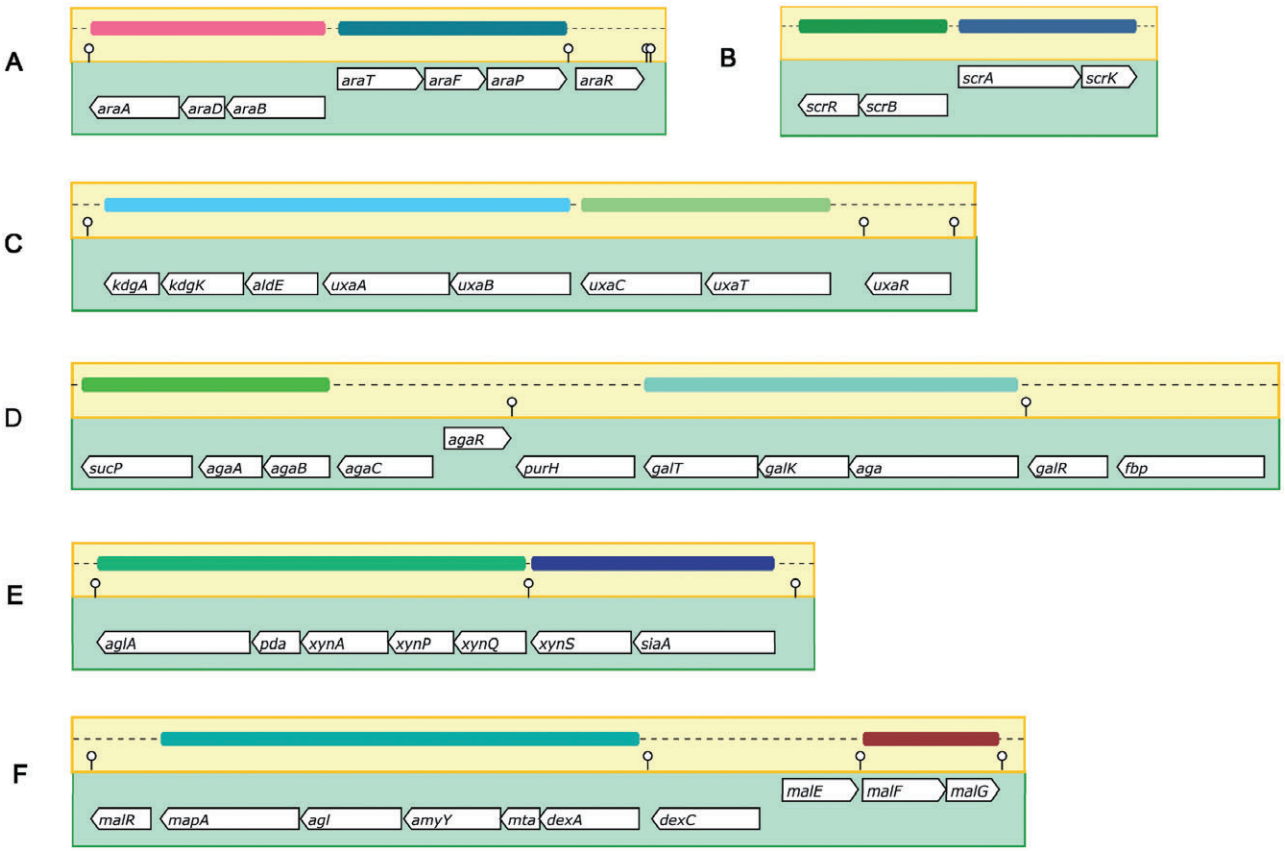

Fig. 3. Variable gene clusters involved in sugar breakdown. As no gene order is known for the query strains, the representative clusters present in the reference genome *L. lactis* KF147 are shown. (A) Arabinose metabolism; (B) sucrose metabolism; (C) galacturonate metabolism; (D) α-galactoside metabolism; (E) xylan breakdown; (F) starch breakdown. Coloured bars indicate operon predictions of two or more genes; stalks indicate predicted terminators. Images made using MINOMICS (Brouwer *et al*., 2009). Gene annotations are in Table S4.

2000). In plant strains, the conjugative element is smaller and lacks the nisin genes. Here, the sucrose gene cluster (Fig. 3B) was found in all plant strains, except N42, M20 and E34. In an earlier study, plant strains KF147 and KF282 were already found to grow on sucrose, in contrast to dairy strains IL1403 and SK11 (Siezen *et al*., 2008). However, three dairy strains do contain the operon: NCD0895, LMG14418 and V4. This suggests that the ability to ferment sucrose is not plant-specific.

• *Galacturonate metabolism.* Previously, the plant strains KF147 and KF282 were shown to grow on glucuronate, which is a building block of the complex sugar xylan, found in plant cells (Siezen *et al*., 2008). All four *L. lactis* strains (KF147, KF282, SK11 and IL1403) described in that study were found to contain a gene cluster for uptake and degradation of D-glucuronate: *kdgR–uxuB–uxuA–uxuT–hypAE–uxaC–kdgK–kdgA*. Only strain KF147 was found to have an additional gene cluster for uptake and degradation of D-galacturonate, a compound that is formed by the epimerization of glucuronate, which is a building block of pectin (Fig. 3C). In the present study, the D-glucuronate cluster was found to be present in all strains, except the *hordniae* strain LMG8520. The additional D-galacturonate cluster described for KF147 was found to be only present in some other plant strains, i.e. KF146, KF196, KF67, LMG8526 and LMG9446. This suggests that these six plant strains are able to metabolize both pectin and xylan, while the rest of the plant strains can only metabolize xylan.

• *α-Galactoside metabolism.* α-Galactosides, such as raffinose, melibiose and stachyose, are oligosaccharides typical for plants. In a previous study comparing four *L. lactis* strains, only plant strain KF147 was found to grow on α-galactosides (Siezen *et al*., 2008). In agreement with this observation, only strain KF147 was then found to possess a gene cluster for α-galactosides uptake, breakdown and subsequent D-galactose conversion: *fbp–galR–aga–galK–galT–purH–agaRCBA–sucP* (Fig. 3D). The present analysis predicts that three other plant strains also contain this gene cluster, i.e. strains KF146, LMG9449 and B2244B. This α-galactoside gene cluster

resides on a 51 kb transposon in strain KF147, which could be conjugally transferred to strain MG1363 (Machielsen *et al*., 2010) and is spontaneously lost upon prolonged growth in milk (Bachmann, 2009). The entire transposon appears to be present in strain B2244B, and parts of the transposon are present in strains LMG9449, KF146, KF67, M20, UC317 and N42.

