High genetic variability of the Streptococcus thermophilus cse central part, a repeat rich region required for full cell segregation activity

Frédéric Borges, Séverine Layec, Annabelle Fernandez, Bernard Decaris, Nathalie Leblond-Bourget

Antonie van Leeuwenhoek (2006) 90:245–255
DOI 10.1007/s10482-006-9079-5

ORIGINAL PAPER

High genetic variability of the *Streptococcus thermophilus cse* central part, a repeat rich region required for full cell segregation activity

Frédéric Borges · Séverine Layec · Annabelle Fernandez · Bernard Decaris · Nathalie Leblond-Bourget

Received: 30 January 2006 / Accepted: 11 April 2006 / Published online: 11 August 2006

Abstract The *cse* gene of *Streptococcus thermophilus* encodes an extracytoplasmic protein involved in cell segregation. The Cse protein consists of two putative domains: a cell wall attachment LysM domain and a catalytic CHAP domain. These two domains are spaced by an interdomain linker, known as Var-Cse, previously reported to be highly divergent between two *S. thermophilus* strains. The aim of this study was to assess the extent of this intraspecific variability and the functional involvement of the *var-cse* region in cell segregation. Analysis of the *var-cse* sequence of 19 different strains allowed detection of 11 different alleles, varying from 390 bp to 543 bp, all containing interspersed and tandem nucleotides repeats. Overall, 11 different repeat units were identified and some series of these small repeats, named supermotifs, form large repeats. Results suggested that *var-cse* evolved by deletion of all or part of the repeats and by duplication of repeats or supermotifs. Moreover, sequence analysis of the whole *cse* locus revealed that the *cse* ORF is mosaic suggesting that *var-cse* polymorphism resulted from horizontal transfer. The partial deletion of the *var-cse* region of the *S. thermophilus* strain CNRZ368 led to the lengthening of the number of cells per streptococcal chain, indicating that this region is required for full cell segregation in *S. thermophilus* strain CNRZ368.

Keywords Cell separation · Horizontal transfer · Mosaic · Mutation · Repeat

Abbreviations
CR Coding Repeat
Indel Insertion/deletion

F. Borges · S. Layec · A. Fernandez · B. Decaris · N. Leblond-Bourget (✉)
Laboratoire de Génétique et Microbiologie, UMR INRA 1128, IFR 110, Faculté des Sciences et Techniques de l'Université Henri Poincaré Nancy 1, BP 239, 54506 Vandoeuvre-lès-Nancy, France
e-mail: bourget@nancy.inra.fr

F. Borges
Department of Microbiology and Biotechnology, University of Ulm, Albert-Einstein-Allee 11, D-89069 Ulm, Germany

Introduction

The most relevant property of repeat sequences, from an evolutionary point of view, is their genetic plasticity. For intragenic repeats, known as coding repeats (CRs), the consequences of this instability, regarding the Open Reading Frame (ORF) integrity and the encoded amino acid sequence, are different according to their size and

their respective reading phase positioning. When the repeat motif size is not a multiple of three nucleotides, genetic instability of the repeat rich region can induce, at a high frequency, protein shortening or lengthening, inactivating or activating the ORF. Usually, genetic instabilities of these kinds of repeats induce switching between activation and inactivation of the gene and is a mechanism of gene expression regulation by phase variation (Hallet 2001). When the repeat size is a multiple of three and when the repeats are positioned in the same phase, the size and the amino acid sequences are modified but ORFs are always finished by the same stop codon whatever the allele and therefore are not inactivated by frameshifts. Intragenic repeats from the last category are widely studied in pathogenic bacteria but less in non-pathogenic bacteria. Furthermore, these genetically instable regions are mainly used as molecular tags for typing pathogenic strains, but their biological function and evolution mechanism are not often studied (van Belkum et al. 1998; van Belkum 1999).

The size of CRs can vary from one to one hundred nucleotides. Genes, and especially those encoding surface proteins, can contain numerous sets of repeats. In most cases, CRs are organized in direct tandem repeat and are mainly built with one type of motif (Fischetti et al. 1991; Dramsi et al. 1993; van Belkum et al. 1998). Exceptions, however, can be found such as the gene encoding the G protein from the GX7805 strain of group G *Streptococcus* that contains 4 different types of motifs that are organized as alternated interspersed repeats (Filpula et al. 1987). One of the most complex reported examples of CRs is found in the *cse* gene from *Streptococcus thermophilus* involved in Cell Segregation (Borges et al. 2005). It encodes an extracytoplasmic protein with one putative cell wall attachment (LysM) domain and a putative Cysteine Histidine-dependent Amidohydrolase/Peptidase (CHAP) domain. These two domain-encoding sequences are spaced by a region named *var-cse*. Analysis and comparison of the *var-cse* allele from two *S. thermophilus* strains (CNRZ368 and LMG18311) revealed that the *var-cse* sequence: (i) is almost completely built with numerous (up to 8) different types of small interspersed and tandem direct repeats, (ii) encodes an amino acid sequence of low complexity, and (iii) is highly divergent. The sequence of the two analyzed *S. thermophilus* strains diverge from approximately 40%. However, the comparison of two alleles only provided little information concerning the origin of this variability. Moreover, the replacement of the *var-cse* region of strain LMG18311 by that of strain CNRZ368 had little or no effect on cell segregation activity of Cse (Borges et al. 2005). This suggested that this region may be dispensable for cell segregation activity of Cse.

The aim of this study was to assess the extent of *var-cse* variability and to determine if this region is required for Cse activity. For this purpose, the sequence of *var-cse* alleles from numerous strains of *S. thermophilus* were analyzed and compared. In addition, the impact of a partial deletion of *var-cse* on cell segregation was investigated.

