cse, a Chimeric and Variable Gene, Encodes an Extracellular Protein Involved in Cellular Segregation in Streptococcus thermophilus

Frédéric Borges, Séverine Layec, Annabelle Thibessard, Annabelle Fernandez, Brigitte Gintz, Pascal Hols, Bernard Decaris, Nathalie Leblond-Bourget

To cite this version:

HAL Id: hal-01631252
https://hal.univ-lorraine.fr/hal-01631252
Submitted on 9 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
cse, a Chimeric and Variable Gene, Encodes an Extracellular Protein Involved in Cellular Segregation in Streptococcus thermophilus

Frédéric Borges, Séverine Layec, Annabelle Thibessard, Annabelle Fernandez, Brigitte Gintz, Pascal Hols, Bernard Decaris, and Nathalie Leblond-Bourget

Laboratoire de Génétique et Microbiologie, UMR INRA 1128, IFR 110, Faculté des Sciences et Techniques de l’Université Henri Poincaré Nancy 1, Vandoeuvre-lès-Nancy, France, and Unité de Génétique, Institut des Sciences de la Vie, Université catholique de Louvain, Louvain-La-Neuve, Belgium

Received 2 September 2004/Accepted 4 January 2005

The isolation of a Streptococcus thermophilus CNRZ368 mutant displaying a long-chain phenotype allowed us to identify the cse gene (for cellular segregation). The N terminus of Cse exhibits high similarity to Streptococcus agalactiae surface immunogenic protein (SIP), while its C terminus exhibits high similarity to S. thermophilus PcsB. In CNRZ368, deletion of the entire cse open reading frame leads to drastic lengthening of cell chains and altered colony morphology. Complementation of the Δcse mutation with a wild-type allele restored both wild-type phenotypes. The central part of Cse is a repeat-rich region with low sequence complexity. Comparison of cse from CNRZ368 and LMG18311 strains reveals high variability of this repeat-rich region. To assess the impact of this central region variability, the central region of LMG18311 cse was exchanged with that of CNRZ368 cse. This replacement did not affect chain length, showing that divergence of the central part does not modify cell segregation activity of Cse. The structure of the cse locus suggests that the chimeric organization of cse results from insertion of a duplicated sequence deriving from the pcsB 3′ end into an ancestral sip gene. Thus, the cse locus illustrates the module-shuffling mechanism of bacterial gene evolution.

Materials and Methods

Bacterial strains, growth conditions, and plasmids. All strains and plasmids used in this work are presented in Table 1. Depending on the experiments, S. thermophilus and its derivatives were cultivated in milk medium, TYP (tryptose, 7 g·liter⁻¹; Proteose peptone, 7 g·liter⁻¹; yeast extract, 2 g·liter⁻¹; lactose, 20 g·liter⁻¹) (5), or M17 (51) medium without shaking. Milk medium was used for strain storage. M17 was used for mutant generation, and TYP was employed for phenotypic analysis. Phenotypic analyses were performed at 42°C, the optimal growth temperature of S. thermophilus. Erythromycin was added at 2 μg·ml⁻¹ when required. S. thermophilus derivative strains containing pGB9 (27) or pNST260+ plasmid derivatives were cultivated at 30°C when plasmid self-maintenance was required and at 42°C for selection of clones with the chromosome’s integrated plasmid. S. thermophilus strains containing pFUN or its derivatives were cultivated at 30°C. L. lactis strains were grown at 30°C in M17 medium (51) supplemented with 0.5% glucose. Erythromycin was added at 5 μg·ml⁻¹ when required.

Recombinant plasmids derived from pGB9 were transformed into Escherichia coli EC101, a TGI strain containing a chromosomal copy of the pWV01 repA gene (6), and selected at 37°C on Luria-Bertani (LB) (47) containing 150 μg of erythromycin ml⁻¹. Recombinant plasmids derived from pFUN, which was kindly provided by J. Poquet from INRA Jouy en Josas, France (38) were transformed into E. coli DH5α and selected on LB medium containing 100 μg of ampicillin ml⁻¹.

DNA manipulations. Preparation of chromosomal and plasmid DNA and Southern analysis were performed according to standard protocols (47). Se-

Among gram-positive cocci, some species grow as single cells and others grow as grouped cells, linked together into various arrangements. In particular, cocci from several genera, such as Streptococcus, Lactococcus, Peptostreptococcus, Ruminococcus, and Coprococcus, grow as cell chains (48). The number of cells per chain varies considerably among species and strains (14). Indeed, chain length can range from two cells, as in Streptococcus pneumoniae, to around 100 cells in Streptococcus mitior (14). Few genes have been implicated in short-chain phenotypes, since their inactivation leads to an increased number of cells per chain. One of the most striking phenotypes is that of an S. pneumo-

...
sequencing was performed by using dye terminator chemistry on an ABI Prism 377 genetic analyzer (PE Biosystems). Sequence data were analyzed with BLAST (1, 23). Oligonucleotides used in this study were purchased from MWG Biotech AG (Ebersburg, Germany) and are listed Table 2. (i) Translational fusion with Δspnuc. To obtain the plasmid pFUN::cse, the region covering the putative promoter region and the open reading frame (ORF) except its stop codon, was amplified by PCR and introduced into pFUN, leading to cseΔspnuc fusion. For pFUN::Δspnuc, the regions flanking the sequence encoding the putative signal peptide (the 81 bp following the ATG putative initiation site of translation) were amplified by PCR and introduced into pFUN. This led to a cseΔspnuc fusion where the sequence encoding the putative signal peptide of Cse was replaced by a HindIII restriction site. Following ligation together and with pGh9, after introduction of the recombinant plasmid into S. thermophilus, two crossovers, upstream and downstream of the deleted region, were selected (52). Thus, only the first and the last three codons of the cse open reading frame were conserved as previously described (52).