Complex sugar metabolism

• *Xylan breakdown.* Xylan is the main component of hemicelluloses, which are heteropolymers frequently encountered in plant material. Xylan is composed of D-xylose units, which can be substituted with side groups, such as L-arabinose, D-galactose or acetyl. It is a complex structure, requiring multiple enzymes acting together for breakdown. Xylose is subsequently converted into xylulose-5-phosphate, which can enter the pentose phosphate pathway. Earlier studies revealed the presence of a gene cluster predicted to be involved in xylan breakdown in plant strains KF147 and KF282 (Siezen *et al*., 2008) (Fig. 3E). In the current study this gene cluster was only found to be present in some ssp. *lactis* strains, mostly plant-derived strains but also in two dairy *lactis* strains (Table 6).

• *Starch/maltose breakdown.* A large gene cluster, *malR–mapA–agl–amyY–maa–dexA–dexC–malEFG,* involved in breakdown of starch and its building block maltose is present in all four sequenced reference *L. lactis* strains: IL1403, MG1363, SK11 and KF147 (Fig. 3F). The CGH data predict that the entire cluster is absent only in the *cremoris* strains HP, FG2 and LMG6897T, while the maltose transporter genes *malEFG* are absent in 10 *lactis* strains. The genes for starch breakdown and subsequent uptake and conversion of oligo/monosaccharides are probably lost in these three *cremoris* strains as a consequence of living in a lactose-rich dairy environment.

Amino acid metabolism

• *Glutamate metabolism.* Glutamate decarboxylase activity is one of the phenotypic traits used to distinguish ssp. *cremoris* from ssp. *lactis* strains (Nomura *et al*., 1999; 2000; 2002). CGH analysis indicates that all strains of ssp. *cremoris* and ssp. *lactis* appear to have a large gene cluster for glutamate metabolism, including the genes *gadRCB* and *gltBD*. The glutamate decarboxylase gene *gadB* of *cremoris* strain SK11 is inactive due to a frameshift mutation (Wegmann *et al*., 2007), while the *gadB* gene of *cremoris* strain MG1363 is complete and was shown to be active (Sanders *et al*., 1998). Our CGH analysis can only predict whether genes are present, and not whether they are active or inactive. Therefore we conclude that presence/absence of *gadB* genes or their activity is not suitable to distinguish ssp. *cremoris* from ssp. *lactis*.

• *Arginine metabolism.* Arginine deiminase activity is another phenotypic trait used to distinguish ssp. *cremoris* from ssp. *lactis* strains. Gene clusters *argFBDJC*, *argGH* and *argRS–arcABD1C1C2TD2* for arginine metabolism are predicted by CGH analysis to be present in all analysed *L. lactis* strains. The arginine deiminase gene *arcA* of *cremoris* strain SK11 is inactive due to a frameshift mutation (Wegmann *et al*., 2007), while the *arcA* gene of *cremoris* strain MG1363 is complete and has been shown to be functional (Budin-Verneuil *et al*., 2006). Therefore, as described for the *gadB* genes, the presence/absence of the *arcA* gene does not appear to be a good predictor to distinguish between ssp. *cremoris* and *lactis*.

• *Branched-chain amino acid metabolism.* Degradation products from branched-chain amino acids play a major role in cheese flavour formation (Smit *et al*., 2005). A large cluster *leuABCD–ilvDBHCA* involved in branched-chain amino acid metabolism was found to be absent in dairy *L. lactis* strain ML8, and incomplete in strains LM8520 and N41. Therefore, all three strains are probably incapable of synthesizing branched-chain amino acids. However, auxotrophy in dairy *L. lactis* strains may also be due to simple mutations in these genes, as has been demonstrated for strain IL1403 (Godon *et al*., 1993).

Citrate metabolism. Citrate utilization, with final production of acetoin and diacetyl, is an interesting phenotypic trait for the dairy industry. Diacetyl production is the criterion for naming of the biovar *diacetylactis* strains. The genes required are *citP* for citrate permease (usually plasmid-located; see below) and operon *citMCDEFXG* encoding the enzymes for metabolism of citrate (Garcia-Quintans *et al*., 2008). Indeed, the chromosomal gene cluster was detected only in strains belonging to the biovar *diacetylactis* included in our analysis: IL1403 (plasmid-free derivative of a *diacetylactis* strain), DRA4 and M20. Only strain DRA4 has the plasmid-encoded citrate permease gene *citP* (see below).

Survival/stress response

• *Manganese transport.* Manganese functions in protection against oxidative stress, as has been described for *Bacillus subtilis* (Inaoka *et al*., 1999) and *Lactobacillus plantarum* (Groot *et al*., 2005). Studies with tellurite-resistant *L. lactis* mutants showed that manganese stimulates iron transport and reduces oxidative stress (Turner *et al*., 2007). A manganese ABC-transporter operon *mtsACB* was identified in most strains, except

lactis strain LMG9446 and dairy *cremoris* strains V4, LMG6897T, HP and FG2. The gene cluster shows high sequence similarity to genes in enterococci and streptococci (60–98% amino acid identity). As iron excess is believed to generate oxidative stress, it is possible that these strains are less resistant to oxidative stress because they are unable to transport iron efficiently and consequently have higher intracellular iron levels.

• *Tolerance to high osmolarity. Lactococcus lactis* strains from the ssp. *cremoris* have been described to be more sensitive to osmotic stress than ssp. *lactis* strains. The mechanism of osmo-dependent repression by the glycine/betaine transporter encoded in the *bus* operon in *L. lactis* has been described in a recent study (Romeo *et al*., 2007). Reduced growth of *cremoris* strains at high osmolality has been shown to relate to absence or reduced activity of the *bus* operon (Obis *et al*., 2001). In our CGH analysis, both the *busRAB* operon and a gene cluster encoding a choline transporter (*choQS*) and glutathione reductase (*gshR*) were found to be absent only in *cremoris* strains HP, FG2 and LMG6897T. A high-affinity K+ transport system *kdpDEABC* (two-component regulator and ATPase) is absent in all *cremoris* strains and in the *hordniae* strain, but present in all *lactis* strains except *diacetylactis* strains IL1403 and DRA4 (Table 6). These findings suggest that in particular many of the *cremoris* strains cannot cope well with a high-osmolarity environment, such as high salt concentrations.

• *Non-ribosomal peptide/polyketide synthesis.* Several soil bacteria, such as *Bacillus* and *Streptomyces* species, are known to contain gene clusters involved in non-ribosomal peptide or polyketide biosynthesis (Finking and Marahiel, 2004; Siezen and Khayatt, 2008). Non-ribosomal peptide synthetases (NRPS) and polyketide synthases (PKS) are of great interest, because they produce numerous therapeutic agents and have a great potential for engineering novel compounds. These multi-module proteins are the largest enzymes known. In recent years, NRPS and NRPS/PKS gene clusters have also been identified in the lactic acid bacteria *L. plantarum* WCFS1 (Kleerebezem *et al*., 2003) and *L. lactis* KF147 (Siezen *et al*., 2008; 2010). It was hypothesized that the NRPS/PKS product in *L. lactis* functions in microbe–plant interactions (defence or adhesion) or that it facilitates iron uptake from the environment. Here, the complete NRPS/PKS gene cluster of 13 genes from strain KF147 has been found to be present in five of the *L. lactis* strains, i.e. the plant strains KF147, KF146, KF134, KF196 and Li-1, suggesting that all these plant strains are capable of synthesizing this as yet unknown NRPS/PKS product.