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, M17 (Terzaghi and Sandine 1975) or HJL (Stingele and Mollet 1996) media, without shaking. Milk medium was used for strain storage, M17 was used for mutant generation and HJL was employed for phenotypic analyses. Phenotypic analyses were performed at 42°C, the optimal *S. thermophilus* growth temperature. Erythromycin was added at 2 μg ml^{-1} when required. *S. thermophilus* derivative strains containing pGh9 (Maguin et al. 1992) plasmid derivatives were cultivated at 30°C when plasmid self-maintenance was required and at 42°C for selection of clones containing chromosomally integrated plasmid.

Recombinant plasmids derived from pGh9 were transformed into *Escherichia coli* EC101, a TG1 strain containing a chromosomal copy of the pWV01 *repA* gene (Buist et al. 1995), and were selected at 37°C on Luria-Bertani (LB) medium

Table 1 Strains and plasmids used in this study

Strain or plasmid	Relevant characteristic(s)	Source or reference
Strains		
Streptococcus thermophilus		
CNRZ368	Wild-type strain	INRA-CNRZ, strain collection
CNRZ385	Wild-type strain	INRA-CNRZ, strain collection
CNRZ445	Wild-type strain	INRA-CNRZ, strain collection
CNRZ702	Wild-type strain	INRA-CNRZ, strain collection
CNRZ308	Wild-type strain	INRA-CNRZ, strain collection
CNRZ7	Wild-type strain	INRA-CNRZ, strain collection
CNRZ1066	Wild-type strain	INRA-CNRZ, strain collection
CNRZ464	Wild-type strain	INRA-CNRZ, strain collection
CNRZ388	Wild-type strain	INRA-CNRZ, strain collection
CNRZ1402	Wild-type strain	INRA-CNRZ, strain collection
CNRZ302	Wild-type strain	INRA-CNRZ, strain collection
CNRZ307	Wild-type strain	INRA-CNRZ, strain collection
NST1	Wild-type strain	Laboratoire de Génétique et Microbiologie, Nancy
A054	Wild-type strain	(Mercenier et al. 1987)
ATCC19258	Wild-type strain	American Type Culture Collection
IP6757	Wild-type strain	Institut Pasteur strain collection
Sfi6	Wild-type strain	Nestlé strain collection
LMG18311	Wild-type strain	BCCM^{TM}/LMG, strain collection
CNRZ368-$\Delta_{67\text{-}411}$*var-cse*	*S. thermophilus* CNRZ368 with deletion of 345bp of *var-cse*.	This work
CNRZ368-Δ*cse*	*S. thermophilus* CNRZ368 with Δ*cse* mutation	(Borges et al. 2005)
Escherichia coli		
EC101	*supE hsd-5 thi* Δ(lac-proAB) F'*(traD6 proAB*$^{+}$ *lacI*^{q} *lacZΔM15) repA*$^{+}$, derivative of TG1 strain	(Leenhouts 1995) (Sambrook et al.1989)
DH5α	*supE*44 Δ*lac*U169 (*ϕ*80 *lacZ* ΔM15) *hsdR*17 *end*A1 *gyr*A96 *thi*-1 *rel*A1	(Sambrook et al. 1989)
Plasmids		
pGh9	Em^{R}, thermosensitive replication origin from pVE6002	(Maguin et al. 1992)
pGh9::$\Delta_{67\text{-}411}$*var-cse*	pGh9 with $\Delta_{129\text{-}411}$var-cse and CNRZ368 surrounding regions	This work

(Sambrook et al. 1989) containing 150 μg ml^{-1} of erythromycin.

DNA manipulations

Preparation of chromosomal and plasmid DNA, and Southern analysis were performed according to standard protocols (Sambrook et al. 1989). Sequences were aligned with CLUSTALW (Thompson et al. 1994). DOT PLOTs were constructed by using the Nucleic Acid Dot Plot from the Molecular Toolkit (http://www.arbl.cvmbs.colostate.edu/molkit/dnadot/index.html). The *cse* region encompassing *var-cse* was amplified by PCR using the primers 5′-CTGTAGTAGCAGAATCTAAC-3′ and 5′-GCACTAGCAATCCAGTCTT-3′. The reaction mixture was incubated at 95°C for 5 min. This was followed by 30 cycles of 45 s at 95°C, 45 s at 50°C , and 45 s at 72°C. After the last cycle, extension of the products was made at 72°C for 5 min. The *cse* locus was amplified by PCR using the following primers 5′-TTGTTGGCGCGCTTTAGCCGATTTGGCTTTTGAAG-3′ and 5′-TTGTTGGCGCGCCCCATTATTTTCTCAGGATGAAT-3′. Amplifications were performed as follow: one cycle of 5 min at 95°C followed by 10 cycles of 45 s at 95°C, 45 s at 50°C and 90 s at 72°C. This was followed by 20 cycles of 45 s at 95°C, 45 s at 60°C and 90 s at 72°C. At last, the reaction mixture was incubated 5 min at 72°C. Taq polymerase (Sigma) was used for amplification. To avoid the sequencing of PCR induced mutations, four independent PCR reactions were performed in parallel for each *S. thermophilus* strain, and the products were pooled before sequencing.

