Research article

Lactobacillus plantarum gene clusters encoding putative cell-surface protein complexes for carbohydrate utilization are conserved in specific gram-positive bacteria

Roland Siezen*[1,2,3], Jos Boekhorst[2], Lidia Muscariello[1,3], Douwe Molenaar[1,3], Bernadet Renckens[2,3] and Michiel Kleerebezem[1,3]

Address: [1]Wageningen Centre for Food Sciences (WCFS), Wageningen, The Netherlands, [2]Centre for Molecular and Biomolecular Informatics (CMBI), Radboud University Nijmegen, The Netherlands and [3]NIZO food research, Ede, The Netherlands

Email: Roland Siezen* - r.siezen@cmbi.ru.nl; Jos Boekhorst - j.boekhorst@cmbi.ru.nl; Lidia Muscariello - muscar@igb.cnr.it; Douwe Molenaar - douwe.molenaar@nizo.nl; Bernadet Renckens - b.renckens@cmbi.ru.nl; Michiel Kleerebezem - michiel.kleerebezem@nizo.nl

* Corresponding author

Abstract

Background: Genomes of gram-positive bacteria encode many putative cell-surface proteins, of which the majority has no known function. From the rapidly increasing number of available genome sequences it has become apparent that many cell-surface proteins are conserved, and frequently encoded in gene clusters or operons, suggesting common functions, and interactions of multiple components.

Results: A novel gene cluster encoding exclusively cell-surface proteins was identified, which is conserved in a subgroup of gram-positive bacteria. Each gene cluster generally has one copy of four new gene families called *cscA, cscB, cscC* and *cscD*. Clusters encoding these cell-surface proteins were found only in complete genomes of *Lactobacillus plantarum*, *Lactobacillus sakei*, *Enterococcus faecalis*, *Listeria innocua*, *Listeria monocytogenes*, *Lactococcus lactis ssp lactis* and *Bacillus cereus* and in incomplete genomes of *L. lactis ssp cremoris*, *Lactobacillus casei*, *Enterococcus faecium*, *Pediococcus pentosaceus*, *Lactobacillius brevis*, *Oenococcus oeni*, *Leuconostoc mesenteroides*, and *Bacillus thuringiensis*. These genes are neither present in the genomes of streptococci, staphylococci and clostridia, nor in the *Lactobacillus acidophilus* group, suggesting a niche-specific distribution, possibly relating to association with plants. All encoded proteins have a signal peptide for secretion by the Sec-dependent pathway, while some have cell-surface anchors, novel WxL domains, and putative domains for sugar binding and degradation. Transcriptome analysis in *L. plantarum* shows that the *cscA-D* genes are co-expressed, supporting their operon organization. Many gene clusters are significantly up-regulated in a glucose-grown, *ccpA*-mutant derivative of *L. plantarum*, suggesting catabolite control. This is supported by the presence of predicted CRE-sites upstream or inside the up-regulated *cscA-D* gene clusters.

Conclusion: We propose that the CscA, CscB, CscC and CscD proteins form cell-surface protein complexes and play a role in carbon source acquisition. Primary occurrence in plant-associated gram-positive bacteria suggests a possible role in degradation and utilization of plant oligo- or poly-saccharides.

Background

Most Gram-positive bacteria are known to produce a multiplicity of extracellular proteins, many of which are destined to become attached to the cell surface [1-5]. These surface-exposed proteins serve to communicate and interact with the environment. Particularly in pathogenic streptococci, staphylococci and *Listeria*, they are often of primary importance in bacterial adhesion, invasion and interaction with host cells [6-8]. Cell-surface proteins are also known to play an essential role in providing nutrition to the cell through binding, degradation and uptake of carbon and nitrogen substrates. Many cell-surface proteins have a multi-domain architecture, and share various structural features including secretion signal peptides, cell-anchoring domains or motifs, cell-wall spanning regions, and repeated domains of various functions. In some cases, multiple proteins join forces to form large extracellular complexes that provide both binding and enzymatic functionalities, such as the cellulosomes of anaerobic bacteria (e.g. *Clostridium*, *Ruminococcus*) for degradation of and growth on cellulose, the main structural component of plant cell walls [9-13].

Even though the function of a variety of extracellular proteins of Gram-positive bacteria has been characterized experimentally, recent genome sequencing efforts have led to the prediction of hundreds of encoded extracellular proteins of unknown function. Many of these appear to belong to conserved homologous families of hypothetical extracellular proteins, suggesting common functions in different bacterial species. While it is often possible to detect known cell-anchoring domains in these proteins, such as (i) amino- or carboxy-terminal membrane-spanning anchors, (ii) peptidoglycan anchors covalently bound through their LPxTG motif [4,14-18], (iii) amino-terminal lipid-bound anchors [19], and (iv) a variety of domains binding non-covalently to peptidoglycan, teichoic acids [20] or surface polysaccharides, the main function(s) of these encoded cell-surface proteins in their interaction with the environment remains elusive.

Lactobacillus plantarum is a gram-positive bacterium that is encountered in many different environmental niches, as it is associated with various plants [21-24], it occurs in several food and feed fermentations [25-28], and it is a natural inhabitant of the gastrointestinal tract of humans and animals [29,30]. Analysis of the 3.3 Mbp genome sequence of *L. plantarum* WCFS1 revealed over 200 putative extracellular proteins based on the presence of an N-terminal signal peptide [31]. The vast majority of these proteins contained at least one of the cell-anchoring motifs described above. A new C-terminal domain designated WxL was found in 19 proteins of *L. plantarum*. More recently, fifteen proteins with a WxL-like domain were identified in the genome of *Lactobacillus sakei* 23 K [32],

and found to be encoded in gene clusters that potentially encode a multicomponent complex of unknown function on the bacterial surface. In search of putative functions for the encoded hypothetical extracellular proteins, and their possible relation to niche adaptation, we have now discovered that 35 of the cell-surface proteins of *L. plantarum* are encoded in nine paralogous gene clusters. Four different types of novel protein families are represented in these gene clusters. We present bioinformatics and experimental evidence that the encoded proteins are functionally coupled and possibly form a cell-surface protein complex that could play a role in sugar metabolism. A genome-wide search revealed similar gene clusters in a specific subgroup of mainly plant-associated Gram-positive bacteria, and we therefore postulate a role in degradation of (complex) plant polysaccharides.