Cell envelope

• *Exopolysaccharide (EPS) biosynthesis.* Bacteria living in plant environments are often found in biofilms, using exopolysaccharides (EPS) to adhere to plants (Danhorn and Fuqua, 2007). As a consequence, genes involved in the physical interaction with the plant cells are expected to be present in the plant-derived *L. lactis* strains. EPS-producing strains are interesting for the dairy industry, as they are used to improve the texture and viscosity of fermented products. Our CGH results show some remarkable variability in chrOG distribution of EPS genes.

A large EPS biosynthesis cluster I of about 25 genes includes *rmlACBD* and *rgpABCDEF* that are responsible for the formation of rhamnose-glucose polysaccharides (Fig. 4A) (Table S4). This EPS gene cluster consists of three separate parts: (i) the first part of seven to eight genes (*rmlA–rgpB*) appears to be present in all ssp. *lactis* and *cremoris* strains, (ii) the second part of seven to eight genes (*rgpC–ycbC*) is present in all *cremoris* strains, but only in *lactis* strains KF7, KF147 and IL1403, while (iii) the third part of nine genes is completely different in the *cremoris* and *lactis* reference strains (see genes and their functions in Table S4). This third set of *cremoris*-like genes appears to be present in all *cremoris* strains and *lactis* strain KF282, while the third set of *lactis* genes, presumably involved in glycerophosphate-containing lipoteichoic acid biosynthesis, is again only present in *lactis* strains KF7, KF147 and IL1403 (Table S4). This variability in the composition of genes in this large EPS cluster suggests that a variety of different EPS structures can be made by *L. lactis* strains.

A second large cluster II for EPS biosynthesis in the plant-derived strain KF147 consists of 13 genes, *epsX-ABCDEFGHIJKL* (Fig. 4B) (Siezen *et al*., 2008; 2010). In the present study, this complete cluster was found to be present only in plant strains KF147 and KF146, while parts of the cluster (usually including the genes *epsXABC*, which possibly encode a basic EPS backbone structure) are present in the plant strains N41, KF134, KF196, KF67, KF7, LMG8526 and B2244B (Table 6). Therefore, this EPS gene cluster and its variants appear to be more specific for plant-derived strains, and could encode biosynthesis of EPS which are beneficial for survival in the plant environment.

This remarkable variability of EPS cluster genes in *L. lactis* confirms other observations on diversity already reported in *Streptococcus thermophilus* (Rasmussen *et al*., 2008), again suggesting a rich variety in structures of the produced EPS in these LAB species.

• *Teichoic acid (TA) biosynthesis.* A teichoic acid (TA) biosynthesis gene cluster *tagL–tagB* is quite variable in the four reference strains (Table 6). The reference

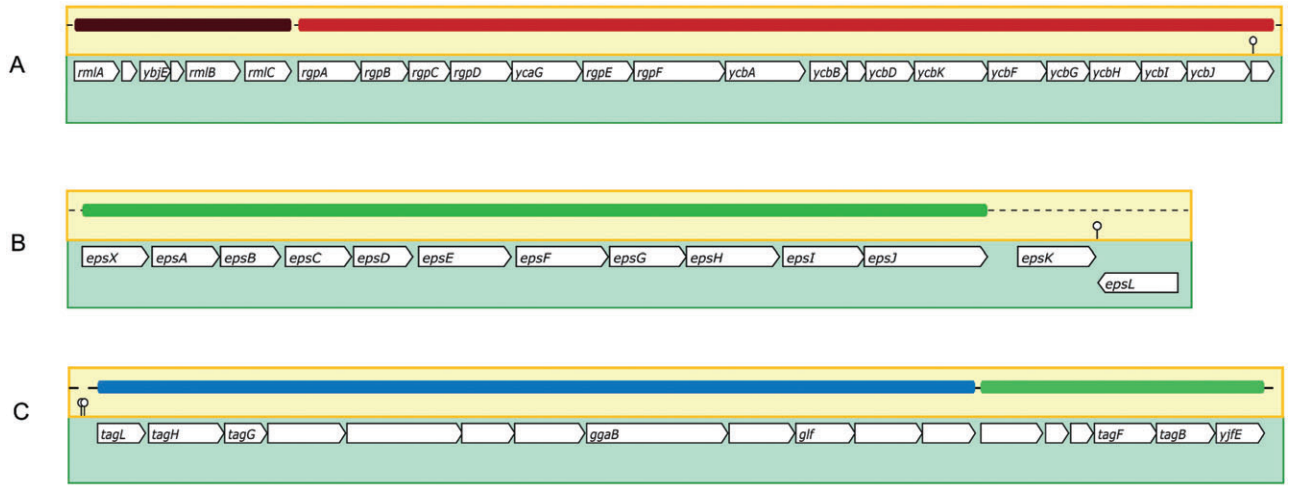

Fig. 4. Variable gene clusters for cell-envelope biosynthesis. As no gene order is known for the query strains, the representative clusters present in the reference genome *L. lactis* KF147 are shown. (A) Exopolysaccharide (EPS) biosynthesis cluster I; (B) exopolysaccharide (EPS) biosynthesis cluster II; (C) teichoic acid biosynthesis cluster. Coloured bars indicate operon predictions of two or more genes; stalks indicate predicted terminators. Images made using MINOMICS (Brouwer *et al.*, 2009). Gene annotations are in Table S4.

cremoris strain MG1363 and *lactis* strain KF147 have the most similar TA cluster, sharing 14 syntenous genes (out of 17 genes in KF147 and 19 in MG1363) (Fig. 4C), while strain IL1403 shares only 7 (out of 15) genes with MG1363 and KF147. In reference strain SK11, all the genes between *tagB* and *tagL* have been replaced by pseudogenes encoding transposases and a putative lipopolysaccharide-1,2-glucosyltransferase. All these types of TA clusters are predicted to be present in the larger set of *L. lactis* strains analysed in this study, with the IL1403-type TA cluster being the most common (Table 6). The variability in the composition of this TA biosynthesis gene cluster suggests that different types of teichoic acids and their derivatives may be made by *L. lactis* strains.