Analysis of *cse* mosaicism

A sequence identity matrix was elaborated, on the basis of sequence alignments of the *cse* locus, by using the BioEdit program (Hall 1999). One matrix was constructed for each of the A, B, D and E regions of the *cse* locus. The boundary separating the A and B regions was positioned on the 5′ side of the first variable nucleotide position that separates the sequences as groups I and II. The boundary between the D and E regions was positioned similarly. The boundaries flanking the C (or *var-cse*) region were positioned so that the highly variable repeat region is completely comprised in the C region. A mean and standard deviation were then calculated with the values resulting from the comparison of sequences that belong to the same group as well as from different groups. However, since group I was only represented by two sequences (from CNRZ368 and CNRZ445), the values presented for the group I does not correspond to a mean.

Construction of a $\Delta_{67\text{-}411}$*var-cse* mutant strain of *S. thermophilus*

The *var-cse* region from nucleotide 67 to 411 (with nucleotide 1 as the first 5′ nucleotide of the *var-cse* region) was deleted by allelic replacement as previously described (Thibessard et al. 2004). Briefly, the two regions flanking the *cse* locus to be deleted were independently amplified by PCR using the following primers pairs: 5′-CCCCCCAAGCTTTGAATGTTTTGGCTAATATC-3′ and 5′-CCGGAATTCTTCAGTTGTTTCAGTTGC-3′, 5′-CCGGAATTCCTAGCAGCTACATACGA-3′ and 5′-CCCCCTGCAGTTATGGATAAATATAATATAC-3′. The PCR products were digested by appropriate restriction enzymes, joined together and ligated with the plasmid pGh9. The ligation products were used to transform *E. coli* EC101. After introduction of the recombinant plasmid into *S. thermophilus*, deleted mutants displaying an EcoRI restriction site instead of the 345 bp of the *var-cse* region were selected as previously described (Thibessard et al. 2004). The *var-cse* partial deletion was checked by PCR, Southern hybridization and sequencing (data not shown).

Microscopy

Cells were observed with a Nikon OPTIPHOT microscope mounted with phase contrast equipment "Ph" at ×100 magnification, using the condenser turret at position Ph4, or at magnification ×1000 by phase contrast.

Nucleotide sequence accession numbers

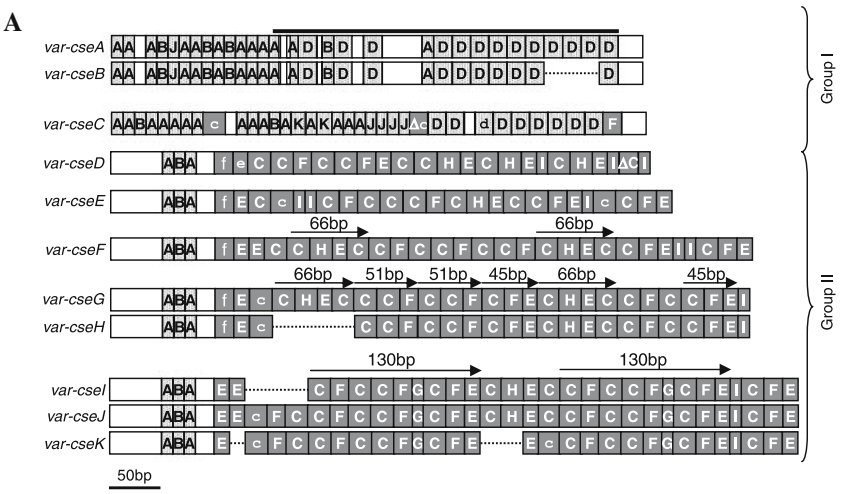
DNA sequences reported in this paper have been deposited in GenBank under accession numbers

AY730642, AY695844, CP000024 and from DQ102334 to DQ102345.

Results

Structure of the *var-cse* alleles

In order to characterize the genetic variability of *var-cse*, the corresponding sequence of 15 *S. thermophilus* strains was determined. For this purpose, a region of the *cse* ORF comprising *var-cse* was amplified by PCR from genomic DNA. A single PCR product was obtained for each strain and was sequenced. Since the sequence of the *cse* locus of 4 strains (CNRZ368, LMG18311, CNRZ1066 and LMD-9) was already available (Bolotin et al. 2004; Borges et al. 2005; http://www.genome.jgi-psf.org/draft_microbes/strth/strth.home.html), a total of 19 sequences were analyzed. Among these sequences, 11 alleles (named *var-cseA* to *var-cseK*) differing by their sequence and their size were



B

Motif name	Nucleic acid sequence consensus[a]	Amino acid sequence consensus
A	GAA ACA WCW	E T (S/T)
B	GCA CAA GCT	A Q A
C	TCA GAD GYD CCW GMD RDM	S (E/D) (A/V) P (A/E) (E/S/V)
D	GAA GYW CYA KYW GAA	E (A/V) (P/L) (A/V/S) E
E	TCA GAA GCT GYA	S E A (A/V)
F	CAA CYA GCA GCD RCA	Q (L/P) A A (A/T)
G	CCA GAG AGC	P E S
H	GTA GYA CCT GTC GCA ACA	V (A/V) P V A T
I	CCR GCT DYW	P A (T/V/A/S)
J	GAA GMA ACT	E (A/E) T
K	GAA GAG GCA TCA	E E A S