Results

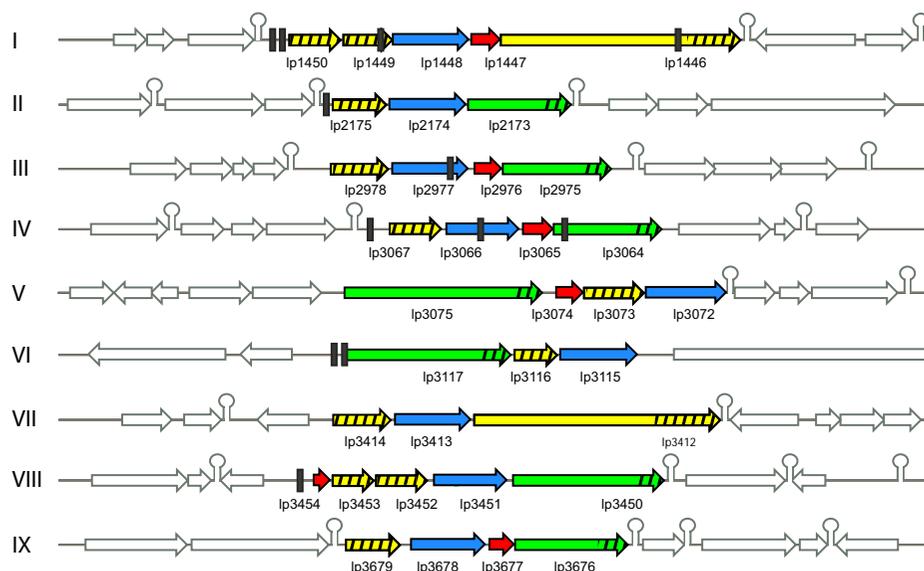## *Cell-surface clusters in Lactobacillus plantarum WCFS1*

Analysis of the chromosome indicated that many of the predicted extracellular proteins are encoded in clusters of 3–6 genes [31]. A closer inspection reveals that nine clusters encode proteins which can be divided into 4 different classes or families based on amino acid sequence similarity, domain and motif characteristics (Table 1; Fig. 1; see details in additional files 1, 2). All of the 35 encoded Csc proteins (cell-surface complex) have normal signal peptides for secretion via the Sec-dependent pathway [33] and processing by the signal peptidase I. Most of the Csc proteins and their domains are of unknown function since they do not have significant similarity to proteins of known function (see below for details). The four families can be easily distinguished based on domain composition. The CscA proteins are all predicted to contain a conserved domain of unknown function (PFAM: DUF916) as well as a C-terminal transmembrane anchor. CscB and CscC proteins are characterized by a novel domain of 160–190 residues, which we have termed WxL since it contains two characteristic conserved sequence motifs containing the WxL signature (Fig. 1)[31]. The CscB proteins are on average 240 amino acids in size and consist almost entirely of the WxL domain, while the CscC proteins are much larger with an average size of 800 amino acids and have a variable N-terminus. Since the WxL domains of the CscB and CscC proteins can be distinguished based on sequence characteristics such as the distance between the conserved WxL residues, they were considered as two different families (WxL1 for the CscB proteins, WxL2 for the CscC proteins). Finally, members of the CscD family all have a C-terminal LPxTG-type motif for sortase-mediated covalent anchoring to the peptidoglycan layer [4,14], and are uncharacteristically small for LPxTG-anchored proteins. Figure 2 summarizes the characteristics of the four Csc family members. The individual families will be discussed in more detail below.

Table 1: Occurrence of cell-surface clusters and genes in genomes

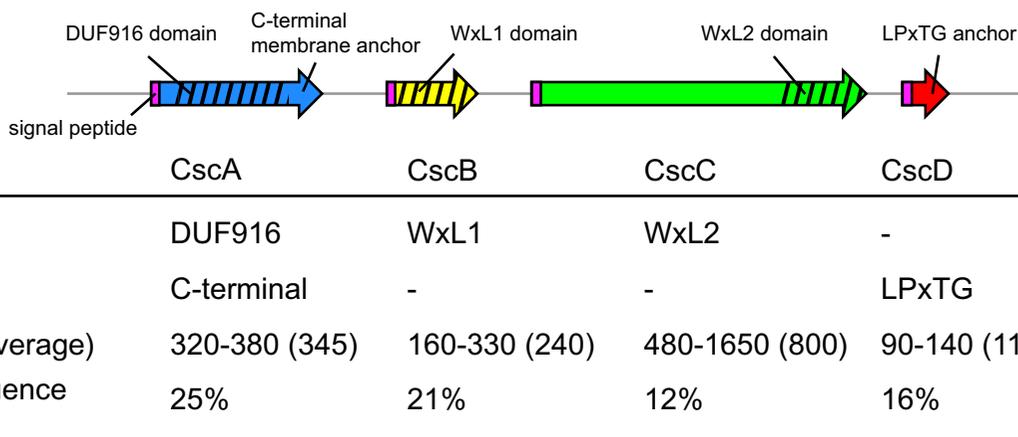| Species	clusters	genes#			
		cscA	*cscB*	*cscC*	*cscD*
Complete genome sequences					
Lactobacillus plantarum WCFS1	9	9	13	8	6
Lactobacillus sakei 23K	8	10*	8	6*	2
Enterococcus faecalis V583	6	7	17	4	5
Lactococcus lactis IL1403	3	4 *	6	2	3
Listeria innocua Clip11262	3	4	3	3	2
Listeria monocytogenes EGD-e	2	2	2	2	1
Bacillus cereus ZK	1	1	2	1	1
Bacillus anthracis A2012 *(plasmid)*	1	1*	1	1	0
Incomplete genome sequences					
Lactococcus cremoris SK11	5	5 *	7	3	3
Lactobacillus casei ATCC367	4	5*	5*	3*	3
Enterococcus faecium DO	3	4	3 *	3	1
Pediococcus pentosaceus ATCC25745	2	2	2	1	1
Leuconostoc mesenteroides ATCC8293	1	1	1	1	1
Oenococcus oeni PSU-1	1	1 *	1	0	0
Bacillus thuringiensis ATCC35646	1	1	2	1	1
Lactobacillus brevis ATCC367	?	0	1	6	0
Total	50	57	74	45	30

Seven of these *csc* genes are found outside the gene clusters in complete genomes (details in additional file 1).

* Some genes contain frame shifts, stop codons or truncations (details in additional file 2).

Figure 1
***Csc* gene clusters found in *Lactobacillus plantarum* WCFS1**. Genes are color-coded according to family: *cscA* (blue), *cscB* (yellow), *cscC* (green), *cscD* (red); other genes are not coloured. Positions of encoded WxL1 domains (in CscB) and WxL2 domains (in CscC) are striped. Predicted CRE sites are indicated by black vertical bars (see also Table 3). Predicted terminators are indicated by loop symbols.

Cell-surface clusters in other bacteria

	CscA	CscB	CscC	CscD
domain	DUF916	WxL1	WxL2	-
anchor	C-terminal	-	-	LPxTG
size range (average)	320-380 (345)	160-330 (240)	480-1650 (800)	90-140 (115)
average sequence identity	25%	21%	12%	16%

Figure 2
Schematic summary of the characteristics of Csc families. The summary is based on all proteins found in the genomes listed in Table 1. The size range refers to the entire proteins. Approximate position and size of domains (DUF916, WxL1, WxL2), identified by Hidden Markov models, are indicated by stripes. The average sequence identity refers to the entire proteins, and is particularly low for the CscC proteins, which only have WxL2 domains in common, and for CscD proteins, which only have the LPxTG anchors in common.

The NCBI and ERGO genome databases were searched for the presence of Csc family members and *csc*-like gene clusters. Clusters encoding these cell-surface proteins were found in the complete genomes of *Lactobacillus plantarum* (9 clusters)[31], *Lactobacillus sakei* (8)[32], *Enterococcus faecalis* (6), *Listeria innocua* (3), *Listeria monocytogenes* (2), *Lactococcus lactis ssp lactis* (3), *Bacillus cereus* ZK (1), *Bacillus cereus* 10987 (1, on plasmid) and *Bacillus anthracis* (1, on plasmid) (Table 1). The *csc* clusters were also found in the incomplete genomes of *L. lactis ssp cremoris* (5 clusters, of which one cluster on a plasmid), *Lactobacillus casei* (3), *Enterococcus faecium* (3), *Pediococcus pentocaseus* (2), *Oenococcus oeni* (1), *Leuconostoc mesenteroides* (1), and *Bacillus thuringiensis* (1). Details of all *csc* gene clusters and encoded proteins can be found in additional files 1, 2, 3. In several cases *csc* genes are still unidentified in incomplete genomes because the clusters are on small contigs. Each gene cluster generally has one copy each of the 4 new gene families *cscA*, *cscB*, *cscC* and *cscD*, although some variation is observed. A single copy of the *cscA* is always present, while 1–4 different *cscB* genes occur in the gene clusters. Although single *cscC* and *cscD* genes are usually present, they are missing in a few clusters. All encoded proteins have a regular signal peptide for secretion by the Sec-dependent pathway.