Diversity of plasmid-encoded genes

Dairy strains often contain several plasmids to provide the functions needed to survive and thrive in a milk environment (McKay, 1983; Davidson *et al.*, 1996; Siezen *et al.*, 2005). All known plasmid-located genes of *L. lactis* were represented on the CGH array (Table S5) which allowed us to assess there occurrence and distribution in the *L. lactis* strains analysed in our study. The presence or absence of corresponding genes, rather than OGs, in the 39 *L. lactis* strains was evaluated from the CGH data, and is available in Table S6. In this case, initial clustering into 'plasmid OGs' did not provide any advantage due to the large variability in types of known plasmids and their encoded proteins. Moreover, direct analysis of the much smaller set of plasmid genes was computationally easier, and allowed a direct analysis of their presence/absence in context of functional gene clusters.

Overall, dairy strains appear to contain many known plasmid-encoded functions, while plant strains contain few or none (Table 7). These functions include lactose metabolism (*lacRABCDFEGX* genes), external proteolysis (*prtP, prtM*), copper resistance (*lcoCRS*), cadmium resistance (*cadAC*) and manganese transport (*mntH*). Dairy strains harbouring multiple genes for replication and partitioning presumably contain multiple plasmids encoding these functions (Table 7). Interestingly, strains N41 and N42, of soil and grass origin, appear to have very similar plasmid-encoded functions compared with the dairy strains. Moreover, they both cluster with dairy strains based on chromosome content (Fig. 1), and may therefore originally be from dairy sources.

Several plant-derived *L. lactis* strains also appear to contain plasmids, but the encoded genes could not be predicted because our pan-genome microarray specified probes to many known dairy plasmids, whereas few plasmids from plant isolates have been described and thus were not included on the array. Therefore our present analysis clearly underestimates the plasmid-encoded genes of plant *L. lactis* strains. The presence of genes for EPS biosynthesis in many plant strains does not always correlate with the presence of replication/partitioning genes, so those EPS genes may be chromosomally located (Table 6). Gel electrophoresis confirmed that most dairy strains contained multiple plasmids, while these plant strains contained very few or no plasmids (Fig. 5).

Discussion

The present study supports the view of *L. lactis* as a genomically very flexible species. Different genetic events

Table 7. Diversity of putative plasmid-encoded genes and functions.

Strain	Subspecies	Dairy	Replication/ partitioning	Mobilization/ conjugation	Proteolysis (*prtP, prtM*)	Copper resistance	Cadmium resistance	Manganese transport	Lactose metabolism	Citrate uptake (*citP*)	Glu dehydrogenase	EPS synthesis
LMG6897T	C	D	+	+/–	+			+	+			
HP	C	D	+	+				+	+			
FG2	C	D	+	+	+			+	+			
SK11	C	D	+	+/–	+	+		+	+			
AM2	C	D	+	+/–	+	+		+	+			+
NCD0763	C	D	+		+		+		+			
MG1363	C	D*										
N41	C		+	+	+	+	+	+	+			+/–
V4	C	D							+			
KW10	C		+	+/–	+/–		+/–			+		+/–
B2244B	L											+/–
LMG8526	L						+/–	+				+/–
Li-1	L		+	+/–			+		+		+	
K231	L		+	+/–			+					
KF7	L		+				+					+/–
LMG9449	L		+	+		+/–	+					
KF24	L		+									
KF146	L											+/–
KF134	L											+/–
KF196	L			+								+/–
KF67	L											+/–
KF201	L											
E34	L											
K337	L											
M20	L											
LMG8520	H		+									
UC317	L	D	+	+	+	+	+	+	+			+/–
NCD0895	L	D	+	+			+	+	+			
ML8	L	D	+	+	+		+	+	+			
LMG14418	L	D										
N42	L		+	+	+	+/–	+	+	+			
IL1403	L	D*										
DRA4	L	D	+	+/–			+	+	+	+		+/–
LMG9446	L		+	+/–		+/–	+				+	
KF147	L		+									+/–
ATCC19435^T	L	D	+						+			
KF282	L		+									
P7304	L		+	+/–		+			+/–	+		+/–
P7266	L		+	+/–					+/–	+	+	+/–

Predicted presence of plasmid-encoded genes and their functions in the *L. lactis* strains. L: ssp. *lactis*; C: ssp. *cremoris*; D: dairy; * denotes plasmid-cured strain; + denotes the presence of all or most of the required genes, +/– denotes the presence of some of the required genes. Genes that are known to be both chromosomally and plasmid-encoded are not included in this analysis, e.g. transposases, intergrases/recombinases, restriction/modification system (*hsdM*, *hsdR*, *hsdS*), proteolytic system (*pcp*, *pepO*, *pepF*, *oppACBFD*), cold shock proteins and all plasmid-encoded genes that hybridized with the plasmid-free strains IL1403 or MG1363.

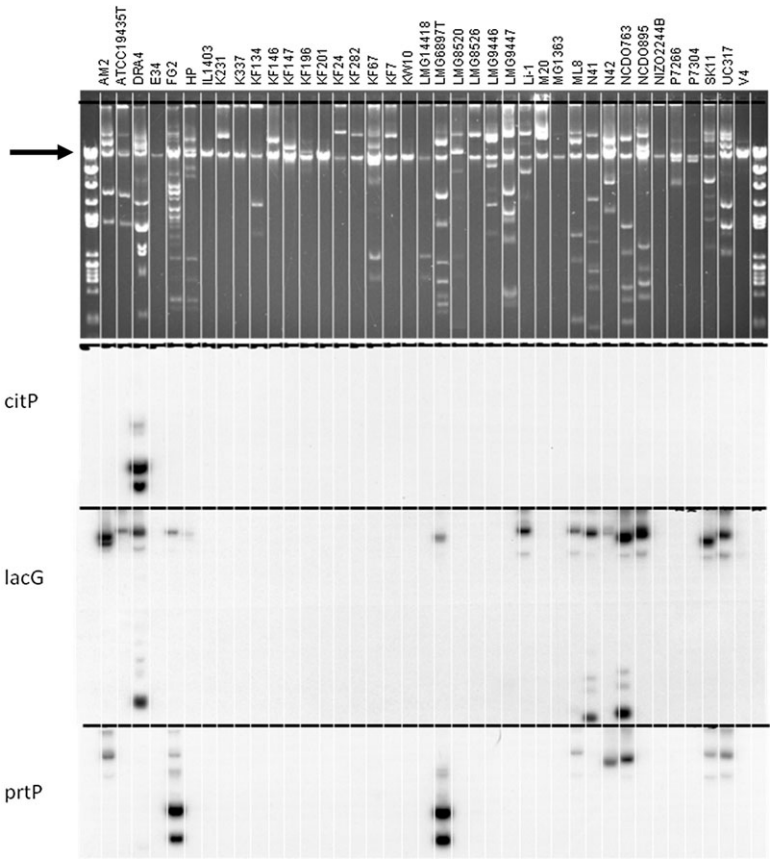

Fig. 5. Polyacrylamide gel electrophoresis of plasmid DNA in *L. lactis* strains. Far left and right lanes contain molecular weight markers. The lower three panels are Southern blots of the same gel as at top, using probes for the *citP, lacG* and *prtP* genes. The arrow indicates an artefact band, present in all lanes, and presumably due to contaminating chromosomal DNA.