[a] W : A/T ; R : A/G ; Y : C/T ; M : A/C ; K : G/T ; D : A/G/ T

Fig. 1 Repetitive structure of *var-cse* alleles. (**A**) Schematic representation of the *var-cse* region of the following *S. thermophilus* strains: CNRZ368 NST1 and CNRZ385 (*var-cseA* allele); A054 (*var-cseB* allele); CNRZ445 (*var-cseC* allele); ATCC19258 (*var-cseD* allele); CNRZ702 (*var-cseE* allele); CNRZ308 (*var-cseF* allele); IP6757, CNRZ7, CNRZ1066 and Sfi6 (*var-cseG* allele); CNRZ464 (*var-cseH* allele); LMG18311 and CNRZ388 (*var-cseI* allele); LMD9 and CNRZ1402 (*var-cseJ* allele); CNRZ302 and CNRZ307 (*var-cseK* allele). Each repeat unit is represented by a letter-containing box (referring to part B). Repeat units in accordance with the consensus sequences presented in panel B are represented in uppercase letters and those that are slightly divergent from the consensus are represented by lowercase letters. The Δ symbol refers to partially deleted repeats. Repeats that are characteristic of group I and group II are represented by light colored grey and dark grey rectangles respectively. Aligned sequences are represented by bracketing one above the other the schematic representations of the sequences aligned by CLUSTALW. Dotted lines indicate gaps in the alignments. Notice that because of the repeat richness of these regions, the positioning of gaps can not be favored between two possibilities and so representation of one of them was arbitrarily chosen. The repeated supermotifs corresponding to nucleotide duplications greater than 45 bp with less than 1% degeneracy are indicated by arrows. The region of *var-cseA* deleted in the CNRZ368-Δ67-411*var-cse* mutant strain is indicated by a thick line. (**B**) Table listing the consensus sequences of the repeated motifs identified in the *var-cse* alleles

detected. The size of the *var-cse* sequences varies from 390 to 543 bp, i.e. a maximum size variation of 153 bp was found. The repeat content was assessed for each allele. This analysis revealed that they are almost all built-up of direct repeats, with tandem and dispersed arrangement (Fig. 1A). A total of 11 different repeat motifs (named A to K), with sizes from 9 to 18 bp, were detected (Fig.1B). All of the repeated units are multiples of three and are thus in frame. Consequently, all the nucleic acid repeat sequences correspond to an amino acid repeat in Cse (Fig.1B). The sequence of 8 of these 11 motifs (A, C, D, E, F, H, I, and J) are degenerated. The most degenerated is the most frequent C motif, with 9 degenerated positions among 18.

Consensus sequence comparison of the motifs showed that some of them are similar to parts of others. For example, the nucleotide sequence encoding the PES part of the SEAPES amino acid sequence (one possible C motif) is identical to the G motif, whose corresponding amino acid sequence is also PES (Fig. 2).

Nucleotide sequence analysis by dot plot indicated that the *var-cse* alleles can contain repeats larger than 18 bp, corresponding to the size of the largest repeats described Fig. 1B. Indeed, the *var-cseF* and *var-cseG* alleles contain a perfectly duplicated 66 bp region (Fig. 1A). This region corresponds to the CHEC series of motifs, or supermotif. This supermotif is also present in the other alleles but has degenerated. Other examples are the CCF (51 bp) and CFE (45 bp) supermotifs that are perfectly duplicated in *var-cseG* and *var-cseH* (Fig. 1A). The largest duplication identified is 130 bp long, comprising the CFCCFGCFE motif block. This region is duplicated almost perfectly (the two copies exhibits one divergent nucleotide) in *var-cseI* to *K* (Fig. 1A). In conclusion, the small repeated motifs can be organized as repeated supermotifs.

Motif name	Amino acid sequence	Nucleotide sequence
C:	SEA**PAA**	TCAGAAGCA**CCAGCTGC**A
I:	**PAA**	**CCAGCTGC**T
C:	SEA**PES**	TCAGAGGCA**CCAGAGAGC**
G:	**PES**	**CCAGAGAGC**
C:	**SEA**PA	**TCAGAAGCT**CCTGCTAGC
E:	**SEA**A	**TCAGAAGCT**GCA
E:	**SEA**V	**TCAGAAGCT**GTA
C:	S**EAPAE**	TCA**GAAGC**G**CCAGCAGAA**
D:	**EAPAE**	**GAAGC**A**CCAGCAGAA**

Fig. 2 Similarity between repeat motifs. Similar repeat motifs were aligned. Identical amino acids or nucleotides are indicated by bold face

Features of *var-cse* intraspecific variability

Alignment of *var-cse* alleles was obtained using CLUSTALW. Taking into account their high divergence (nucleotide or amino acid), only the most related sequences could be aligned and are represented in Fig. 1A. Related sequences differed only by their repeat copy number. For instance, the D motif is found 13 times in *var-cseA* but only 10 times in *var-cseB*. Additionaly, the number of supermotifs can also differ from one allele to the other. Indeed, *var-cseG* and *var-cseH* differ only by one copy of the CHEC supermotif.

In addition, the *var-cse* alleles can be classified in two groups on the basis of their repeat type content. The group I, bringing together *var-cseA* to *C*, comprises all of the D, J and K motifs, and the majority (74%) of the A motifs (Fig. 1A). Group II, bringing together *var-cseD* to *K*, comprises all of the E, G, H, and I motifs, and the majority (98%) of the C and F motifs.

cse is mosaic

Two alternative hypotheses could explain the origin of the *var-cse* alleles in these two groups. First, the common ancestor of the two corresponding groups of *S. thermophilus* strains is ancient each has since evolved by severe mutation accumulation. Alternatively, a *var-cse* allele was acquired in *S. thermophilus* by horizontal sequence transfer resulting in a mosaic *cse* locus.