Evidence of gene clusters as functional units
There are many indications that these gene clusters are functional units, i.e. that the genes are transcribed coordinately, and that the encoded gene products function together in a pathway or protein complex.

C*sc* genes are nearly exclusively found in these gene clusters, with very few exceptions outside the clusters. The clusters rarely contain other genes than the *csc* family members, as based on the criteria of correct gene orientation, small intergenic distance and absence of predicted termination sequences. In all *csc* clusters, the genes are oriented in the same transcriptional direction and usually have intergenic regions smaller than 100 nucleotides, suggestive of an operon structure. In general, the *csc* gene clusters are bounded by terminators on both sides (Fig. 1). One complete gene cluster (LLX-I) on the *L. lactis ssp lactis* IL1403 chromosome is exactly bordered by IS981 elements, and several other clusters are flanked on one side by IS elements, suggesting that some of these gene clusters have been transferred as a unit. Moreover, complete *csc* gene clusters are found on plasmids of *L. lactis* SK11 [34], *B. anthracis* and *B. cereus* (see additional file 1), suggesting that these genes can be transferred between strains or species via these mobile genetic elements.

Comparative DNA microarray-based genotyping analysis of 20 strains of *Lactobacillus plantarum* revealed considerable variation in the presence/absence of different DNA regions in individual strains as compared to strain WCFS1 [35]. In general, the *csc* clusters of *L. plantarum* WCFS1 appear to be highly conserved in other strains. However, the entire cluster LPL-IX (LPL3676-3679) appears to be missing in 3 of the 20 strains analyzed, while the genes flanking this cluster appear to be present. Again, this suggests that the entire cluster can be excised or inserted as a functional unit.

Domain and function prediction of Csc proteins

CscA family

The CscA proteins are found to belong to the PF06030 Pfam family (or DUF916, bacterial proteins of unknown function). In addition to the N-terminal signal peptide, these proteins all contain a predicted C-terminal transmembrane helix, which presumably serves to anchor them in the cell membrane (see full sequence alignment in additional file 6). Each *csc* gene cluster generally encodes only a single CscA protein (see additional file 1). The CscA-family members are fairly uniform in size (320–380 residues), and the large majority are predicted to be very basic proteins with a pI above 9.0 (see additional file 2).

CscB family

The CscB family members are also fairly uniformly sized (190–280 residues, with a few exceptions), and typically have an acidic pI of 4–5. These proteins are not yet described in the Pfam or COG databases. We have defined the C-terminal domain of about 160–190 residues as the "WxL1" domain (Fig. 1; see full sequence alignment in additional file 7) since it contains two highly conserved sequence motifs Trp-x-Leu. Preceding the first Trp-x-Leu motif is a highly conserved Asp-x-Arg-Gly sequence. Most family members have a short Pro-rich region between the signal peptide and the WxL1-domain. The four exceptions are much larger proteins of *L. plantarum* (LPL1446, LPL3412) and *E. faecalis* (EF0405, EF0406) that have the C-terminal WxL1 domain in common; the larger N-terminal parts of these *L. plantarum* proteins are similar to each other, but have no known other domains, whereas the two *E. faecalis* proteins are also similar to each other and have L-domain-like repeats (see below).

CscC family

The CscC family members are much larger than CscA or CscB proteins, and more heterogeneous in size (500–900 residues, with some exceptions). They are multi-domain proteins, all characterized by a C-terminal domain of about 130–140 residues, defined as the "WxL2" domain since it is very similar to the WxL1 domain but differs in overall size, in conserved residues and in the distance between the two WxL motifs (see full alignment of WxL2 domains in additional file 8). Based on these differences, the WxL1 and WxL2 domains can be distinguished as different domain variants, which is also supported by Hidden Markov Models: CscB proteins were recognized by a Hidden Markov model based on the WxL1 domain without false positive hits in CscC proteins, and vice versa.

In addition, other domains could be identified in some CscC proteins with homology to different kinds of binding domains, albeit often with weak homology (see additional file 4). The clearest domain-homologue identified is an N-terminal domain of about 300 residues with structural similarity to concanavalin A-like lectins/glucanases. This superfamily includes a diverse range of carbohydrate-binding domains and glycosyl hydrolase enzymes that share a common structural fold (see Pfam clan CL0004) [36-38]. Lectins and glucanases exhibit the common property of reversibly binding to specific (complex) carbohydrates. This ConA-like domain was found in ten CscC proteins from six different species, and is characterized by several conserved aromatic residues, most of which are tryptophans (see full sequence alignment in additional file 9). Aromatic residues of starch-binding domains have been shown to be involved in the binding of saccharide rings by stacking with indole and phenyl rings [39]. Various (semi)-conserved Asp and Glu residues are potential metal ion ligands, including an ExD motif, as also found in glycosyl hydrolases of this superfamily (see Pfam clan CL0004). The ConA-like domains of CscC proteins show distinct sequence similarity to each other, but much less to other families of the large concanavalin A-like lectin/glucanase superfamily, suggesting that they may represent a new subfamily. The best sequence similarity is with leguminous plant lectins, including the known metal ion binding residues (alignment in additional file 13).

CscD family

The CscD family is not characterized by sequence similarity, but rather by the presence of both a signal peptide for secretion, and by an LPxTG-type motif for covalent anchoring to the peptidoglycan matrix. CscD proteins form a very unusual group among the LPxTG-proteins [14], [], since they are extremely short (90–140 residues) and have only 40–60 residues between the signal peptide (which is removed by signal peptidase I) and the LPxTG-anchoring motif (which is cleaved by sortase). This implies that only a short peptide of that length would become attached to the peptidoglycan. These peptides have very low sequence homology to each other, and multiple sequence alignment is not informative. We propose that they play a role in anchoring the other Csc proteins to the cell surface through as yet unknown interactions.

Cluster evolution

Family tree analysis of the CscA, CscB and CscC proteins (see additional files 10, 11, 12) suggests first that the clusters have evolved as units without shuffling, as the three trees are basically the same. Secondly, some cluster duplications are of early origin as they precede several speciation events. Other cluster duplications are of more recent origin, as cluster members from the same species are grouped in the same branch, as can be clearly seen in species with many clusters, i.e. *L. plantarum*, *L. sakei*, *E. faecalis* and *L. lactis*. Also, the gene order in clusters of these more recent duplications has changed little, compared to older

duplications (see additional file 3). Finally, multiple copies of *cscB* genes in clusters appear to be the most recent duplications, as they are most similar to members within the same cluster (see additional files 1, 11).

Co-expression and regulation of cluster genes

Several previous transcriptome investigations aimed at elucidation of *L. plantarum* response under various stress conditions have indicated that the transcription of specific *csc* genes is regulated in response to bile, salt and lactate stress [41,42]. In several cases, the expression of entire *csc* gene clusters was observed to change significantly.