– some reversible, some irreversible – influence phenotypes, which are the interactions between the bacterium and the environment it encounters. Genetic transfer has been demonstrated to be possible between strains of the two *L. lactis* subspecies (Rademaker *et al*., 2007) and also with other bacteria (Bolotin *et al*., 2004). Also, literature data on amino acids auxotrophy (e.g. Delorme *et al*., 1993) and on carbohydrate metabolism, e.g. maltose degradation shown in the present study, confirm that auxotrophy is either due to mutations/frameshifts or due to deletions. This further demonstrates the flexibility of *L. lactis* genomes, and their diversification related to niche adaptation. This is important also in the taxonomic perspective (Pace, 2009), as previous work and our study demonstrate that nomenclature based only on phenotype is unreliable. In fact, some phenotypic tests differentiating type strains of *lactis* and *cremoris* are due to severe gene deletions in the *cremoris* type strain and in a few other strains, but due to simple point mutations in other strains (e.g. SK11), which could be reversible. From the current study we conclude that species *lactis* diversity can best be described through a combination of 16S rRNA sequence, genotypic markers and selected phenotypic tests. Therefore, we suggest that nomenclature of this species should be based on genotypic tests, e.g. fingerprinting techniques or specific gene sequence analysis, completed with classical phenotypic tests, to guarantee the continuity with classical taxonomy.

Our data support the theory that the ancestor of the species originally inhabited the plant niche, but was able to successfully colonize other habitats due to its genomic flexibility (Quiberoni *et al*., 2001). The first event in evolution appears to be subspeciation into the *lactis* and *cremoris* subspecies, with no evident differences between gene gain and gene loss, which generated the two subspecies. Adaptation to milk was a more recent event, and therefore appears to have happened independently in the two subspecies. Considering that very few ssp. *cremoris* strains are known outside the dairy environment, speciation and adaptation to milk for this subspecies could have happened at the same time, while adaptation in ssp. *lactis* could be a more recent event. Interestingly, the two sequenced *cremoris* strains, SK11 and MG1363, display genomic inversions (Wegmann *et al*., 2007). Therefore, structural events could have influenced speciation and/or adaptation to milk in this subspecies. Also, mobile elements could have played a crucial role, as witnessed by the plasmid location of genes responsible for lactose degradation and oligopeptide transport in strain SK11.

Our CGH analysis of presence or absence of gene clusters can be used to match phenotypic traits to specific genes or gene clusters, i.e. find correlations between gene content and functional properties. However, gene-trait matching is not straightforward as, for instance, many genes encode proteins of yet unknown function, genes can be inactivated or differentially expressed, and phenotypic test results can often be ambiguous. On the other hand, our extensive data set is an obvious starting point for further research to investigate gene-trait matching in *L. lactis* strains and to move further in the genome annotation procedure. In this sense, the genes need to be seen in their genomic and biological context and, in particular, in the context of cellular metabolic pathways (Teusink *et al*., 2005). Therefore, innovative bioinformatics tools, such as Random Forest methods, are currently being used to investigate gene-trait matching and to evaluate these data in a functional perspective (J. Bayjanov, R.J. Siezen and S.A.F.T. van Hijum, in preparation).

Experimental procedures

Strain selection and DNA preparation

Lactococcus lactis strains were selected from a large set of phenotypically and genotypically characterized strains (Rademaker *et al*., 2007) to represent the diversity of the species in terms of taxonomy and ecology. They belong phenotypically to both subspecies *lactis* (29 strains) and *cremoris* (10 strains) and were isolated from different sources (Table 1). The source, growth conditions and typing of the selected *L. lactis* strains, using 16S rRNA typing and other standard methods and using outgroups such as *L. plantarum* and *Enterococcus casseliflavus*, have been described in detail previously (Rademaker *et al*., 2007). These authors concluded that the two very divergent strains P7304 and P7266 belong to the *L. lactis* species, but that these strains follow a different lineage. DNA was prepared from *L. lactis* strains (Table 1) using the QiaAmp DNA Mini Kit (Qiagen GmbH, Hilden, Germany) according to the manufacturer's protocol for the isolation of genomic DNA from Gram-positive bacteria.

Microarray design, data acquisition and normalization.

All *L. lactis* genomic, plasmid and single gene or operon DNA sequences (1988 sequences present in July 2005, constituting 10.7 Mb) were collected from the NCBI CoreNucleotide database. This included the complete genome sequences of *L. lactis* strain IL1403 (2.35 Mb, Accession No. AE005176) and the incomplete genome of strain SK11 (2.43 Mb, GenBank record GI:62464763). Additionally, draft genome sequences consisting at that time of 547 contigs (2.3 Mb) of *L. lactis* ssp. *lactis* strain KF147 (NIZOB2230) and 961 contigs (2.6 Mb) of *L. lactis* ssp. *lactis* KF282 (B2244W) were added. Redundant stretches of DNA were removed, where a DNA fragment was defined as redundant if it differed from another fragment by at most 2 nucleotides over a window of 100 nucleotides.

For the remaining non-redundant 7 Mb of DNA, on each of the sequences, 32 bp probes were defined with a sliding window of 19 nucleotides, resulting in a total of 386 298 probes. We also designed 3181 random probes with their sequence absent in the non-redundant 7 Mb of *L. lactis* DNA, and these were randomly located on the array. Details of array production, DNA hybridization (NimbleGen Systems, Madison, WI, USA), data normalization and data submission to GEO are described in Bayjanov and colleagues (2009). Briefly, array normalization was performed using the fields package (Fields Development Team; http://www.image.ucar.edu/Software/Fields/) using the statistical programming language R (R Development Core Team, 2006). Description of the array platform with probe information and hybridization data of 39 *L. lactis* strains have been deposited in the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo) with the Accession No. GPL7231.

The annotations (gene definitions and putative protein function descriptions) were extracted from the GenBank files for publicly available sequences; for the draft sequences Glimmer (Salzberg *et al*., 1998) and InterProScan (Zdobnov and Apweiler, 2001) were used. For selected genes the annotation was improved using the ERGO Bioinformatics Suite (Overbeek *et al*., 2003).