To test these hypotheses, we compared the entire *cse* ORF and flanking regions from 10 *S. thermophilus* strains belonging to group I (CNRZ368, CNRZ445) and group II (LMG18311, LMD9, ATCC19258, CNRZ1066, CNRZ302, CNRZ308, CNRZ702, CNRZ1402). Sequence alignments revealed that the regions encompassing the 436 bp B region and the 228 bp D region upstream and downstream *var-cse* respectively, could be classified into group I and II, on the criterion of sequence identity. Indeed,

sequence identity matrixes showed that along the B and D regions, percentage sequence identities are much higher between sequences belonging to the same group than between sequences that belong to different groups (Fig. 3). On the other hand, along the regions flanking BCD, *i.e.* the A and E regions, all the sequences are almost identical. Therefore, it is within the B, C (or *var-cse*) and D regions that the sequences can be classified into two groups. We conclude that these results show that the *cse* locus is mosaic.

Involvement of *var-cse* in cell segregation

The high intraspecific variability of the *var-cse* region could suggest that it is a dispensable region of the gene. As Cse is involved in the separation of *S. thermophilus* cells (Borges et al. 2005), the involvement of the *var-cse* region in cell segregation was investigated.

For this purpose, 80% of the *var-cse* region was deleted in-frame in the CNRZ368 strain (Fig. 1A) using an allelic replacement strategy. The resulting CNRZ368-$\Delta_{67\text{-}411}$*var-cse* strain was grown, in parallel with CNRZ368 and CNRZ368-Δ*cse*, in liquid medium without shaking. The CNRZ368-Δ*cse* strain is a *cse* null mutant that was previously constructed by in-frame deletion of the *cse* ORF (Borges et al. 2005).

Phase-contrast photonic microscopy observation revealed that the CNRZ368-$\Delta_{67\text{-}411}$*var-cse* cell chains were shorter than those of the Δ*cse* mutant (Fig. 4A). This was reinforced by counting the number of cells per chain for each strain. All the cells of the CNRZ368-$\Delta_{67\text{-}411}$*var-cse* mutant formed chains of less than 351 cells, while more than 99% of the CNRZ368-Δ*cse* mutant cells were found in chains of more than 1000 cells (Fig. 4B). This result shows that a residual cell segregation activity is encoded by the *cse* mutant gene devoid of the $_{67\text{-}411}$*var-cse* region. On the other hand, 80% and 23% of the WT and the CNRZ368-$\Delta_{67\text{-}411}$*var-cse* mutant cells, respectively, compose chains of 1 to 50 cells (Fig. 4B). Moreover, none of the WT chains contained more than 151 cells, yet 25% of mutant chains exhibited this length or longer. Thus, the CNRZ368-$\Delta_{67\text{-}411}$*var-cse* chains are typically longer than those of the WT, indicating that the mutant with a partial deletion of the *var-cse* region is impaired in cell segregation. These data show that the *var-cse* region is required for full cell segregation in *S. thermophilus* strain CNRZ368.

Discussion

The *var-cse* variability: a possible consequence of repeat richness

Analysis of the *var-cse* sequence from 19 *S. thermophilus* strains has revealed the extent of the intraspecific variability of this *cse* region. This variability is emphasized when considering that *S. thermophilus* strains are closely related (Bolotin et al. 2001; P. Renault, personal communication).

The *var-cse* polymorphism can result from repeat number variability. Such variability was

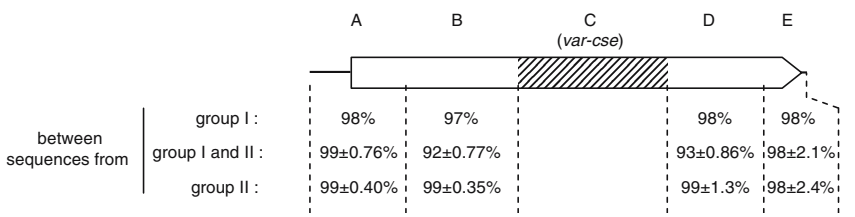


Fig. 3 The *cse* locus is mosaic. The sequences of the *cse* locus were aligned with CLUSTALW. Based on the alignment, a sequence identity matrix was built and the mean value and standard deviation corresponding to comparison between sequences from group I, sequences from group I and group II, and sequences from group II was reported for regions A, B, D and E of the *cse* locus. Values without standard deviation resulted from comparison of two sequences. Such analysis is not reported for the region 3 (corresponding to *var-cse*) since alignment is not valuable for this region. The open arrow represent the *cse* ORF and indicate its reading direction