TABLE 1. Strains and plasmids used in this work

<table>
<thead>
<tr>
<th>Strain or plasmid</th>
<th>Relevant phenotype(s) or genotype(s)</th>
<th>Source or reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. thermophilus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNRZ368</td>
<td>Wild-type strain</td>
<td>INRA-CNRZ, strain collection</td>
</tr>
<tr>
<td>16D10</td>
<td>Em’ S. thermophilus CNRZ368</td>
<td>This work</td>
</tr>
<tr>
<td>CNRZ368-Δcse</td>
<td>Em’ S. thermophilus CNRZ368 with Δcse mutation</td>
<td>This work</td>
</tr>
<tr>
<td>CNRZ368-Δcse-T</td>
<td>Em’, CNRZ368-Δcse with pNST260Δ+::cse integrated into the chromosome</td>
<td>This work</td>
</tr>
<tr>
<td>CNRZ368-Δcse-C</td>
<td>Wild-type strain</td>
<td>This work</td>
</tr>
<tr>
<td>LMG18311</td>
<td>S. thermophilus LMG18311 with Δcse mutation</td>
<td>This work</td>
</tr>
<tr>
<td>LMG18311-Δcse</td>
<td>Derivative of MG1363 wild-type strain containing 701-bp SacI-SpeI chromosomal deletion in acmA gene</td>
<td>6</td>
</tr>
<tr>
<td>L. lactis subsp. cremoris</td>
<td>MG1363acmAΔcse</td>
<td>This work</td>
</tr>
<tr>
<td>E. coli</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EC101</td>
<td>supE hsd-5 thi (lac-proAB) F’ (traB6 proAB’ lacI1 lacZΔM15) repA’, derivative of TG1 strain (47)</td>
<td>23</td>
</tr>
<tr>
<td>DHLα</td>
<td>supE44 ΔlacU169 (d80 lacZ ΔM15) hisD17 endA1 gyrA96 thi-1 relA1</td>
<td>47</td>
</tr>
<tr>
<td>BL21 (DE3)</td>
<td>hsdR gal (lacI857 indI1 sam7 nin5 lacU1V5-T7 gene I)</td>
<td>47</td>
</tr>
<tr>
<td>Plasmids</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pGh9</td>
<td>Em’, thermosensitive replication origin from pVE6902</td>
<td>27</td>
</tr>
<tr>
<td>pGh9::ISSI</td>
<td>pGh9 with ISSI</td>
<td>28</td>
</tr>
<tr>
<td>pGh9::ΔcseCNRZ368</td>
<td>pGh9 with Δcse and CNRZ368 surrounding regions</td>
<td>This work</td>
</tr>
<tr>
<td>pGh9::ΔcseLMG18311</td>
<td>pGh9 with Δcse and LMG18311 surrounding regions</td>
<td>This work</td>
</tr>
<tr>
<td>pGh9::cseLMG18311varCNRZ368</td>
<td>pGh9 with chimeric allele cseLMG18311varCNRZ368</td>
<td>This work</td>
</tr>
<tr>
<td>pNST260Δ+::cse</td>
<td>pNST260Δ+ with CNRZ368 cse ORF, its putative promoter and terminator</td>
<td>This work</td>
</tr>
<tr>
<td>pFUN</td>
<td>Em’ and Amp’, derivative of PI252, with Δspnuc gene</td>
<td>38</td>
</tr>
<tr>
<td>pFUN::cse</td>
<td>pFUN with cseΔspnuc translational fusion</td>
<td>Novagen</td>
</tr>
<tr>
<td>pET15b::cse</td>
<td>Protein expression vector</td>
<td>This work</td>
</tr>
<tr>
<td>pET15b::cseΔ183</td>
<td>pET15b with hexa-His-Cse translational fusion</td>
<td>This work</td>
</tr>
<tr>
<td>pET15b::cseΔ183+183</td>
<td>pET15b with hexa-His-CseΔ183+183 translational fusion</td>
<td>This work</td>
</tr>
</tbody>
</table>