Defining orthologous groups of genes (OGs)

During the course of our work, the complete sequences of *L. lactis* ssp. *cremoris* strains SK11, MG1363 and KF147 were published (Makarova *et al*., 2006; Wegmann *et al*., 2007; Siezen *et al*., 2010), and we re-mapped the microarray probes to the annotated genes in these genomes. In order to predict orthology among genes, the chromosome sequence of the four fully sequenced public *L. lactis* strains (ssp. *lactis* IL1403, ssp. *lactis* KF147, ssp. *cremoris* SK11, ssp. *cremoris* MG1363) were used. The orthology prediction program InParanoid (Remm *et al*., 2001) was run to find orthologous genes among these genomes. InParanoid's default minimum bit score value of 50 and a minimum identity value of 80 were used for grouping genes into OGs. All possible pairwise comparisons between the genes of the four chromosomes were performed and iteratively combined to groups of chromosomal orthologous genes (chrOGs). In cases where inconsistencies were found between the InParanoid predictions (i.e. homologous genes from the four reference genomes were not all bidirectional best hits to each other), genes were regarded as not being orthologous and each treated as single genes in an orthologous group of size 1. The genes from plasmids were not categorized into OGs, but were studied as single genes (828 genes).

We compared our chrOGs with the complete annotated list of LaCOGs available at ftp://ftp.ncbi.nih.gov/pub/wolf/lacto (file LaCOGS_table.xls) (Makarova *et al*., 2006).

Determination of gene conservation in the strains

A novel genotype-calling algorithm PanCGH was developed to determine the presence/absence of orthologous groups

of genes in strains with unknown genome sequence (Bayjanov *et al*., 2009; 2010). Briefly, a threshold score of 5.5 was defined based on presence/absence of orthologous groups in the four sequenced strains. This score was then used in the genotype-calling algorithm applied to normalized hybridization signals of DNA from query strains. Thus, presence/absence of genes was determined on the basis of signal intensities and orthologue distribution. Applying the PanCGH algorithm to the CGH data results in a binary matrix, in which the rows represent the chrOGs and the columns the different strains. For each strain, a '1' denotes the presence of an orthologue in the strain and '0' denotes the absence of an orthologue. 'NA' signifies that presence or absence of an orthologue in a strain could not be estimated from the data due to too few valid probe signals of the chrOG members. The PanCGH algorithm assumes a minimum of 10 aligned probes, and hence CGH signal data for 622 chrOGs were not considered, as these genes were represented by less than 10 probes on the array. The hybridization results for these chrOGs were excluded from further data analysis.

Presence or absence of plasmid-encoded genes was analysed separately. Probes for all published plasmids of *L. lactis* (Table S5) were also present on the array. PanCGH was used to predict presence/absence in query strains of the known plasmid-encoded genes from their hybridization signals. Genes that are known to be plasmid- and chromosome-encoded were not included in this analysis of putative plasmid genes, e.g. genes encoding transposases, intergrases/recombinases, restriction/modification (R/M) system (*hsdM, hsdR, hsdS*), proteolytic system (*pcp, pepO, pepF, oppACBFD*), cold shock proteins and all plasmid-encoded genes that hybridized with the plasmid-free strains IL1403 or MG1363.

Hierarchical clustering of strains

To study the evolutionary relatedness and differences in genes and gene clusters that could have contributed to *L. lactis* strain diversification, a hierarchical clustering was performed by comparing the presence/absence profiles of chrOGs of the different strains to each other. Of the original 3877 chrOGs, the 622 chrOGs containing 'NA' values were omitted from this clustering (see above). A tree was constructed using the statistical programming language R, with the average linkage clustering method based on the binary distance metric.

Determining gene clusters contributing to strain diversification

By combining both the tree plot and the presence/absence profiles ('NA' values were again omitted), genes were identified that might be important for the diversification of the strains. Since plasmid genes are frequently exchanged between bacteria, these genes were not considered in this analysis. A Perl-script was developed that identifies features (chrOGs) that cause a clear separation between branches in a tree, encoded in the Newick format. The script parses the tree according to the depth-first search principle, in which the tree is traversed from the root to each leaf. At each split in the tree the presence/absence patterns of the strains in the two branches are evaluated. For each chrOG the fraction of presence in the two sub-branches is calculated and only those chrOGs with a difference in presence of more than 70% are selected. This allows identification of chrOGs that are (almost) fully absent in one branch and (almost) fully present in the other. From this analysis a list of chrOGs that are important for each split in the tree was obtained. This list was used to identify gene clusters in the strains, which were projected on the chromosomes of the four reference genomes: MG1363, IL1403, SK11 and KF147. Gene clusters can be (parts of) operons or functional groups of genes, involved in a certain trait. Per split in the tree, the genes of the reference genomes constituting a chrOG were retrieved. For these genes the locations in the respective genome were retrieved and groups of adjacent genes were identified. Furthermore, an operon prediction was performed for the chromosomes of the four reference strains using the Operon web-tool of the Molecular Genetics group of the University of Groningen (http://bioinformatics.biol.rug.nl/websoftware/operon/). The default settings were used for the predictions (maximum spacing between ORFs of 100 bp and maximum energy/deltaG of 0).

Identifying subspecies-specific or niche-specific OGs

Strains were divided into two categories according to their subspecies or niche assignment. We used a hypergeometric test in order to find OGs that are mostly present in one category of strains (e.g. in ssp. *lactis* strains) but almost absent in all strains of the other category (e.g. ssp. *cremoris* strains). The resulting *P*-values were corrected for false discovery rate and only OGs that have a *P*-value below 0.05 were considered to be specific.

Plasmid gel electrophoresis

Isolation of plasmid DNA was performed as previously described (de Vos *et al*., 1989). Standard SDS-polyacrylamide gel electrophoresis was performed as described by Sambrook and colleagues (1989). Southern hybridization was performed using probes designed to detect the typical plasmid-located genes *citP* (encoding citrate permease for citrate uptake), *lacG* (encoding 6-P-β-galactosidase carried on the lactose plasmid) and *prtP* (encoding cell-wall proteinase).

Acknowledgements

We thank Ingrid van Alen-Boerrigter for experimental support. This work was supported by a BSIK grant through the Netherlands Genomics Initiative (NGI) and was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC). Additional funding was obtained from NGI as part of the Kluyver Centre for Genomics of Industrial Fermentation.

References

Bachmann, H. (2009) *Regulatory and Adaptive Responses of* Lactococcus lactis *in Situ*. Wageningen, the Netherlands: Wageningen University.

Bayjanov, J.R., Wels, M., Starrenburg, M., van Hylckama Vlieg, J.E., Siezen, R.J., and Molenaar, D. (2009) PanCGH: a genotype-calling algorithm for pangenome CGH data. *Bioinformatics* **25:** 309–314.

Bayjanov, J.R., Siezen, R.J., and van Hijum, S.A. (2010) PanCGHweb: a web tool for genotype calling in pangenome CGH data. *Bioinformatics* **26:** 1256–1257.