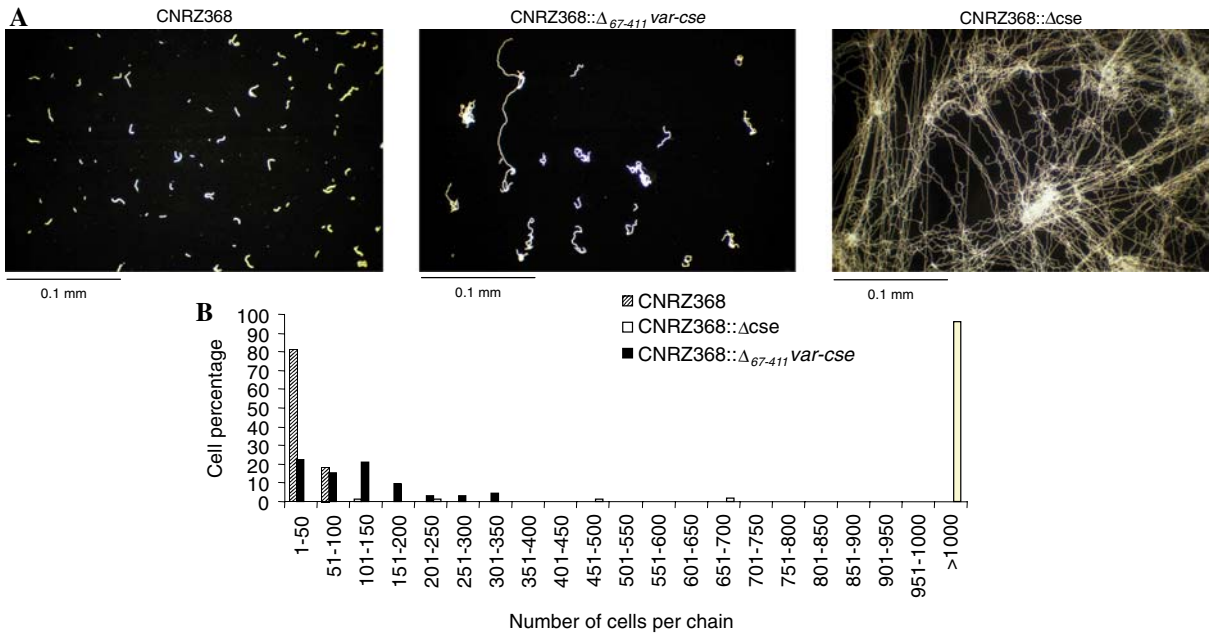


Fig. 4 Influence of *var-cse* partial deletion on cell segregation activity. (**A**) Cell chain morphology of *S. thermophilus* strains. (**B**) Results from counting the number of cells per chain. At least 1000 cells were counted from three independent cultures. Cells were photographed and counted in stationary phase after 20 h of growth

already widely described for repeat regions (Yang and Gabriel 1995; van Belkum et al. 1998; Wilton et al. 1998; Lukomski et al. 2000; Schupp et al. 2000; Rasmussen and Bjorck 2001; Areschoug et al. 2002; Koroleva et al. 2002; Karatzas et al. 2003; de Benito et al. 2004). It is likely that insertions/deletions (indels) of repeat units result from polymerase slippage, during DNA replication, or by single strand annealing (SSA) (Michel 1999). These two mechanisms of illegitimate recombination imply pairing of ectopic complementary sequences. Sequence alignment of *varcseI, J, K* suggests that an indel of an E motif occurred (Fig. 1A). This indel would result from ectopic pairing of two sequences, each one constituting a whole E repeat unit. Interestingly, considering the similarities between repeat motifs, it is likely that illegitimate recombination events are implicated in pairing of sequences from different repeat motifs. This hypothesis is supported by our *var-cseI, J, K* alignment. This alignment revealed an indel of a CH unit (Fig.1A), potentially resulting from pairing of the 5′ end of the inserted/deleted C motif, with the 5′ end of the E motif downstream the CH indel (Fig. 1A). Thus, the *var-cse* region could have evolved by recombination between identical repeats but also between different repeats. The ability of different repeats to undergo recombination would have increased the variety of recombination events in *var-cse* and could therefore be a factor responsible for the high polymorphism of this region.

Origin of the diversity of the repeats

Partial pairing of repeated motifs could be responsible for indels of blocks of whole repeats units, but also they could induce partial deletion of repeat units. Indeed, since the G motif is entirely comprised within the C repeat (Fig. 2), and taking into account the abundance of the latter, a simple hypothesis is that the G motif has been generated by partial deletion of an ancestral C motif (Fig. 5). A probable scenario would implicate an ectopic pairing of the GCA trinucleotide repeats: one constituting the 3′ end of the F repeat unit upstream of the hypothetical C repeat unit and the other encoding the alanine amino acid of the C hypothetical repeat unit (Fig. 5).

This ectopic pairing would have generated the deletion of the SEA encoding region of the hypothetical C repeat unit, generating the G motif. It is notable that at the end of this scenario, the G motif is not yet a repeat.

The alleles can contain large nucleotide duplications. Three alleles, grouping together the *var-cse* sequence of 11 *S. thermophilus* strains, exhibit an interspersed duplication of the CFCCFGCFE block of motifs (Fig.1). Two hypotheses could explain this duplication: that the same succession of repeat motifs appeared independently twice following multiple indels or, alternatively, that a single mutation resulted in the duplication of one of the two copies. Taking into account the high conservation of the CFCCFGCFE supermotif (1 nucleotide divergence between the 2 copies along 130 bp), it is more probable that the CFCCFGCFE duplication resulted from a unique mutation. Thus, the *var-cse* region would have evolved by duplication of motif blocks.

Interestingly, the CFCCFGCFE duplication would have duplicated the enclosed G motif that would have been created by partial deletion of a C hypothetical motif. This hypothesis would explain how the G motif, initially generated as a single copy, was duplicated and thus became a repeat. This hypothesis is supported by the observation that all the G motifs identified here are included in the CFCCFGCFE supermotifs (Fig. 1A). A similar mechanism of repeat generation by recombination between existing interspersed repeats has already been suggested for the G protein from two strains of Group G streptococci (Fahnestock et al. 1986; Filpula et al. 1987; Kehoe 1994).

In conclusion, a mutation series consisting of partial deletion of repeats followed by supermotif duplication could be responsible in part for the high repeat variety of the *var-cse* region.