TABLE 2. Oligonucleotides used in this work

<table>
<thead>
<tr>
<th>Oligonucleotide</th>
<th>Sequence* (5’–3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>delcse.E3</td>
<td>CCCCCCCCAAGCTTCAAGATCGATCGTTTGAAC</td>
</tr>
<tr>
<td>delcse.I5</td>
<td>CCCCCCGAATTCTTGATATAGAATTATATCATTC</td>
</tr>
<tr>
<td>delcse.I3</td>
<td>CCCCCGGAATTTATTTGCAATAGATGACATTAC</td>
</tr>
<tr>
<td>delcse.E3</td>
<td>CCCCCCGTGCAGCAGCCTTGGTGGTTGAG</td>
</tr>
<tr>
<td>fcompl</td>
<td>TTTGTGCGGCGCTTAAGCGGTTGTTTGGTTTGAAG</td>
</tr>
<tr>
<td>rcompl</td>
<td>TTTGTGGCGGCGCTTAATTATCTTCTTCTGAGGATAT</td>
</tr>
<tr>
<td>race.E3</td>
<td>CCCCCCGAGCTTGAGATAGTGGTGGTGGTGAATATC</td>
</tr>
<tr>
<td>race.E5</td>
<td>CCCCCGATATCGATATATATATATATATATATAT</td>
</tr>
<tr>
<td>ravar.S</td>
<td>CAGTGGTCATGACGACAAAC</td>
</tr>
<tr>
<td>ravar.I3</td>
<td>GCATATTCCGATTCATGACCA</td>
</tr>
<tr>
<td>fncse.E5</td>
<td>CCCCCCGGATTTCTGATAGGCTAGGGGCTT</td>
</tr>
<tr>
<td>fncse.I5</td>
<td>CCCCCCGGATTTCTGATAGGCTAGGGGCTT</td>
</tr>
<tr>
<td>fncse.I3</td>
<td>CCCCCCGGATTTCTGATAGGCTAGGGGCTT</td>
</tr>
<tr>
<td>fncse.I3</td>
<td>CCCCCCGGATTTCTGATAGGCTAGGGGCTT</td>
</tr>
<tr>
<td>sex.E5</td>
<td>CCCCCCGCTAACATTTTACCAATACCAATACCA</td>
</tr>
<tr>
<td>sex.I5</td>
<td>CCCCCCGCTAACATTTTACCAATACCAATACCA</td>
</tr>
<tr>
<td>sex.I3</td>
<td>CCCCCCGCTAACATTTTACCAATACCAATACCA</td>
</tr>
<tr>
<td>sex.E3</td>
<td>CCCCCCGCTAACATTTTACCAATACCAATACCA</td>
</tr>
</tbody>
</table>

* Restriction enzyme sites are underlined.
ORF were kept, and the remainder of the ORF was replaced by an EcoRI restriction site. The cse deletion was checked by PCR and Southern hybridization (data not shown).

(iii) Δcse complementation. Complementation of the Δcse mutation was carried out by inserting the wild-type allele in trans. For this purpose, the pNST260+/cse plasmid (G. Guédon, personal communication), carrying a gene encoding the ICE1s1 integrase and its attL attachment site, was used (7). This plasmid can integrate into the S. thermophilus CNRZ368 chromosome at the attL site of ICE1s1 (G. Guédon, personal communication). The entire cse ORF with its putative promoter and terminator was amplified by PCR and ligated into pNST260+. Following selection in E. coli EC101, the construct was introduced into S. thermophilus by electroporation. As a control, S. thermophilus was also transformed with empty pNST260+. Integration of pNST260+ and pNST260+cse into the chromosome was selected at 42°C, the replication-restrictive temperature, in the presence of erythromycin. Plasmid integration into the chromosome attachment site was checked by Southern hybridization (data not shown).

(iv) Allelic replacement of the var-cse region. The approach for var-cse replacement involved the wild-type sequence containing GAT and ATC sequences localized, respectively, upstream and downstream of var-cse. These sequences allow an EcoRV restriction site formation if they are assembled. The plasmid used for replacement of the var-cse region of LMG18311 by that of CNRZ368 was prepared in a two-step cloning procedure. The var-cse flanking regions from LMG18311 strain were amplified by PCR. After digestion by appropriate restriction enzymes, PCR fragments were ligated together, generating an EcoRV restriction site, and inserted into pGh9. This EcoRV restriction site was generated by assembling the GAT and ATC sequences that flank the wild-type var-cse sequence. The plasmid carrying the two PCR fragments was selected in E. coli EC101. Then the region localized between the GAT and ATC sequences included a previously region from CNRZ368 (as defined in Fig. 2D) was amplified by PCR. The PCR product was next treated at 72°C for 30 min with Pfu polymerase for generation of blunt ends. This last step removes the 3′ adenosine overhang added during the PCR. The polished fragment was inserted into the plasmid carrying the var-cse flanking regions previously digested by EcoRV to generate blunt ends. The resulting plasmid was selected in E. coli EC101. The correct insert orientation was checked by sequencing. Then the plasmid was introduced into S. thermophilus LMG18311 by electroporation. The transformed strain was checked by Southern hybridization with a specific probe for the var-cse region, which were selected to generate the LMG18311-cseΔ::cseΔ mutant.

(v) Plasmid constructs for protein overproduction. The whole cse ORF and the regions encoding Cse parts from amino acid 1 to 183 and from amino acid 184 to 461 were amplified by PCR. The PCR products were independently introduced into the pET15b plasmid (Novagen), ending in an in-frame fusion downstream of the hexa-His encoding sequence. After selection in E. coli DE3S5a, the constructs were introduced into E. coli BL21(DE3) for protein overproduction.

Microscopy. Colonies and cells were observed with a Nikon OPTIPHOT microscope mounted with phase contrast equipment (Ph). Colony examination was done at a magnification of ×100. Cells were observed at a magnification of ×1,000, with the condenser turret at position Ph6, or at a magnification of ×1,000 by phase contrast.

Protein overproduction. E. coli BL21(DE3) strains transformed with pET15b and derivatives were grown in LB medium at 37°C supplemented with 50 μg of ampicillin ml⁻¹. At an optical density at 600 nm (OD600) of 0.6, isopropyl-β-D-thiogalactopyranoside (IPTG) was added at a final concentration of 100 μM during 4 h to induce expression of N-terminally hexa-His-tagged proteins. To extract proteins, cell pellets resuspended in sample electrophoresis buffer were heated for 5 min at 95°C. For cell lysis testing in nondenaturing conditions, proteins were extracted by sonication.