In the present study, seven of the nine *csc* gene clusters of *L. plantarum* appeared to be significantly up-regulated as a consequence of a replacement mutation in the *ccpA* gene (encoding catabolite control protein A, CcpA) when grown on glucose as the main energy and carbon source (Table 2; Figure 3). These data strongly suggest that these gene clusters are part of the catabolite control regulon that is controlled by the central regulator CcpA. To further substantiate this, a MAST-motif search was performed to identify putative CRE sites, for binding of CcpA [43,44], within the *csc* gene clusters and their upstream regions. Putative CRE sites could be identified for six out of the seven up-regulated *csc* clusters, generally upstream of the first gene of the cluster and in three clusters also inside *csc* genes (Figure 1, Table 3). In contrast, no significant CRE-like sites could be identified within or upstream of the residual *csc* gene clusters, supporting a functional role of the identified CRE-site candidate sequences in regulation of these clusters.

Taken together these data strongly support the consistent coordinated expression of the L. plantarum csc clusters,

while a putative role for specific subsets of these clusters in stress survival/adaptation or in carbon source acquisition can be anticipated.

Discussion

Conserved gene clusters encoding extracellular proteins belonging to four distinct new families have been found in several gram-positive bacteria. Based on the experimental evidence and predictions provided above that the CscA, CscpB, CscC and CscD proteins are functionally coupled, we propose that they form a cell-surface protein complex. Two components are presumably bound to cell-wall components, i.e. the CscA is membrane-anchored and CscD is bound to peptidoglycan. The CscB and CscC proteins have novel WxL domains which could function in binding to CscA/CscD proteins, or to other components of the cell-surface (peptidoglycan, polysaccharides, teichoic acids, etc). The occurrence of these *csc* clusters in a limited number of gram-positive bacteria suggests a niche adaptation. All of the species in Table 1 are free-living bacteria found in the environment. Several of these bacteria are known to be associated with plants and plant fermentations, and many are used for making a variety of fermented products such as sauerkraut, sourdough, olives, silage, soy milk, wine and cheese, or can be found as contaminants of these products. *L. sakei* is more often associated to meat products [32]. It is noteworthy that these gene clusters are neither present in the many sequenced genomes of (mostly pathogenic) streptococci, staphylococci, and clostridia, nor in the *Lactobacillus acidophilus* subgroup of the lactobacilli, which are typical gut bacteria.

Experimental characterization of a Csc family protein has demonstrated its cell-surface location [45]. A *cscB* gene product called Cpf (Co/aggregation-Promoting Factor) of
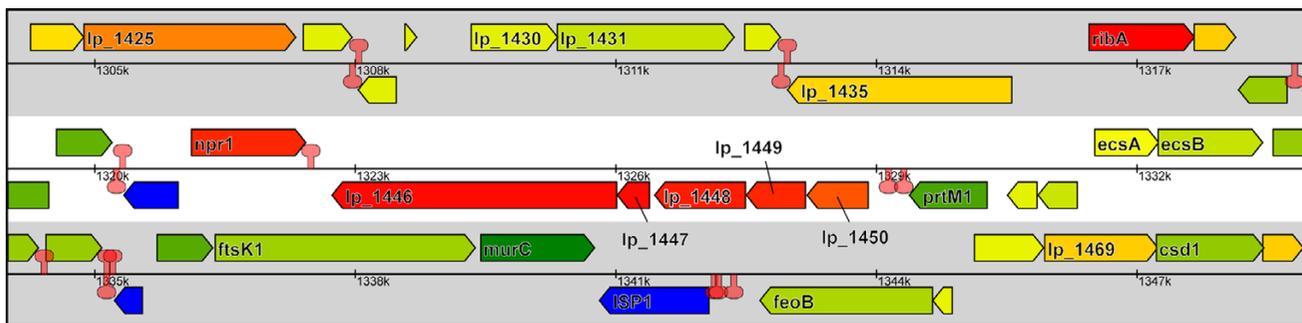
Figure 3
Example of coordinated gene expression of *csc* gene clusters found in *Lactobacillus plantarum* WCFS1. Cluster LPL-I (genes lp1446–lp1450) is shown with flanking genes. Genes are color-coded according to relative gene expression (measured by transcriptome analysis) in a comparison of the wild-type strain and a *ccpA* knock-out mutant upon growth on glucose. A sliding color scheme is used from down regulation (dark green) to up regulation (dark red) of genes in mutant *vs* wild type. Blue genes were not measured. Predicted terminators are indicated by loop symbols. Figure were made with the Microbial Genome Viewer [79, 80].

Table 2: Gene expression data of *L. plantarum*, growth of *ccpA* mutant vs wild-type

				ccpA mutant/wild-type	
cluster	gene	Csc family	product	Ma	P-value
I	LPL1450	B	extracellular protein	1.00	0.0008
	LPL1449	B	extracellular protein	1.22	0.0003
	LPL1448	A	cell surface protein precursor	1.24	0.0002
	LPL1447	D	cell surface protein precursor	1.36	0.0002
	LPL1446	B	extracellular protein	1.20	0.0006
II	LPL2175	B	extracellular protein	1.35	0.0005
	LPL2174	A	cell surface protein precursor	0.79	0.0048
	LPL2173	C	extracellular protein	0.71	0.0032
III	LPL2978	B	extracellular protein	0.08	0.7133
	LPL2977	A	cell surface protein precursor	0.90	0.0005
	LPL2976	D	cell surface protein precursor	0.59	0.0077
	LPL2975	C	extracellular protein	0.37	0.0466
IV	LPL3067	B	extracellular protein	2.00	0.0001
	LPL3066	A	cell surface protein precursor	1.85	0.0001
	LPL3065	D	cell surface protein precursor	1.49	0.0002
	LPL3064	C	extracellular protein	1.51	0.0001
V	LPL3075	C	cell surface protein (putative)	-0.16	0.0780
	LPL3074	D	cell surface protein precursor	0.32	0.0357
	LPL3073	B	extracellular protein	0.01	0.0600
	LPL3072	A	cell surface protein precursor	-0.68	0.0016
VI	LPL3117	C	cell surface protein (putative)	1.54	0.0002
	LPL3116	B	extracellular protein	1.76	0.0002
	LPL3115	A	cell surface protein precursor	1.85	0.0000
VII	LPL3414	B	extracellular protein	1.21	0.0006
	LPL3413	A	cell surface protein precursor	0.86	0.0079
	LPL3412	B	extracellular protein	0.63	0.0056
VIII	LPL3454	D	cell surface protein (putative)	0.67	0.0087
	LPL3453	B	extracellular protein	0.79	0.0013
	LPL3452	B	extracellular protein	1.26	0.0002
	LPL3451	A	cell surface protein precursor	1.13	0.0008
	LPL3450	C	extracellular protein	1.46	0.0006
IX	LPL3679	B	extracellular protein	-0.17	0.0159
	LPL3678	A	cell surface protein precursor	0.12	0.1040
	LPL3677	D	cell surface protein precursor	0.49	0.0220
	LPL3676	C	extracellular protein	0.70	0.0018

a M = ^{2}log of the expression ratio's, calculated as the average expression level observed in the *ccpA* mutant divided by the average expression ratio observed in the wild-type.

Lactobacillus coryniformis DSM20001^T, a species commonly found in agricultural habitats and food products, was purified and found to mediate coaggregation with and aggregation of other bacterial species. Cpf could be removed from the surface of *Lactobacillus* cells by treatment with high salt (5 M LiCl), and could subsequently be reattached by removal of salt resulting in restoration of

the co/aggregation property. This indicates that CscB proteins are non-covalently bound to the bacterial cell surface, supporting our hypothesis.