Implication of horizontal transfer

Comparison of the *var-cse* region from 19 *S. thermophilus* strains and the flanking regions of 10 of them revealed that *cse* is mosaic. The sequences from group I exhibit 7% nucleotide divergence with the sequences from group II, along the B and C regions of *cse*, regions that are not supposed to evolve rapidly since they are not composed of repeats. It was previously reported that the divergence average between *S. thermophilus* strains sequences is 0.2% (Bolotin et al. 2001; P. Renault, personal communication). Therefore, the intraspecific divergence noticed in the *cse* locus is unusual for *S. thermophilus* sequences. Numerous mosaic genes have been previously described, especially in streptococci (Whatmore and Kehoe 1994; Kapur et al. 1995; Whatmore and Dowson 1999; Filipe et al. 2000; Hakenbeck 2000). These genes are typically formed by allelic recombination between a genomic locus and another allele acquired from horizontal transfer. It is likely that the *cse* mosaic nature has resulted from this mechanism.

Extent of the tolerance for var-cse variability regarding the *cse* cell segregation activity

Our partial *var-cse* deletion led to a decrease of cell segregation activity. Although the *var-cse* region is required for full cell segregation activity,

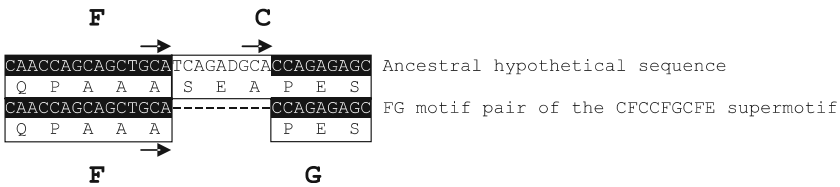


Fig. 5 Sequence alignment illustrating the hypothetical mechanism resulting in the origin of the G motif. Identical nucleotides between aligned sequences are represented by highlighted white letters. The trinucleotide sequence duplicated in the hypothetical ancestral sequence and not duplicated in the actual sequence is indicated by an arrow. The hypothesis implies the generation of the G motif by partial deletion of the hypothetical and ancestral C motif. Partial deletion would have occurred by illegitimate recombination implicating an ectopic pairing of two copies of the duplicated GCA sequence

it exhibits high variability. This suggests that the function fulfilled by *var-cse* does not lead to counter-selection against this variability and has allowed the genomic fixation of a large number of alleles. This hypothesis is supported by results from *var-cse* allelic replacement experiments (Borges et al. 2005). Indeed, the replacement of the *var-cse* sequence of strain LMG18311 by that from CNRZ368, which results in shortening by 60 bp and the alteration of approximately 60% of the remaining amino acid sequence, did not have significant consequence on chain length. However, since the $\Delta_{67\text{-}411}$*var-cse* mutant is affected in cell segregation, this variability tolerance could be restricted to a minimal size of the *var-cse* region. Thus, the role of the *var-cse* region could be to maintain an appropriately orientated CHAP domain within the cell envelope i.e. with the Cse protein anchored by the LysM domain and the *var-cse* region acting as a spacer. This interpretation is consistent with the results from our previous *var-cse* allelic replacement experiments (Borges et al. 2005) since these would be predicted to maintain an appropriately displayed CHAP domain in the Cse constructs.

Analysis of the sequenced streptococcal genomes revealed two other LysM-CHAP domain proteins: SMU_367 from *Streptococcus mutans* and BAB61101 from *Streptococcus intermedius*, whose functions are unknown. In SMU_367 and BAB61101 proteins, the LysM and the CHAP domains are spaced by a short sequence of 31 amino acids and 36 amino acids, respectively. For Cse, the shortest linker sequence identified was that of *S. thermophilus* strain A054 which is 236 amino acids long, about seven times longer than the spacer in the other two LysM-CHAP proteins. Thus, the sequences spacing LysM and CHAP domains are not always long, suggesting that the functionality of this combination of domains does not necessarily require a long interdomain linker.

Acknowledgements We thank P. Renault (Institut National de la Recherche Agronomique, Jouy-en-Josas, France) for personal communications concerning the relatedness of *S. thermophilus* strains. S.L. and A.F. were supported by grants from the Ministère de l'Education Nationale de l'Enseignement Supérieur et de la Recherche. F.B. was supported by a grant from the Institut National de la Recherche Agronomique.

References

Areschoug T, Linse S, Stalhammar-Carlemalm M, Heden LO, Lindahl G (2002) A proline-rich region with a highly periodic sequence in Streptococcal beta protein adopts the polyproline II structure and is exposed on the bacterial surface. J Bacteriol 184:6376–6383

Bolotin A, Quinquis B, Renault P, Sorokin A, Ehrlich SD, Kulakauskas S, Lapidus A, Goltsman E, Mazur M, Pusch GD, Fonstein M, Overbeek R, Kyprides N, Purnelle B, Prozzi D, Ngui K, Masuy D, Hancy F, Burteau S, Boutry M, Delcour J, Goffeau A, Hols P (2004) Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*. Nat Biotechnol 22:1554–1558

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

Borges F, Layec S, Thibessard A, Fernandez A, Gintz B, Hols P, Decaris B, Leblond-Bourget N (2005) cse, a Chimeric and variable gene, encodes an extracellular protein involved in cellular segregation in Streptococcus thermophilus. J Bacteriol 187:2737–2746

Buist G, Kok J, Leenhouts KJ, Dabrowska M, Venema G, Haandrikman AJ (1995) Molecular cloning and nucleotide sequence of the gene encoding the major peptidoglycan hydrolase of *Lactococcus lactis*, a muramidase needed for cell separation. J Bacteriol 177:1554–1563

de Benito I, Cano ME, Aguero J, Garcia Lobo JM (2004) A polymorphic tandem repeat potentially useful for typing in the chromosome of *Yersinia enterocolitica*. Microbiology 150:199–204

Dramsi S, Dehoux P, Cossart P (1993) Common features of gram-positive bacterial proteins involved in cell recognition. Mol Microbiol 9:1119–1121