SDS-PAGE and renaturing SDS-PAGE. Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) was performed as previously described (22) with 10% (wt/vol) polyacrylamide. Lytic activity was tested by SDS-PAGE, performed with gels containing 10% polyacrylamide (wt/vol) and 0.2% (wt/vol) S. thermophilus cells or Micrococcus lysothrophicus cells, followed by a protein renaturation step. The S. thermophilus cells used as substrates for lytic activity were prepared as previously described (39) with the following modifications: after hypothermophilization, cells were resuspended in distilled water at a 10% final concentration and autoclaved for 20 min at 120°C. S. thermophilus proteins were extracted by glass bead disruption of cells as previously described (17). After electrophoresis, gels were gently shaken in 100 ml of water at 4°C for 1 h. Then water was replaced by 100 ml of 20 mM Tris-HCl (pH 7) containing 1% (vol/vol) Triton X-100 for overnight incubation at 42°C. Bands of lytic activity were visualized as previously described (6).

Nucleotide sequence accession numbers. DNA sequences reported in this paper have been deposited in GenBank under accession numbers AY695844, AY730642, and AY730643.

RESULTS
cse involvement in colony morphology and cell segregation. The 16D10 mutant, displaying a long-chain phenotype in addition to a variant colony morphology, was selected from an S. thermophilus CNRZ368 mutant collection. This collection was previously obtained by insertional mutagenesis (53) with the pGh9:ISS1 plasmid (28). The locus disrupted by pGh9:ISS1 in the 16D10 mutant was identified by cloning and sequencing the 2,697 bp flanking the disruption site. Sequence analysis revealed that pGh9:ISS1 disrupted an ORF of 1,386 bp, named cse for cell segregation. The CNRZ368 cse ORF encodes a putative 462-amino-acid protein of unknown function with a deduced molecular mass of approximately 48.5 kDa. The first 28 amino acids display Sec-dependent signal peptide characteristics as predicted by signalP software (35), suggesting that Cse is exported. The Cse central part, containing repeats (see below), is highly enriched in alanine, glutamate, threonine, and serine (20.4, 32.9, 15.3, and 8.3%, respectively) residues compared to the mean content of S. thermophilus proteins (6.1, 5.9, 5.7, and 4.6%, respectively), according to the Codon Usage Database (32). This central region is potentially hydrophilic (due to its high glutamate, serine, and threonine content) and acidic (due to its glutamate enrichment).

To confirm the involvement of cse in these phenotypes, a cse null mutant was constructed by in-frame deletion of the cse ORF, resulting in a CNRZ368-cseΔ strain. Colonies of this mutant were flatter, larger, and less opaque than wild-type ones (Fig. 1A). Phase-contrast photonic microscopy observations revealed that CNRZ368-cseΔ chains were much longer than wild-type ones (Fig. 1B). This was reinforced by counting the number of cells per chain for each strain. Only 15% of wild-type cells formed chains containing more than 100 cells in stationary phase, whereas 100% of CNRZ368-cseΔ cells did (Fig. 1C). During exponential growth (OD600 = 0.2), 76.5% of the wild-type cells counted formed chains containing less than 40 cells and none of them formed chains containing more than 80 cells. On the contrary, 87.5% of the CNRZ368-cseΔ cells formed chains containing more than 100 cells. These results show that the long-chain phenotype of the CNRZ368-cseΔ mutant is not growth phase dependent. No major modification of the cell shape or size was observed in the CNRZ368-cseΔ mutant. Moreover, mutant and wild-type strain growth curves showed no significant difference in either doubling time or in stationary phase entry OD600 value (data not shown).

Complementation was performed by using the pNST260+ integrative plasmid (G. Guédon, personal communication). This plasmid can integrate into the S. thermophilus CNRZ368 chromosome in the attR site of ICE1s1 (G. Guédon, personal communication), a resident integrative element (7). The Δcse mutation was complemented in trans by chromosomal integration of the cse wild-type gene carried by the pNST260+cse plasmid, resulting in the CNRZ368-cseΔ-cse strain. The negative control consisted of the chromosomal insertion of empty pNST260+, resulting in the CNRZ368-cseΔ-cseΔ-T strain. The
complemented strain exhibited a wild-type phenotype with respect to colony morphology and number of cells per chain, while the negative control maintained a mutant phenotype (Fig. 1C and colony morphology data not shown). These results demonstrate that cse is involved in colony morphology and cell segregation.

Central part of cse is repeat rich and variable. Alignment of the cse nucleotide sequence from strains CNRZ368 and LMG18311 (cse\textsubscript{CNRZ368} and cse\textsubscript{LMG18311}) showed that the first 600 bp and the last 346 bp of these ORFs display 94 and 96% identity, respectively (Fig. 2A). In contrast, both cse allele central parts, consisting of 443 bp in cse\textsubscript{CNRZ368} and 503 bp in cse\textsubscript{LMG18311}, show only 61% identity. Because of its variability, this region was called var-cse. Both var-cse regions show a high level of redundancy consisting of short direct repeats, with tandem or dispersed arrangement (Fig. 2B). As a consequence of its high content in short repeats, it is a low-complexity region as detected by the SEG program. Nine different repeat units (named A to I) that varied in size and sequence were distinguished (Fig. 2C). All of the repeated units have a size divisible by three and are in frame. The direct consequence is that all nucleic acid repeat sequences correspond to amino acid repeats in the cse product (Fig. 2B). The comparison of var-cse from S. thermophilus CNRZ368 and LMG18311 showed that var-cse\textsubscript{LMG18311} is 60 bp longer than var-cse\textsubscript{CNRZ368}. Additionally, the repeat content of both var-cse alleles is qualitatively different (Fig. 2D). In conclusion, the central part of cse is almost exclusively built with repeated sequences and displays a high degree of intraspecies variability.