Transcriptomics experiments show that at least six of the *csc* gene clusters of *L. plantarum* are under catabolite repression, as they are up-regulated in a *ccpA*-knockout

Table 3: Putative CRE sites of *L. plantarum csc* clusters

cluster	gene	position	sequence	E-score
I	LPL1450	starts 66 and 42 bp upstream of gene	TGATTATCGTTACCA TGATCACCGCAGGCA	n.a.
	LPL1449	inside gene	TGTAAGCGTCACCA	3,60E-05
	LPL1446	inside gene	TGGAACCGCTGGCA	6,80E-06
II	LPL2175	starts 71 bp upstream of gene	TGAAAGCGGAATCA	2,60E-05
III	LPL2977	inside gene	TGATAACGGCATCA	5,00E-06
IV	LPL3067	starts 263 bp upstream of gene	TGTAACCGTTATCC	8,80E-05
	LPL3066	inside gene	TGGAACCCTTAACA	6,30E-05
	LPL3064	inside gene	TGCAAGCGTATCCA	1,60E-06
V	none			
VI	LPL3117	starts 62 and 31 bp upstream of gene	TGTGAGCGCTATCA AGATTACGCTGTCA	7,80E-06 7,80E-05
VII	none			
VIII	LPL3454	starts 121 bp upstream of gene	TGGAATCGCTGTCA	1,20E-05
IX	none			
consensus *Bacillus*			TGAAAGCGTTTTCA	

n.a. = clear palindromes which are possible CRE sites, but score is lowered due to insertion of 1 extra nucleotide

strain grown on glucose, and they contain CRE elements for binding of the global regulator CcpA. This regulatory clue suggests a functional link of the Csc proteins with sugar metabolism. Furthermore, some CscC proteins contain ConcanavalinA-like lectin/glucanase domains. ConA-like domains are often found in proteins involved in cell recognition and adhesion, and lectins and glucanases are known to reversibly bind to specific complex carbohydrates. Bacterial and fungal glucanases and xylanases with ConA-like domains can degrade complex polysaccharides like beta-glucans, kappa-carrageenans, xylans and cellulose [36-38,46]. Hence, the presence of ConA-like domains in CscC proteins would support a role of the proposed Csc cell-surface protein complex in binding and/or degradation of complex (plant-derived) oligo- or poly-saccharides. Plant cell-wall polysaccharides are an abundant source of carbon and energy for many free-living micro-organisms, which exploit such polysaccharides from decaying plant material, i.e. in compost, soil, and sewage.

It is striking that the genome of *Lactobacillus plantarum* has the most *csc* gene clusters. *L. plantarum* is frequently found on plants [21,23] and fermented plant material [47], and it is used in plant fermentations [48,49] and silage [22,24]. On plant surfaces, *L. plantarum* should be in close association with other bacteria (or fungi) which are capable of plant polysaccharide degradation and *L. plantarum* could make use of the liberated oligosaccharide units. In addition, or alternatively, *L. plantarum* could have its own extracellular enzyme systems for breakdown of complex polysaccharides, and we hypothesize that the newly described Csc system could be one of such systems.

Extracellular protein complexes for degradation of complex polysaccharides are already known in other groups of bacteria, but they are completely different in protein composition from the putative Csc protein complexes. Some anaerobic bacteria such as *Clostridium* and *Ruminococcus* have an elaborate system called the cellulosome, a large extracellular enzyme complex, to break down plant cell walls. In clostridia, the components of cellulosomes are encoded in large gene clusters [50-52], which are coordinately expressed and regulated by catabolite repression [53]. *Bacteroides thetaiotaomicron*, found in the distal intestine (colon) of the GI-tract, has an outer-membrane-associated multi-protein complex called the starch-utilization system (Sus), consisting of different starch-binding proteins and sugar degradation enzymes encoded in gene clusters [54-57]. Hence, it is not unlikely that during evolution different extracellular protein complexes have arisen in subgroups of bacteria, each specific for a particular environmental niche with its characteristic carbohydrate sources.

Conclusion

We have presented bioinformatics and experimental evidence that the extracellular CscA, CscB, CscC and CscD proteins are functionally coupled and possibly form a cell-surface protein complex that could play a role in sugar acquisition. Based on the occurrence of these gene clusters in many environmental Gram-positive bacteria, we postulate a role in degradation and utilization of (complex) plant polysaccharides, and possibly other food polysaccharides. Our hypotheses provide a guide for experimental work in any of these bacteria to investigate the location and composition of these protein complexes, their polysaccharide specificity and degradation properties, or the effect of knock-out mutants on the survival of the strain(s) grown on different substrates.

Methods

Bioinformatics analysis

Sequence information was obtained from the NCBI bacterial genome database [58] and the ERGO database [59]. The ERGO gene nomenclature was used; conversions to SwissProt nomenclature, where possible, is provided in additional file 5. Genome context was visualised in ERGO and with the Artemis viewer [60]. Terminators were determined with TransTerm [61]. Multiple alignments were created using ClustalW [62] and MUSCLE [63]. Signal peptides were predicted with SignalP [64], and transmembrane helices were detected with TMHMM 2.0 [65]. Conserved sequence patterns and novel domains and motifs were identified with MEME [66] and MAST [67]. Previously described domains were identified by scanning protein sequences with Hidden Markov Models (HMMs) from the PFAM [68], SMART [69] and SUPERFAM [70] databases using the HMMER package. HMMs were compared with HHsearch [71]. Protein family trees were made with LOFT (Rene van der Heijden, personal communication).

Motifs representing catabolite-responsive elements (CRE) were searched by first constructing a MEME profile [66] using 22 established CRE-containing sequences from *B. subtilis* [44]. With this profile, the program MAST [67] was used to detect CRE sites in the *L. plantarum* WCFS1 genome.

Members of the Csc families (see below) were searched for in the NCBI and ERGO databases using BLASTP and Hidden Markov Models (HMMs), starting with the *L. plantarum* Csc protein sequences as seeds, followed by iterative rounds of searches until saturation was reached. Subsequently, we used gene context to search the neighborhood of identified *csc* genes to find additional members of the *csc* gene clusters. This step involved searching in the encoded proteins for signal peptides, LPxTG-type anchoring motifs, and domains containing the WxL

motifs (using Hidden Markov Models). In several cases, the correct CDSs were only found after making corrections for missed ORFs, incorrect start codons, frame shifts, etc (see additional files 1, 2).

Strains, growth conditions, and transcriptome profiling

L. plantarum strain LM3 [72] is a close relative of the sequenced strain WCFS1 [31,35] and previous CGH analyses have shown that DNA microarrays based on the genome of strain WCFS1 can be used for transcriptome profiling in this strain: 92% of the probes on the array hybridized with LM3 DNA (D. Molenaar, unpublished data; [35]). Strain LM3 appears to contain all nine *csc* clusters that were identified in the WCFS1 genome, as concluded from array-based genotyping efforts [35] The LM3 strain was used in these studies because a *ccpA*-mutant derivative of this strain is available, LM3-2 (*ccpA::cat*) [72]. Both the parental strain LM3 and its *ccpA* derivative LM3-2 were grown in the 0.25 × MRS medium (prepared without carbon source; [42]) supplemented with 2% glucose. The 1 liter vessel chemostat (Applikon Dependable Instruments, Schiedam, The Netherlands) was operated with 500 ml working volume at 37°C, pH 6.0, 125 rpm, and a flow rate of 120 ml h^{-1} [73]. The aerobic condition was maintained by sparging the vessel with air at a rate of 29 ml min^{-1}. The culture pH was controlled automatically by the addition of 0.5 N HCl or 0.5 N NaOH. The cultures were inoculated with 20 ml of an overnight culture and grown as a batch culture until mid-exponential phase, when continuous feeding of fresh medium was initiated. Samples for RNA extraction were drawn when steady state was reached, that was assumed to require five residence times.