Bolotin, A., Wincker, P., Mauger, S., Jaillon, O., Malarme, K., Weissenbach, J., *et al*. (2001) The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403. *Genome Res* **11:** 731–753.

Bolotin, A., Quinquis, B., Sorokin, A., and Ehrlich, D.S. (2004) Recent genetic transfer between *Lactococcus lactis* and enterobacteria. *J Bacteriol* **186:** 6671–6677.

Brouwer, R.W., van Hijum, S.A., and Kuipers, O.P. (2009) MINOMICS: visualizing prokaryote transcriptomics and proteomics data in a genomic context. *Bioinformatics* **25:** 139–140.

Budin-Verneuil, A., Maguin, E., Auffray, Y., Ehrlich, D.S., and Pichereau, V. (2006) Genetic structure and transcriptional analysis of the arginine deiminase (ADI) cluster in *Lactococcus lactis* MG1363. *Can J Microbiol* **52:** 617–622.

Campo, N., Dias, M.J., Daveran-Mingot, M.L., Ritzenthaler, P., and Le Bourgeois, P. (2002) Genome plasticity in *Lactococcus lactis*. *Antonie Van Leeuwenhoek* **82:** 123–132.

Danhorn, T., and Fuqua, C. (2007) Biofilm formation by plant-associated bacteria. *Annu Rev Microbiol* **61:** 401–422.

Davidson, B.E., Kordias, N., Dobos, M., and Hillier, A.J. (1996) Genomic organization of lactic acid bacteria. *Antonie Van Leeuwenhoek* **70:** 161–183.

Delorme, C., Godon, J.J., Ehrlich, S.D., and Renault, P. (1993) Gene inactivation in *Lactococcus lactis*: histidine biosynthesis. *J Bacteriol* **175:** 4391–4399.

Earl, A.M., Losick, R., and Kolter, R. (2007) *Bacillus subtilis* genome diversity. *J Bacteriol* **189:** 1163–1170.

Finking, R., and Marahiel, M.A. (2004) Biosynthesis of nonribosomal peptides. *Annu Rev Microbiol* **58:** 453–488.

Garcia-Quintans, N., Repizo, G., Martin, M., Magni, C., and Lopez, P. (2008) Activation of the diacetyl/acetoin pathway in *Lactococcus lactis* subsp. *lactis* bv. *diacetylactis* CRL264 by acidic growth. *Appl Environ Microbiol* **74:** 1988–1996.

Godon, J.J., Delorme, C., Bardowski, J., Chopin, M.C., Ehrlich, S.D., and Renault, P. (1993) Gene inactivation in *Lactococcus lactis*: branched-chain amino acid biosynthesis. *J Bacteriol* **175:** 4383–4390.

Groot, M.N., Klaassens, E., de Vos, W.M., Delcour, J., Hols, P., and Kleerebezem, M. (2005) Genome-based in silico detection of putative manganese transport systems in *Lactobacillus plantarum* and their genetic analysis. *Microbiology* **151:** 1229–1238.

Han, Y.H., Liu, W.Z., Shi, Y.Z., Lu, L.Q., Xiao, S., Zhang, Q.H., and Zhao, G.P. (2007) Comparative genomics profiling of clinical isolates of *Helicobacter pylori* in Chinese populations using DNA microarray. *J Microbiol* **45:** 21–28.

van Hylckama Vlieg, J.E., Rademaker, J.L., Bachmann, H., Molenaar, D., Kelly, W.J., and Siezen, R.J. (2006) Natural diversity and adaptive responses of *Lactococcus lactis*. *Curr Opin Biotechnol* **17:** 183–190.

Inaoka, T., Matsumura, Y., and Tsuchido, T. (1999) SodA and manganese are essential for resistance to oxidative stress in growing and sporulating cells of *Bacillus subtilis*. *J Bacteriol* **181:** 1939–1943.

Kelly, W., and Ward, L. (2002) Genotypic vs. phenotypic biodiversity in *Lactococcus lactis*. *Microbiology* **148:** 3332–3333.

Kelly, W.J., Davey, G.P., and Ward, L.J. (2000) Novel sucrose transposons from plant strains of *Lactococcus lactis*. *FEMS Microbiol Lett* **190:** 237–240.

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

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

La, M.V., Crapoulet, N., Barbry, P., Raoult, D., and Renesto, P. (2007) Comparative genomic analysis of *Tropheryma whipplei* strains reveals that diversity among clinical isolates is mainly related to the WiSP proteins. *BMC Genomics* **8:** 349.

Lan, R., and Reeves, P.R. (2000) Intraspecies variation in bacterial genomes: the need for a species genome concept. *Trends Microbiol* **8:** 396–401.

McBride, S.M., Fischetti, V.A., Leblanc, D.J., Moellering, R.C., Jr, and Gilmore, M.S. (2007) Genetic diversity among *Enterococcus faecalis*. *PLoS ONE* **2:** e582.

Machielsen, R., Siezen, R.J., van Hijum, S.A., and van Hylckama Vlieg, J.E. (2010) Molecular description and industrial potential of Tn*6098*-mediated conjugative transfer of alpha-galactosides metabolism in *Lactococcus lactis*. *Appl Environ Microbiol* **77:** 555–563.

McKay, L.L. (1983) Functional properties of plasmids in lactic streptococci. *Antonie Van Leeuwenhoek* **49:** 259–274.

Makarova, K.S., and Koonin, E.V. (2007) Evolutionary genomics of lactic acid bacteria. *J Bacteriol* **189:** 1199–1208.

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

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005) The microbial pan-genome. *Curr Opin Genet Dev* **15:** 589–594.

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

Nomura, M., Nakajima, I., Fujita, Y., Kobayashi, M., Kimoto, H., Suzuki, I., and Aso, H. (1999) *Lactococcus lactis* contains only one glutamate decarboxylase gene. *Microbiology* **145** (Part 6): 1375–1380.

Nomura, M., Kobayashi, M., Ohmomo, S., and Okamoto, T. (2000) Inactivation of the glutamate decarboxylase gene in *Lactococcus lactis* subsp. *cremoris*. *Appl Environ Microbiol* **66:** 2235–2237.

Nomura, M., Kobayashi, M., and Okamoto, T. (2002) Rapid PCR-based method which can determine both phenotype and genotype of *Lactococcus lactis* subspecies. *Appl Environ Microbiol* **68:** 2209–2213.

Obis, D., Guillot, A., and Mistou, M.Y. (2001) Tolerance to high osmolality of *Lactococcus lactis* subsp. *lactis* and *cremoris* is related to the activity of a betaine transport system. *FEMS Microbiol Lett* **202:** 39–44.

Overbeek, R., Larsen, N., Walunas, T., D'Souza, M., Pusch, G., Selkov, E., Jr, *et al*. (2003) The ERGO genome analysis and discovery system. *Nucleic Acids Res* **31:** 164–171.