Secondary structures were predicted for Var-Cse by using the PSIPRED method (19). Almost the entire Var-Cse\textsubscript{CNRZ368} (93%) region and the whole Var-Cse\textsubscript{LMG18311} region are predicted to adopt coil structures, suggesting that this region is a nonglobular domain. However, the confidence value of the prediction is low (3.56/11006 1.79 for Var-Cse\textsubscript{CNRZ368} and 3.45/11006 1.83 for Var-Cse\textsubscript{LMG18311}, on a 0 to 10 scale). The predicted nonglobular structure of this region could explain the tolerance for its high genetic variability.

cse is chimeric. The nucleotide sequence of cse and its flanking regions in S. thermophilus LMG18311 was used in a BLASTN search against the genome of the same strain. Results revealed a region of 619 bp, named HRC1 for homologous region, copy 1, sharing 93% identity with another region of 618 bp, localized elsewhere in the genome and named HRC2 for homologous region, copy 2 (Fig. 3A). HRC1 includes the last 350 bp of cse, the following intergenic region, and the first 58 bp of the downstream \textit{orf1} ORF. HRC2 includes the last 350 bp of \textit{pcsB}, the following intergenic region, and the first 58 bp of the downstream \textit{rppk} ORF (encoding a putative ribose-phosphate pyrophosphokinase) (Fig. 3A). Thus, the C-terminal part of Cse is homologous to the C-terminal part of PcsB from S. thermophilus LMG18311 (Fig. 3A). BLAST searches within nonredundant databases showed that a PcsB orthologue is present in the \textit{Streptococcus} and \textit{Lactococcus} genera, whereas no protein with significant similarity to the whole Cse protein was found in these organisms. PcsB is essential in the maintenance of cell shape in\textit{Streptococcus} species (9, 33, 34, 42, 43). Searches of the Conserved Domain Database show that both PcsB and Cse possess a region homologous to the CHAP domain (cysteine histidine-dependent aminohydrolase/peptidase) (Fig. 3A) which possesses a glutathionylspermidine amidase activity in \textit{E. coli} (4). The CHAP domain is widely distributed in extracellular proteins, where it is supposed to be involved in peptidoglycan hydrolysis (3, 45).
FIG. 2. *var-cse* variability and repeat content of *cse*, *pcsB*, and *orf1*. (A) Schematic representation of the *cse* locus from *S. thermophilus* strains CNRZ368 and LMG18311. Open arrows represent ORFs and indicate their reading direction, broken arrows indicate putative promoters, and hairpin loops symbolize putative rho-independent terminators. The black arrowhead indicates the insertion site of pGh9:IS1 within the genome of the 16D10 mutant. The percentage of sequence identity (id.) is indicated between 5' ends, central parts, and 3' ends of *cse* from CNRZ368 and LMG18311. Hatched boxes represent repeat-containing regions (*var-cse*). (B) Nucleic acid and amino acid sequences of *var-cseCNRZ368* and the proximal region. The regions flanking *var-cse* are italicized. Each repeat unit is boxed, and the name of the repeat unit is indicated above the box. (C) Table of consensus repeat sequences represented in *cse*, *pcsB*, and *orf1*. (D) Schematic representation of repeat-containing regions. Each repeat unit is represented by a letter-containing box. Repeat units, in accordance with consensus sequences presented in panel C, are represented by capital letters, while those slightly divergent from the consensus (17% of maximum divergence) are represented by lowercase letters. Empty boxes represent regions without any repeats. Repeats in grey are common between *cse* from *S. thermophilus* strains CNRZ368 and LMG18311, *orf1*, and *pcsB* from *S. thermophilus* strain LMG18311. Hatched boxes represent repeats only found in *cse*. 

2741
In contrast, the N-terminal part of Cse is homologous to the N-terminal part of the surface immunogenic protein (SIP) from *Streptococcus agalactiae* (Fig. 3A). Searches of the Conserved Domain Database showed that the N-terminal parts of both SIP and Cse contain regions homologous to the LysM domain (Fig. 3A), found in extracellular proteins and involved in their attachment to the cell wall (49). This domain is probably responsible for preferential localization of the SIP surface protein at the cell poles (46, 49).

Interestingly, the C-terminal region of SIP is homologous to the C-terminal part of a putative protein (ORF1) (Fig. 3A) encoded by the gene immediately downstream of *cse* (Fig. 3A).

The CDART database contains two hypothetical proteins from *Streptococcus mutans* and *Streptococcus intermedius* with one LysM domain and one CHAP domain (GenBank accession numbers NP_720819 and BAB61101, respectively). These proteins have a LysM domain at the N terminus and a CHAP domain at the C terminus. Both proteins show significant similarity (48% identity and 63% similarity) over 92% of their length. Their CHAP domains show significant similarity with Cse (51% identity and 62% similarity for NP_720819, 60% identity and 72% similarity for BAB61101), contrary to that of the LysM-containing region, which does not exhibit significant similarity with either Cse or SIP.