In order to avoid degradation, conversion and *de novo* synthesis of mRNA molecules during sampling of cell culture, we performed a quenching method for collection and centrifugation of cells [74]. Cell pellet was resuspended in TE buffer and transferred in a chilled 2-ml microcentrifuge tube containing 1 g of 0.1-mm-diameter zirconium beads (Biospect Products), 0.25 g macaloid (Kronos Titan GmbH, Leverkusen), 50 µl SDS 10% and 500 µl phenol. The cells were broken by bead-beating [75] at room temperature for 4 times 30 sec, with intermittent cooling on ice for 3 min. After centrifugation for 10 min at 14,500 × g at 4°C, phenol-chloroform extraction was performed until the water phase was clear. RNA was precipitated overnight at 20°C with 1 volume isopropanol, pelleted by centrifugation at 14,500 × g, 20 min, at 4°C, washed once with 70% ethanol and resuspended in appropriate volume of RNase-free MQ-water. Contaminating chromosomal DNA was removed by digestion with RNase-free RQ1 DNase (1 U/µl; Promega) for 15 min at 37°C followed by RNA precipitation with 0.3 M Na-acetate and two volumes of ethanol. The pellet was resuspended in

RNase-free MQ-water and determination of sample concentration and quality was performed by an A_{260} and A_{280} reading and by agarose gel electrophoresis. RNA preparations were stored at -80°C until used.

RNA samples were labelled according to previously described methods. The labelled RNA samples were hybridized to previously described, clone-based DNA microarrays that cover more than 80 % of the *L. plantarum* WCFS1 genome, representing 88% of the annotated open reading frames [35]. Hybridizations and washing of the slides, as well as scanning and primary data analyses were performed as previously described.

Statistical analysis

Microarrays containing fragments of the *L. plantarum* WCFS1 genome as probes were used to measure the expression of genes. The design and production of these arrays as well as the normalization of spot data was described before [76]. Statistical analysis of the data was performed using the "limma" package for R [77,78]. Averaging of spot data to obtain gene-related data was performed as described before [76]. The eBayes function in the limma package was applied to obtain a cross-probe variance estimation and false discovery rate corrected p-values for the whole set of probes. The weighted geometric mean of the false-discovery rate (FDR) corrected p-values was calculated as an indication of significance, although these means do not equal FDR corrected p-values anymore for the complete list of genes.

Abbreviations

BLASTP Basic Local Alignment Search Tool for Proteins

CRE Catabolite Responsive Element

HMM Hidden Markov Model

MAST Motif Alignment and Search Tool

MEME Multiple Em for Motif Elicitation

PFAM Protein Family database

SMART Simple Modular Architecture Research Tool

TMHMM TransMembrane Hidden Markov Model

Authors' contributions

RS conceived of the study, participated in its design and coordination, and drafted the manuscript. LM performed the growth studies and microarray experiments, DM performed the statistical analysis, and MK supervised experimental work. JB, BR and RS performed the genome data mining and other bioinformatics sequence analyses. MK

and JB contributed to drafting the manuscript and revising it critically for intellectual content.

Additional material

Additional file 1
Table 4: Summary of csc *genes and gene clusters.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-126-S1.pdf]

Additional file 2
Table 5: Predicted properties of encoded Csc proteins.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-126-S2.pdf]

Additional file 3
Table 6: Order of genes in cell-surface clusters.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-126-S3.pdf]

Additional file 4
Table 7: Putative other domains in CscB and CscC proteins.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-126-S4.pdf]

Additional file 5
Table 8: Conversion of ERGO codes to SwissProt codes.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-126-S5.pdf]

Additional file 6
Figure 4: Multiple sequence alignment of CscA proteins.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-126-S6.pdf]

Additional file 7
Figure 5: Multiple sequence alignment of CscB proteins.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-126-S7.pdf]

Additional file 8
Figure 6: Multiple sequence alignment of CscC proteins.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-126-S8.pdf]

Additional file 9
Figure 7: Multiple sequence alignment of ConA-like lectins/glucanases domains of CscC proteins.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-126-S9.pdf]

Additional file 10

Figure 8: Family tree of CscA proteins.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-126-S10.pdf]

Additional file 11
Figure 9: Family tree of CscB proteins.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-126-S11.pdf]

Additional file 12
Figure 10: Family tree of CscC proteins.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-126-S12.pdf]

Additional file 13
Figure 11: Multiple sequence alignment of ConA-like lectins/glucanases domains of CscC proteins with known 3D structures of lectins.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-126-S13.pdf]

Additional file 14
Legends to additional Figures.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-126-S14.doc]