Pace, N.R. (2009) Mapping the tree of life: progress and prospects. *Microbiol Mol Biol Rev* **73:** 565–576.

Peng, J., Zhang, X., Yang, J., Wang, J., Yang, E., Bin, W., *et al*. (2006) The use of comparative genomic hybridization to characterize genome dynamics and diversity among the serotypes of Shigella. *BMC Genomics* **7:** 218.

Quiberoni, A., Rezaiki, L., El Karoui, M., Biswas, I., Tailliez, P., and Gruss, A. (2001) Distinctive features of homologous recombination in an 'old' microorganism, *Lactococcus lactis*. *Res Microbiol* **152:** 131–139.

R Development Core Team (2006) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing [WWW document]. URL http://www.R-project.org.

Rademaker, J.L., Herbet, H., Starrenburg, M.J., Naser, S.M., Gevers, D., Kelly, W.J., *et al*. (2007) Diversity analysis of dairy and nondairy *Lactococcus lactis* isolates, using a novel multilocus sequence analysis scheme and (GTG)5-PCR fingerprinting. *Appl Environ Microbiol* **73:** 7128–7137.

Rasmussen, T.B., Danielsen, M., Valina, O., Garrigues, C., Johansen, E., and Pedersen, M.B. (2008) *Streptococcus thermophilus* core genome: comparative genome hybridization study of 47 strains. *Appl Environ Microbiol* **74:** 4703–4710.

Remm, M., Storm, C.E., and Sonnhammer, E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314:** 1041–1052.

Romeo, Y., Bouvier, J., and Gutierrez, C. (2007) Osmotic regulation of transcription in *Lactococcus lactis*: ionic strength-dependent binding of the BusR repressor to the *busA* promoter. *FEBS Lett* **581:** 3387–3390.

Salzberg, S.L., Delcher, A.L., Kasif, S., and White, O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* **26:** 544–548.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*. New York, USA: Cold Spring Harbor Laboratory Press.

Sanders, J.W., Leenhouts, K., Burghoorn, J., Brands, J.R., Venema, G., and Kok, J. (1998) A chloride-inducible acid resistance mechanism in *Lactococcus lactis* and its regulation. *Mol Microbiol* **27:** 299–310.

Sandine, W.E. (1972) Ecology of lactic streptococci. A review. *J Milk Food Technol* **35:** 179–206.

Siezen, R.J., and Khayatt, B. (2008) Natural products genomics. *Microb Biotechnol* **1:** 275–282.

Siezen, R.J., Renckens, B., van Swam, I., Peters, S., van Kranenburg, R., Kleerebezem, M., and de Vos, W.M. (2005) Complete sequences of four plasmids of *Lactococcus lactis* subsp. *cremoris* SK11 reveal extensive adaptation to the dairy environment. *Appl Environ Microbiol* **71:** 8371–8382.

Siezen, R.J., Starrenburg, M.J., Boekhorst, J., Renckens, B., Molenaar, D., and van Hylckama Vlieg, J.E. (2008) Genome-scale genotype-phenotype matching of two *Lactococcus lactis* isolates from plants identifies mechanisms of adaptation to the plant niche. *Appl Environ Microbiol* **74:** 424–436.

Siezen, R.J., Bayjanov, J., Renckens, B., Wels, M., van Hijum, S.A., Molenaar, D., and van Hylckama Vlieg, J.E. (2010) Complete genome sequence of *Lactococcus lactis* subsp. *lactis* KF147, a plant-associated lactic acid bacterium. *J Bacteriol* **192:** 2649–2650.

Smit, G., Smit, B.A., and Engels, W.J. (2005) Flavour formation by lactic acid bacteria and biochemical flavour profiling of cheese products. *FEMS Microbiol Rev* **29:** 591–610.

Taibi, A., Dabour, N., Lamoureux, M., Roy, D., and Lapointe, G. (2010) Evaluation of the genetic polymorphism among *Lactococcus lactis* subsp. *cremoris* strains using comparative genomic hybridization and multilocus sequence analysis. *Int J Food Microbiol* **144:** 20–28.

Tailliez, P., Tremblay, J., Ehrlich, S.D., and Chopin, A. (1998) Molecular diversity and relationship within *Lactococcus lactis*, as revealed by randomly amplified polymorphic DNA (RAPD). *Syst Appl Microbiol* **21:** 530–538.

Teusink, B., van Enckevort, F.H.J., Francke, C., Wiersma, A., Wegkamp, A., Smid, E.J., and Siezen, R.J. (2005) *In silico* reconstruction of the metabolic pathways of *Lactobacillus plantarum*: comparing predictions of nutrient requirements with those from growth experiments. *Appl Environ Microbiol* **71:** 7253–7262.

Turner, M.S., Tan, Y.P., and Giffard, P.M. (2007) Inactivation of an iron transporter in *Lactococcus lactis* results in resistance to tellurite and oxidative stress. *Appl Environ Microbiol* **73:** 6144–6149.

de Vos, W.M., Vos, P., de Haard, H., and Boerrigter, I. (1989) Cloning and expression of the *Lactococcus lactis* subsp. *cremoris* SK11 gene encoding an extracellular serine proteinase. *Gene* **85:** 169–176.

Wang, X., Han, Y., Li, Y., Guo, Z., Song, Y., Tan, Y., *et al*. (2007) *Yersinia* genome diversity disclosed by *Yersinia pestis* genome-wide DNA microarray. *Can J Microbiol* **53:** 1211–1221.

Wegmann, U., O'Connell-Motherway, M., Zomer, A., Buist, G., Shearman, C., Canchaya, C., *et al*. (2007) Complete genome sequence of the prototype lactic acid bacterium *Lactococcus lactis* subsp. *cremoris* MG1363. *J Bacteriol* **189:** 3256–3270.

Zdobnov, E.M., and Apweiler, R. (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17:** 847–848.

Zhou, M., Theunissen, D., Wels, M., and Siezen, R.J. (2010) LAB-Secretome: a genome-scale comparative analysis of the predicted extracellular and surface-associated proteins of Lactic Acid Bacteria. *BMC Genomics* **11:** 651.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Table S1. List of core chromosomal OGs (chrOGs) of the *L. lactis* genome. The list includes the genes in these chrOGs in the four reference genomes. The three different worksheets are sorted according to (i) our OG number, (ii) LaCOG

number and (iii) functional annotation of the KF147 gene products.

Table S2. List of *L. lactis*-specific core chrOGs and genes.

Table S3. List of subspecies-specific and niche-specific chromosomal OGs.

Table S4. Annotation of variable gene clusters and genes.

Table S5. Known *L. lactis* plasmids represented on the CGH microarray.

Table S6. Predicted presence/absence of important known plasmid-located genes in 39 *L. lactis* strains.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