**pcsB and orf1 contain common repeats with cse.** Considering partial homology between *pcsB* and cse loci, *pcsB* and orf1, *pcsB* and rppk, and orf1 and rppk, were scanned for repeats. The analysis revealed that the *pcsB* central region and orf1 5′ end also contain repeats (Fig. 3A) that are common to both cse alleles (Fig. 2D). More precisely, all repeats found in *pcsB* (A, C, E, H, and I) are also represented in var-cse, and one repeat (A) found in *pcsB* is represented in var-cse. The orf1 5′ end from CNRZ368 and LMG18311 strains (99.6% identity between orf1 alleles) contains C, E, and F repeated sequences, although all but one are degenerate. Additionally, the success of ECCF is found twice in orf1 and displayed once in var-cse (Fig. 2D).

Interestingly, var-cse and the orf1 repeat-rich regions are located closed to HRC1 boundaries (Fig. 3A). Thus, HRC1 is flanked by two repeat-rich regions which contain repeated motifs common to each other and to the *pcsB* repeat-rich region.

**Extracellular localization of Cse.** According to SignalP (35) analysis, Cse contains a canonical putative signal peptide, suggesting that it is exported from the cell. To test this hypothesis, cse was fused to the reporter gene Δnuc carried by the plasmid pFUN (38). The Δnuc gene encodes a *Staphylococcus aureus* nuclease lacking its signal peptide (10). The region containing the putative promoter and the entire cse ORF (except the stop codon) was cloned into the pFUN vector to generate a cse-Δnuc fusion. The resulting pFUN:cse vector was introduced into *E. coli* DH5α and *S. thermophilus* CNRZ368. Following growth, the transformants were overlaid with TDB agar (21), allowing detection of extracellular nuclease activity by visualization of a pink halo. A pink halo was visualized around colonies of each transformed strain (data not shown), indicating that the fusion protein was extracellular. As a negative control, the same region lacking the putative signal peptide encoding sequence was cloned into pFUN and displayed once in var-cse (Fig. 2D).

**Impact of var-cse variability on Cse cell segregation activity.** The variability of var-cse raised the question of a possible variability of Cse cell segregation activity. Analysis of *S. thermophilus* CNRZ368 and LMG18311 cell distribution in chains revealed intraspecies variability of cell chain length (Fig. 4A). Indeed, 85% of CNRZ368 cells formed chains from 1 to 100...
cells, whereas only 39% of LMG18311 cells are in this size range. Thus, a large proportion of LMG18311 cell chains were much longer than those of CNRZ368. Due to the involvement of cse in cell segregation, we hypothesized a correlation between variability of the cell chain length and var-cse variability. A Δcse mutant of LMG18311 was found to exhibit a long-chain phenotype, as the CNRZ368-Δcse strain does (Fig. 4B), confirming that the cseLMG18311 allele was functional. At the cse chromosomal locus, the var-cse region (Fig. 2D) was replaced by that of CNRZ368 in the LMG18311 genetic background. This was done by allelic replacement, taking into account that var-cse homologous flanking regions are not identical even if they show at least 94% identity (Fig. 2A). Thus, the plasmid used for allelic replacement should contain var-cse from strain CNRZ368 flanked by cse homologous regions from LMG18311 strain. No significant difference was seen between LMG18311 and LMG18311-cse varCNRZ368 strains with respect to chain length (Fig. 4A). This indicates that the cell segregation activity of Cse is not affected by the high variability of its central region.

**Cell wall hydrolase activity.** Since it was previously reported that cell segregation results from cell wall hydrolytic activity detectable by zymography (6, 8, 18, 36, 41, 58), Cse was tested for this activity. Thus, wild-type and CNRZ368-Δcse strain protein extracts were tested by zymography. Under our conditions, no difference was detected between the two lytic activity profiles, with either autoclaved *S. thermophilus* or *M. lysodeikticus* cells as a substrate (data not shown). Therefore, we tried to overproduce the Cse protein in *E. coli* with an N-terminal hexa-His tag. However, this approach was unsuccessful for the full-length Cse protein, as in the case of the full-length AcmB protein from *L. lactis* previously reported (16). The authors were able to circumvent this issue by overproducing AcmB fragments. Using the same alternative way, we were able to overproduce two Cse fragments in *E. coli*, one comprising the

![Fig. 4](image-url). Influence of var-cse variability on Cse cell segregation activity. (A) Results of counting numbers of cells per chain. A total of 1,000 cells were counted from three independent cultures. (B) Cell morphology of *S. thermophilus* strain LMG18311. Cells were photographed or counted in stationary phase after 20 h of growth. After this period of growth, OD values reached 0.66 ± 0.074, 1.1 ± 0.014, 1.1 ± 0.032, and 0.87 ± 0.049, respectively, for strains CNRZ368, LMG18311, LMG18311-cse varCNRZ368, and LMG18311-Δcse.
region from amino acid 1 to 183 and the other from amino acid 184 to 461. Thus, the N-terminal region comprised the LysM putative domain, and the C-terminal region comprised both the Var-Cse region and the CHAP domain. Renaturing SDSPAGE, performed with crude protein extract, either on autoclaved \textit{S. thermophilus} or \textit{M. lysodeikticus} cells, did not allow detection of any lytic activity attributable to either the N-terminal or the C-terminal fragment. To exclude that this negative result was not due to the incorrect refolding of the Cse fragments, a cell lysis test was performed under nondenaturing conditions. For this purpose, proteins of \textit{E. coli} Cse fragment overproducer strains were extracted by sonication. The control crude protein extract was obtained from an \textit{E. coli} strain carrying the pET15b plasmid without an insert. After checking that the overproduced proteins were recovered in the soluble fraction, 200 μg of crude protein extract were added to a TPPY suspension of autoclaved \textit{S. thermophilus} or \textit{M. lysodeikticus} cells and incubated at 42°C. OD<sub>600</sub> measurements over time did not reveal any significant differences between crude protein extracts containing the overproduced Cse fragments and the control.