References
1. Cabanes D, Dehoux P, Dussurget O, Frangeul L, Cossart P: **Surface proteins and the pathogenic potential of Listeria monocytogenes.** *Trends Microbiol* 2002, **10**:238-245.
2. Goward CR, Scawen MD, Murphy JP, Atkinson T: **Molecular evolution of bacterial cell-surface proteins.** *Trends Biochem Sci* 1993, **18**:136-140.
3. Karjalainen T, Waligora-Dupriet AJ, Cerquetti M, Spigaglia P, Maggioni A, Mauri P, Mastrantonio P: **Molecular and genomic analysis of genes encoding surface-anchored proteins from Clostridium difficile.** *Infect Immun* 2001, **69**:3442-3446.
4. Navarre WW, Schneewind O: **Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope.** *Microbiol Mol Biol Rev* 1999, **63**:174-229.
5. Schwarz-Linek U, Hook M, Potts JR: **The molecular basis of fibronectin-mediated bacterial adherence to host cells.** *Mol Microbiol* 2004, **52**:631-641.
6. Frick IM, Schmidtchen A, Sjobring U: **Interactions between M proteins of Streptococcus pyogenes and glycosaminoglycans promote bacterial adhesion to host cells.** *Eur J Biochem* 2003, **270**:2303-2311.
7. Marino M, Banerjee M, Jonquieres R, Cossart P, Ghosh P: **GW domains of the Listeria monocytogenes invasion protein InlB are SH3-like and mediate binding to host ligands.** *Embo J* 2002, **21**:5623-5634.
8. Cue D, Lam H, Cleary PP: **Genetic dissection of the Streptococcus pyogenes M1 protein: regions involved in fibronectin binding and intracellular invasion.** *Microb Pathog* 2001, **31**:231-242.
9. Doi RH, Kosugi A, Murashima K, Tamaru Y, Han SO: **Cellulosomes from mesophilic bacteria.** *J Bacteriol* 2003, **185**:5907-5914.
10. Doi RH, Kosugi A: **Cellulosomes: plant-cell-wall-degrading enzyme complexes.** *Nat Rev Microbiol* 2004, **2**:541-551.
11. Gal L, Pages S, Gaudin C, Belaich A, Reverbel-Leroy C, Tardif C, Belaich JP: **Characterization of the cellulolytic complex (cellulosome) produced by Clostridium cellulolyticum.** *Appl Environ Microbiol* 1997, **63**:903-909.
12. Bayer EA, Belaich JP, Shoham Y, Lamed R: **The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides.** *Annu Rev Microbiol* 2004, **58**:521-554.
13. Belaich JP, Tardif C, Belaich A, Gaudin C: **The cellulolytic system of Clostridium cellulolyticum.** *J Biotechnol* 1997, **57**:3-14.
14. Boekhorst J, de Been MW, Kleerebezem M, Siezen RJ: **Genome-wide detection and analysis of cell wall-bound proteins with LPxTG-like sorting motifs.** *J Bacteriol* 2005, **187**:4928-4934.
15. Dhar G, Faull KF, Schneewind O: **Anchor structure of cell wall surface proteins in Listeria monocytogenes.** *Biochemistry* 2000, **39**:3725-3733.
16. Mazmanian SK, Liu G, Ton-That H, Schneewind O: **Staphylococcus aureus sortase, an enzyme that anchors surface proteins to the cell wall.** *Science* 1999, **285**:760-763.
17. Mazmanian SK, Ton-That H, Schneewind O: **Sortase-catalysed anchoring of surface proteins to the cell wall of Staphylococcus aureus.** *Mol Microbiol* 2001, **40**:1049-1057.
18. Ton-That H, Marraffini LA, Schneewind O: **Protein sorting to the cell wall envelope of Gram-positive bacteria.** *Biochim Biophys Acta* 2004, **1694**:269-278.
19. Sutcliffe IC, Harrington DJ: **Pattern searches for the identification of putative lipoprotein genes in Gram-positive bacterial genomes.** *Microbiology* 2002, **148**:2065-2077.
20. Jonquieres R, Bierne H, Fiedler F, Gounon P, Cossart P: **Interaction between the protein InlB of Listeria monocytogenes and lipoteichoic acid: a novel mechanism of protein association at the surface of gram-positive bacteria.** *Mol Microbiol* 1999, **34**:902-914.
21. Mundt JO, Hammer JL: **Lactobacilli on plants.** *Appl Microbiol* 1968, **16**:1326-1330.
22. Gibson T, Stirling AC, Keddie RM, Rosenberger RF: **Bacteriological changes in silage made at controlled temperatures.** *J Gen Microbiol* 1958, **19**:112-129.
23. Ercolani GL: **Distribution of epiphytic bacteria on olive leaves and the influence of leaf age and sampling time.** *Microbial Ecology* 1991, **21**:35-48.
24. Keddie RM: **The properties and classification of lactobacilli isolated from grass and silage.** *Journal of Applied Bacteriology* 1959, **22**:403-416.
25. Ercolini D, Hill PJ, Dodd CE: **Bacterial community structure and location in Stilton cheese.** *Appl Environ Microbiol* 2003, **69**:3540-3548.
26. Gardner NJ, Savard T, Obermeier P, Caldwell G, Champagne CP: **Selection and characterization of mixed starter cultures for lactic acid fermentation of carrot, cabbage, beet and onion vegetable mixtures.** *Int J Food Microbiol* 2001, **64**:261-275.
27. Aymerich T, Martin B, Garriga M, Hugas M: **Microbial quality and direct PCR identification of lactic acid bacteria and nonpathogenic Staphylococci from artisanal low-acid sausages.** *Appl Environ Microbiol* 2003, **69**:4583-4594.
28. Ruiz-Barba JL, Piard JC, Jimenez-Diaz R: **Plasmid profiles and curing of plasmids in Lactobacillus plantarum strains isolated from green olive fermentations.** *J Appl Bacteriol* 1991, **71**:417-421.
29. Ahrne S, Nobaek S, Jeppsson B, Adlerberth I, Wold AE, Molin G: **The normal Lactobacillus flora of healthy human rectal and oral mucosa.** *J Appl Microbiol* 1998, **85**:88-94.
30. Vaughan EE, Heilig HG, Ben-Amor K, de Vos WM: **Diversity, vitality and activities of intestinal lactic acid bacteria and bifidobacteria assessed by molecular approaches.** *FEMS Microbiol Rev* 2005, **29**:477-490.
31. Kleerebezem M, Boekhorst J, van Kranenburg R, Molenaar D, Kuipers OP, Leer R, Tarchini R, Peters SA, Sandbrink HM, Fiers MW, Stiekema W, Lankhorst RM, Bron PA, Hoffer SM, Groot MN,

Kerkhoven R, de Vries M, Ursing B, de Vos WM, Siezen RJ: **Complete genome sequence of Lactobacillus plantarum WCFS1.** *Proc Natl Acad Sci U S A* 2003, **100:**1990-1995.

32. Chaillou S, Champomier-Verges MC, Cornet M, Crutz-Le Coq AM, Dudez AM, Martin V, Beaufils S, Darbon-Rongere E, Bossy R, Loux V, Zagorec M: **The complete genome sequence of the meatborne lactic acid bacterium Lactobacillus sakei 23K.** *Nat Biotechnol* 2005, **23:**1527-1533.

33. van Wely KH, Swaving J, Freudl R, Driessen AJ: **Translocation of proteins across the cell envelope of Gram-positive bacteria.** *FEMS Microbiol Rev* 2001, **25:**437-454.

34. Siezen RJ, Renckens B, van Swam I, Peters S, van Kranenburg R, Kleerebezem M, de Vos WM: **Complete Sequences of Four Plasmids of Lactococcus lactis subsp. cremoris SK11 Reveal Extensive Adaptation to the Dairy Environment.** *Appl Environ Microbiol* 2005, **71:**8371-8382.

35. Molenaar D, Bringel F, Schuren FH, de Vos WM, Siezen RJ, Kleerebezem M: **Exploring Lactobacillus plantarum genome diversity by using microarrays.** *J Bacteriol* 2005, **187:**6119-6127.

36. Coutinho PM, Deleury E, Davies GJ, Henrissat B: **An evolving hierarchical family classification for glycosyltransferases.** *J Mol Biol* 2003, **328:**307-317.

37. Gilkes NR, Henrissat B, Kilburn DG, Miller RCJ, Warren RA: **Domains in microbial beta-1, 4-glycanases: sequence conservation, function, and enzyme families.** *Microbiol Rev* 1991, **55:**303-315.

38. Davies GJ, Gloster TM, Henrissat B: **Recent structural insights into the expanding world of carbohydrate-active enzymes.** *Curr Opin Struct Biol* 2005, **15:**637-645.

39. Rodriguez-Sanoja R, Oviedo N, Sanchez S: **Microbial starch-binding domain.** *Curr Opin Microbiol* 2005, **8:**260-267.

40. Boekhorst J, Kelmer Q, Kleerebezem M, Siezen RJ: **Comparative analysis of proteins with amucus-binding domain found exclusively in lactic acid bacteria.** *Microbiology* 2006, **152:**273-80.

41. Bron PA, Molenaar D, de Vos WM, Kleerebezem M: **DNA microarray-based identification of bile-responsive genes in Lactobacillus plantarum.** *J Appl Microbiol* 2006, **100:**728-738.

42. Pieterse B, Leer RJ, Schuren FH, van der Werf MJ: **Unravelling the multiple effects of lactic acid stress on Lactobacillus plantarum by transcription profiling.** *Microbiology* 2005, **151:**3881-3894.