Fahnestock SR, Alexander P, Nagle J, Filpula D (1986) Gene for an immunoglobulin-binding protein from a group G streptococcus. J Bacteriol 167:870–880

Filipe SR, Severina E, Tomasz A (2000) Distribution of the mosaic structured *murM* genes among natural populations of *Streptococcus pneumoniae*. J Bacteriol 182:6798–6805

Filpula D, Alexander P, Fahnestock SR (1987) Nucleotide sequence of the protein G gene from *Streptococcus* GX7805, and comparison to previously reported sequences. Nucleic Acids Res 15:7210

Fischetti VA, Pancholi V, Schneewind O (1991) Common characteristics of the surface proteins from gram-positive cocci. In: Dunny GM, Cleary PP, Mc Kay LL (eds) Genetics and Molecular Biology of Streptococci, Lactococci, and Enterococci. ASM press, pp 290–294

Hakenbeck R (2000) Transformation in *Streptococcus pneumoniae*: mosaic genes and the regulation of competence. Res Microbiol 151:453–456

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:95–98

Hallet B (2001) Playing Dr Jekyll and Mr Hyde: combined mechanisms of phase variation in bacteria. Curr Opin Microbiol 4:570–581

Kapur V, Kanjilal S, Hamrick MR, Li LL, Whittam TS, Sawyer SA, Musser JM (1995) Molecular population genetic analysis of the streptokinase gene of *Streptococcus pyogenes*: mosaic alleles generated by recombination. Mol Microbiol 16:509–519

Karatzas KA, Wouters JA, Gahan CG, Hill C, Abee T, Bennik MH (2003) The CtsR regulator of *Listeria monocytogenes* contains a variant glycine repeat region that affects piezotolerance, stress resistance, motility and virulence. Mol Microbiol 49:1227–1238

Kehoe MA (1994) Cell-wall-associated proteins in Gram-positive bacteria. In: Hakenbeck J-MGaR (ed) Bacterial Cell Wall. Elsevier Science B.V, pp 217–261

Koroleva IV, Efstratiou A, Suvorov AN (2002) Structural heterogeneity of the streptococcal C5a peptidase gene in *Streptococcus pyogenes*. J Bacteriol 184:6384–6386

Leenhouts K (1995) Integration strategies and vectors. Dev Biol Stand 85:523–530

Lukomski S, Nakashima K, Abdi I, Cipriano VJ, Ireland RM, Reid SD, Adams GG, Musser JM (2000) Identification and characterization of the *scl* gene encoding a group A *Streptococcus* extracellular protein virulence factor with similarity to human collagen. Infect Immun 68:6542–6553

Maguin E, Duwat P, Hege T, Ehrlich D, Gruss A (1992) New thermosensitive plasmid for gram-positive bacteria. J Bacteriol 174:5633–5638

Mercenier A, Robert C, Romero DA, Slos PYL (1987) Transfection of *Streptococcus thermophilus* spheroblasts. In: Ferreti JJ, Curtiss R (eds) Streptococcal genetics. ASM press, pp 234–237

Michel B (1999) Illegitimate recombination in bacteria. In: Charlebois RL (ed) Organization of the prokaryotic genome. American Society for Microbiology, pp 129–150

Rasmussen M, Bjorck L (2001) Unique regulation of *SclB*—a novel collagen-like surface protein of *Streptococcus pyogenes*. Mol Microbiol 40:1427–1438

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning : a laboratory manual. Cold Spring Harbor Laboratory Press

Schupp JM, Klevytska AM, Zinser G, Price LB, Keim P (2000) *vrrB*, a hypervariable open reading frame in *Bacillus anthracis*. J Bacteriol 182:3989–3997

Stingele F, Mollet B (1996) Disruption of the gene encoding penicillin-binding protein 2b (*pbp2b*) causes altered cell morphology and cease in exopolysaccharide production in *Streptococcus thermophilus* Sfi6. Mol Microbiol 22:357–366

Terzaghi B, Sandine W (1975) Improved medium for lactic streptococci and their bacteriophages. Appl environ microbiol 29:807–813

Thibessard A, Borges F, Fernandez A, Gintz B, Decaris B, Leblond-Bourget N (2004) Identification of *Streptococcus thermophilus* CNRZ368 Genes Involved in Defense against Superoxide Stress. Appl Environ Microbiol 70:2220–2229

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 22:4673–4680

van Belkum A (1999) Short sequence repeats in microbial pathogenesis and evolution. Cell Mol Life Sci 56:729–734

van Belkum A, Scherer S, van Alphen L, Verbrugh H (1998) Short-sequence DNA repeats in prokaryotic genomes. Microbiol Mol Biol Rev 62:275–293

Whatmore AM, Dowson CG (1999) The autolysin-encoding gene (*lytA*) of *Streptococcus pneumoniae* displays restricted allelic variation despite localized recombination events with genes of pneumococcal bacteriophage encoding cell wall lytic enzymes. Infect Immun 67:4551–4556

Whatmore AM, Kehoe MA (1994) Horizontal gene transfer in the evolution of group A streptococcal *emm-like* genes: gene mosaics and variation in Vir regulons. Mol Microbiol 11:363–374

Wilton JL, Scarman AL, Walker MJ, Djordjevic SP (1998) Reiterated repeat region variability in the ciliary adhesin gene of *Mycoplasma hyopneumoniae*. Microbiology 144(Pt 7):1931–1943

Yang Y, Gabriel DW (1995) Intragenic recombination of a single plant pathogen gene provides a mechanism for the evolution of new host specificities. J Bacteriol 177:4963–4968