Finally, we tried an indirect way, by checking whether \textit{cse} would be able to suppress a long-chain phenotype resulting from a murein hydrolase defect. For this purpose, \textit{L. lactis} ΔacmA was transformed with pNST260+::\textit{cse}, and the number of cells per chain was counted. No difference in chain length was noticed between ΔacmA and ΔacmA transformed with pNST260+::\textit{cse} (data not shown).

DISCUSSION

Three genes of \textit{S. thermophilus} (\textit{rodA}, \textit{php2b}, and \textit{mreD}) have been previously shown to be involved in cell chain length (50, 52, 54). These genes are also involved in cell shape, since mutants in these genes display rod-shaped cells instead of ovoid cells. In this study, we identified a fourth gene involved in the shortening of cell chains, called \textit{cse} for its role in cellular segregation. Deletion of \textit{cse} and complementation experiments have shown that the \textit{cse} chimeric gene is involved in cell segregation. The Cse extracellular protein contains a putative LysM domain and a putative CHAP domain. LysM domains of AcmA, the major peptidoglycan hydrolase of \textit{L. lactis}, are directly involved in cell wall attachment (49), and CHAP domains are widely found in surface proteins where they are supposed to be involved in peptidoglycan hydrolysis (3, 45). Therefore, Cse may be attached to the cell wall and may exhibit peptidoglycan hydrolase activity.

To test whether Cse has a peptidoglycan hydrolase activity, protein extracts of \textit{S. thermophilus} wild type, \textit{S. thermophilus} Δcse, and \textit{E. coli} Cse fragment overproducer strains were tested by zymography with embedded \textit{S. thermophilus} or \textit{M. lysodeikticus} cells as a substrate. In addition, cell lysis activity was tested in liquid medium, with native crude protein extracts of \textit{E. coli} Cse fragment overproducer strains. In both experiments, no Cse lytic activity was detected. Finally, \textit{cse} was not able to complement the long-chain phenotype resulting from a peptidoglycan hydrolase defect of an \textit{L. lactis} ΔacmA mutant. The absence of a detectable lytic activity was previously reported for PcsB from \textit{S. agalactiae} (42, 43), another protein with a CHAP domain. These results suggest that either Cse has no peptidoglycan hydrolase activity or Cse activity could not be detected by these assays.

The results reported here show that \textit{cse} is a chimeric gene and suggested that its 5′ and 3′ ends originated from sip-like and \textit{pcsB} genes, respectively. Indeed, the region comprising the 3′ end of \textit{cse}, the following intergenic region, and the beginning of \textit{orf1} (region named HRC1) shows 93% identity with the region comprising the \textit{pcsB} 3′ end, the following intergenic region, and the beginning of \textit{rppk} (region named HRC2) (Fig. 3A). Moreover, HRC1 is flanked by two regions encoding amino acid sequences homologous to parts of the SIP protein from \textit{S. agalactiae} (Fig. 3A). Furthermore, SIP-encoding genes had been found in the genomes of all streptococci except that of \textit{S. thermophilus} and \textit{S. mutans}, and a PcsB-encoding gene had been found in the genomes of all streptococci, including \textit{S. thermophilus}. Taken together, all of these data suggest that HRC2 had been subjected to a duplication event. The resulting duplicated sequence would have been inserted into a \textit{sip} ancestral gene, thus creating two chimeric genes, \textit{cse} and \textit{orf1} (Fig. 3B).

Interestingly, \textit{S. mutans} and \textit{S. intermedius} possess one LysM-CHAP protein. The \textit{S. mutans} genome, whose complete sequence is available, does not encode a protein with significant similarity to SIP of other streptococci and encodes a protein, named GbpB, homologous to whole PcsB (9). Thus, the genomic context of \textit{S. mutans} would be similar to that of \textit{S. thermophilus}, suggesting that Cse and the \textit{S. mutans} LysM-CHAP protein have a common ancestor that would have been generated by module shuffling. However, the \textit{S. mutans} LysM-CHAP protein does not show significant similarity to SIP and shows significant similarity with Cse only from its putative CHAP domain. Moreover, the C terminus of the LysM-CHAP protein harbors only 61% identity to the C terminus of GbpB, whereas the C terminus of Cse shows 93% identity with the C terminus of \textit{S. thermophilus} PcsB (Fig. 3A). If the LysM-CHAP protein of \textit{S. mutans} and Cse had been generated by the same DNA duplication event, it would be expected that the percentage of divergence between these LysM-CHAP proteins and their PcsB/GbpB genomic counterparts would be the same. These data support an alternative hypothesis where the LysM-CHAP protein from \textit{S. mutans} and Cse have different origins.