43. Moreno MS, Schneider BL, Maile RR, Weyler W, Saier MHJ: **Catabolite repression mediated by the CcpA protein in Bacillus subtilis: novel modes of regulation revealed by whole-genome analyses.** *Mol Microbiol* 2001, **39:**1366-1381.

44. Miwa Y, Nakata A, Ogiwara A, Yamamoto M, Fujita Y: **Evaluation and characterization of catabolite-responsive elements (cre) of Bacillus subtilis.** *Nucleic Acids Res* 2000, **28:**1206-1210.

45. Schachtsiek M, Hammes WP, Hertel C: **Characterization of Lactobacillus coryniformis DSM 20001T surface protein Cpf mediating coaggregation with and aggregation among pathogens.** *Appl Environ Microbiol* 2004, **70:**7078-7085.

46. Ay J, Gotz F, Borriss R, Heinemann U: **Structure and function of the Bacillus hybrid enzyme GluXyn-1: native-like jellyroll fold preserved after insertion of autonomous globular domain.** *Proc Natl Acad Sci U S A* 1998, **95:**6613-6618.

47. Tamminen M, Joutsjoki T, Sjoblom M, Joutsen M, Palva A, Ryhanen EL, Joutsjoki V: **Screening of lactic acid bacteria from fermented vegetables by carbohydrate profiling and PCR-ELISA.** *Lett Appl Microbiol* 2004, **39:**439-444.

48. Patel HM, Wang R, Chandrashekar O, Pandiella SS, Webb C: **Proliferation of Lactobacillus plantarum in solid-state fermentation of oats.** *Biotechnol Prog* 2004, **20:**110-116.

49. Amoa-Awua WK, Appoh FE, Jakobsen M: **Lactic acid fermentation of cassava dough into agbelima.** *Int J Food Microbiol* 1996, **31:**87-98.

50. Tamaru Y, Karita S, Ibrahim A, Chan H, Doi RH: **A large gene cluster for the Clostridium cellulovorans cellulosome.** *J Bacteriol* 2000, **182:**5906-5910.

51. Belaich A, Parsiegla G, Gal L, Villard C, Haser R, Belaich JP: **Cel9M, a new family 9 cellulase of the Clostridium cellulolyticum cellulosome.** *J Bacteriol* 2002, **184:**1378-1384.

52. Nolling J, Breton G, Omelchenko MV, Makarova KS, Zeng Q, Gibson R, Lee HM, Dubois J, Qiu D, Hitti J, Wolf YI, Tatusov RL, Sabathe F, Doucette-Stamm L, Soucaille P, Daly MJ, Bennett GN, Koonin EV, Smith DR: **Genome sequence and comparative analysis of the solvent-producing bacterium Clostridium acetobutylicum.** *J Bacteriol* 2001, **183:**4823-4838.

53. Han SO, Yukawa H, Inui M, Doi RH: **Transcription of Clostridium cellulovorans cellulosomal cellulase and hemicellulase genes.** *J Bacteriol* 2003, **185:**2520-2527.

54. Cho KH, Salyers AA: **Biochemical analysis of interactions between outer membrane proteins that contribute to starch utilization by Bacteroides thetaiotaomicron.** *J Bacteriol* 2001, **183:**7224-7230.

55. D'Elia JN, Salyers AA: **Effect of regulatory protein levels on utilization of starch by Bacteroides thetaiotaomicron.** *J Bacteriol* 1996, **178:**7180-7186.

56. Shipman JA, Berleman JE, Salyers AA: **Characterization of four outer membrane proteins involved in binding starch to the cell surface of Bacteroides thetaiotaomicron.** *J Bacteriol* 2000, **182:**5365-5372.

57. Reeves AR, Wang GR, Salyers AA: **Characterization of four outer membrane proteins that play a role in utilization of starch by Bacteroides thetaiotaomicron.** *J Bacteriol* 1997, **179:**643-649.

58. **NCBI bacterial genome database** [ftp://ftp.ncbi.nih.gov/genomes/Bacteria/]

59. Overbeek R, Larsen N, Walunas T, D'Souza M, Pusch G, Selkov EJ, Liolios K, Joukov V, Kaznadzey D, Anderson I, Bhattacharyya A, Burd H, Gardner W, Hanke P, Kapatral V, Mikhailova N, Vasieva O, Osterman A, Vonstein V, Fonstein M, Ivanova N, Kyrpides N: **The ERGO genome analysis and discovery system.** *Nucleic Acids Res* 2003, **31:**164-171.

60. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16:**944-945.

61. Jacobs GH, Rackham O, Stockwell PA, Tate W, Brown CM: **Transterm: a database of mRNAs and translational control elements.** *Nucleic Acids Res* 2002, **30:**310-311.

62. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs.** *Nucleic Acids Res* 2003, **31:**3497-3500.

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

64. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Int J Neural Syst* 1997, **8:**581-599.

65. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305:**567-580.

66. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers: ; Menlo Park, California,.** *Volume 2.* AAAI Press; 1994:28-36.

67. Bailey TL, Gribskov M: **Combining evidence using p-values: application to sequence homology searches.** *Bioinformatics* 1998, **14:**48-54.

68. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR: **The Pfam protein families database.** *Nucleic Acids Res* 2004, **32:**D138-41.

69. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P: **SMART 4.0: towards genomic data integration.** *Nucleic Acids Res* 2004, **32:**D142-4.

70. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.** *J Mol Biol* 2001, **313:**903-919.

71. Soding J: **Protein homology detection by HMM-HMM comparison.** *Bioinformatics* 2005, **21:**951-960.

72. Muscariello L, Marasco R, De Felice M, Sacco M: **The functional ccpA gene is required for carbon catabolite repression in Lactobacillus plantarum.** *Appl Environ Microbiol* 2001, **67:**2903-2907.

73. Starrenburg MJ, Hugenholtz J: **Citrate Fermentation by Lactococcus and Leuconostoc spp.** *Appl Environ Microbiol* 1991, **57:**3535-3540.

74. Pieterse B, Jellema RH, van der Werf MJ: **Quenching of microbial samples for increased reliability of microarray data.** *J Microbiol Methods* 2006, **64:**207-216.

75. Lopez de Felipe F, Starrenburg MJC, Hugenholtz J: **The role of NADH-oxidation in acetoin and diacetyl production from glucose in Lactococcus lactis subsp. lactis MG1363.** *FEMS Microbiology Letters* 1997, **156:**15-19.

76. Sturme MH, Nakayama J, Molenaar D, Murakami Y, Kunugi R, Fujii T, Vaughan EE, Kleerebezem M, de Vos WM: **An agr-like two-component regulatory system in Lactobacillus plantarum is involved in production of a novel cyclic peptide and regulation of adherence.** *J Bacteriol* 2005, **187:**5224-5235.

77. Smyth GK, Yang YH, Speed TP: **Statistical issues in microarray data analysis.** In *Functional Genomics: Methods and Protocols Volume 224.* Edited by: Brownstein MJ and Khodursky AB. Totowa, NJ, USA, Humana Press; 2003:111-136.

78. **The R project for Statistical Computing** [http://www.r-project.org]

79. Kerkhoven R, van Enckevort FH, Boekhorst J, Molenaar D, Siezen RJ: **Visualization for genomics: the Microbial Genome Viewer.** *Bioinformatics* 2004, **20:**1812-1814.

80. **Microbial Genome Viewer** [http://www.cmbi.ru.nl/MGV]