The consequence of the suspected DNA rearrangement that originated \textit{cse} is the association of a putative LysM-encoding sequence to a putative CHAP-encoding sequence, giving rise to a new protein generation (Cse) by domain shuffling. Domain shuffling has been proposed as a mechanism of gene evolution in bacteria (13, 15, 25, 44) and implies considering genes to be associations of modules. Each gene module would encode a protein domain which has been defined as “part of a protein that can fold up independently of neighboring sequences” (11). These modules could undergo rearrangements, ending in the creation of new module associations and, consequently, new protein domain associations. As an example, extracellular proteins from \textit{S. pneumoniae} that bind the choline component of the cell wall appear to evolve by domain shuffling. These multidomain proteins consist of a choline-binding domain, triggering protein association to the cell wall, and a catalytic domain that differs from one protein to another (25). Another example is the peptidoglycan hydrolase domain of \textit{Bacillus anthracis}. 
AmiA that is also found in other proteins where it is fused to different attachment signals (31).

The cse central region and the central region of its product are variable and repeat rich. The existence of variable tandemly repeated sequences is a common characteristic of extracellular proteins of gram-positive bacteria (20). This variability may facilitate adaptation to environmental changes (20), for instance, variability in the number of repeats in the alpha C protein allows group B streptococci to escape host immunity (20). In other cases, not restricted to extracellular proteins, genetic variability of repeat-rich regions directly modifies intrinsic biochemical activity of the protein, such as in restriction and modification enzymes encoded by EcoR124 and EcoR124/3 genes (40). These type I restriction enzymes recognize sites GAA(N6)RTCG and GAA(N7)RTCG, respectively, differing only in the length of a spacer. This difference in their specificity is due to two or three copies of a 12-bp sequence localized in their respective gene central parts (40). Therefore, we could not exclude that variations in the Cse repeat-rich region have consequences on cell segregation activity. However, replacement of var-cseLMG18311 by var-cseCNRZ368 at the LMG18311 cse chromosomal locus did not induce significant changes in cell segregation activity on the criterion of cell number per chain. Thus, it is likely that the high variability of the cse central region was not counter-selected because it does not affect the cell segregation activity of Cse.

The question of the role of the Cse central part in cell segregation remains open. This region could be a linker joining the N-terminal part to the C-terminal part of Cse. Such a function is assumed, for instance, by Q-linkers that join functionally distinct domains in nitrogen regulatory proteins (55). Variations in length and sequence of NifA and NtrC Q-linkers have no consequence on the activity of these proteins (55). Similarly, the replacement of var-cseLMG18311 by var-cseCNRZ368 ending in a 60-bp shortening of var-cse, since var-cseCNRZ368 is 60 bp shorter than var-cseLMG18311, did not have consequences on Cse cell segregation activity. A linker role of the Cse central part could explain why its high variability, and especially its length variability, does not affect cellular segregation. This possible function of the Cse central part would also be consistent with its predicted non-globular structure.

Two repeat-rich regions sharing common repeat motifs flank HRC1. This suggests that the insertion event of the duplicated region occurred within an ancestral repeat-rich region localized in the central part of the sip-like gene ancestor. Thus, SIP-encoding genes of other streptococci could also contain a repeat-rich region. Therefore, repeat content and sequence complexity level were analyzed in S. agalactiae SIP protein and orthologous counterparts in S. pneumoniae, Streptococcus pyogenes, Streptococcus gordoni, and Streptococcus suis. No repeat was found in any of these sequences. However, analysis of the amino acid composition showed that these regions have a low sequence complexity. Moreover, as observed for cse, their alignment revealed strong interspecies divergence of the central region (data not shown). Thus, although SIP proteins do not contain repeats, their central part showed low sequence complexity and interspecies variability. These data are in agreement with the existence of a low-complexity region in the central part of the sip-like gene ancestor of cse. We hypothesize that this low-complexity region could have been the site of an HRC1 insertion event responsible for cse creation. Interestingly, other multidomain proteins exhibit junctions with low-complexity sequences between domains (44, 55). For instance, the multiporphosphyl transfer protein of Rhodobacter capsulatus, a permease from a phosphotransferase system, is composed of three domains connected by two similar linker regions of 17 residues. These two linkers are rich in glycine, alanine, and proline residues, lowering the complexity of these regions (57). It has been proposed that the evolution of PTS permease occurred by interdomain shuffling and that this shuffling was allowed by genetic recombination between linker-encoding sequences (57). This example, together with cse, raises the question of a possible evolutionary advantage, at the genetic level, of DNA regions with a low-complexity sequence. Recombination events occurring inside domain-encoding sequences will probably inactivate, in many cases, the domain functionality and will therefore give rise to inefficient domain associations. We speculate that low-complexity regions can be more tolerant targets for genetic recombination events, responsible for domain shuffling, because of their low requirement in sequence and size.

ACKNOWLEDGMENTS

We thank Isabelle Poquet and Emmanuelle Maguin for providing the pFUN and pGh9:IS5 plasmids, respectively, and Valérie Legue for photographic contributions. We also thank Jan Kook for providing L. lactis MG1363:acsΔΔf and Gérard Guédon for allowing us to use pNST260+ prior to publication. We are grateful to Paul Hoskisson for help in preparing the manuscript.

S.L. and A. F. were supported by grants from the Ministère de l'Éducation Nationale de l'Enseignement Supérieur et de la Recherche. F.B. was supported by a grant from the Institut National de la Recherche Agronomique. P.H. is Research Associate at FNRS.

REFERENCES

LytB, a novel pneumococcal murein hydrolase essential for cell separation.