J Mol Evol (2008) 67:13–22
DOI 10.1007/s00239-008-9117-1

Evolution of the Hyaluronic Acid Synthesis (*has*) Operon in *Streptococcus zooepidemicus* and Other Pathogenic Streptococci

Lars M. Blank · Philip Hugenholtz · Lars K. Nielsen

Received: 28 January 2008 / Accepted: 29 April 2008 / Published online: 13 June 2008
© Springer Science+Business Media, LLC 2008

Abstract Hyaluronic acid (HA) is a ubiquitous linear polysaccharide in vertebrates and also is the capsule material of some pathogenic bacteria including group A and C streptococci. In bacteria, the HA synthase occurs in an operon (*has*) coding for enzymes involved in the production of HA precursors. We report two new members of the *has* operon family from *Streptococcus equi* subsp. *zooepidemicus* (*S. zooepidemicus*) and *Streptococcus equi* subsp. *equi* (*S. equi*). The *has* operon of *S. zooepidemicus* contains, in order, *hasA*, *hasB*, *hasC*, *glum*, and *pgi*, whereas these genes are separated on two operons in *S. equi* (*hasA*, *hasB*, *hasC* and *hasC*, *glmU*, *pgi*). The transcription start site and a σ^{70} promoter were experimentally identified 50 bp upstream of *hasA* in *S. zooepidemicus*. We performed a phylogenetic analysis of each of the *has* operon genes to determine the evolutionary origin(s) of the streptococcal *has* operon. In contrast to other capsular and exopolysaccharide operons, *has* operons have undergone no detectable interspecies lateral gene transfers in their construction, instead relying on intragenome gene duplication for their assembly. Specifically, *hasC* and *glmU* appear to have been duplicated into the *S. zooepidemicus has* operon from remotely located but near-identical paralogues most likely to improve HA productivity by gene dosage in this streptococcus. The intragene rearrangements appear to be ongoing events and the two *has* operons of the *S. equi* subspecies represent two alternatives of the same gene arrangement. A scenario for the evolution of streptococcal *has* operons is proposed.

Keywords Hyaluronic acid production · Lateral gene transfer · Gene dosage · Duplicate genes · Homologous recombination · Group C

Electronic supplementary material The online version of this article (doi:10.1007/s00239-008-9117-1) contains supplementary material, which is available to authorized users.

L. M. Blank · L. K. Nielsen
Department of Chemical Engineering, The University of Queensland, Brisbane, Australia

P. Hugenholtz · L. K. Nielsen
ComBinE Group, Advanced Computational Modeling Centre, The University of Queensland, Brisbane, Australia

Present Address:
L. M. Blank (✉)
Faculty of Biochemical and Chemical Engineering, TU Dortmund, Dortmund, Germany
e-mail: lars.blank@bci.tu-dortmund.de

Present Address:
P. Hugenholtz
Microbial Ecology Program, DOE Joint Genome Institute, Walnut Creek, CA, USA

Abbreviations

hasA	Hyaluronate synthase
hasB	UDP-glucose dehydrogenase
hasC	glucose-1-phosphate uridyltransferase
glmU	N-acetyl-glucosamine-1-phosphate uridyltransferase
pgi	glucose-6-phosphate isomerase

Introduction

Hyaluronic acid (HA) is an important polymer in vertebrates, fulfilling a large range of functions such as lubrication, water binding, and cell trafficking (Lee and Spicer 2000). HA is also found as capsule material of pathogenic group A and C streptococci, where it camouflages the bacteria against the

host immune system (Wibawan et al. 1999). The two precursors for this linear polymer are UDP-glucuronic acid and UDP-*N*-acetylglucosamine, synthesized from the glycolysis intermediates glucose-6-phosphate and fructose-6-phosphate, respectively (Fig. 1).

Like other β-polysaccharide synthases, such as cellulose and chitin synthase, the HA synthase is a multifunction, membrane-bound enzyme that assembles the polymer directly from the UDP sugars and extrudes it through the plasma membrane (Kamst and Spaink 1999; Tlapak-Simmons et al. 1999). In contrast, extracellular and capsular polysaccharide production in other streptococci and most pathogenic bacteria is performed by an indirect mechanism, in which the repeating unit is assembled from UDP sugars on an isoprenoid lipid carrier by the action of specific glycosyl transferases, most likely transported across the cell membrane before being polymerized by a membrane-associated polymerase (De Vuyst et al. 2001).

The HA synthase (*hasA*; hyaluronate synthase) was first identified in *Streptococcus pyogenes* (DeAngelis et al. 1993), where it is part of a polycistronic transcription unit that also encodes for two additional genes involved in UDP-glucuronic acid synthesis, namely, UDP-glucose dehydrogenase (*hasB*) and glucose-1-phosphate uridyltransferase (*hasC*) (Dougherty and van de Rijn 1994). More recently, the *has* operon from *Streptococcus uberis* was cloned and found to encode *hasA* and *hasB* but not *hasC* (Ward et al. 2001). Here we characterize the *has* operons of two equine pathogens, *Streptococcus equi* subsp. *zooepidemicus* and *Streptococcus equi* subsp. *equi*, used in the commercial production of HA (Chong et al. 2005).

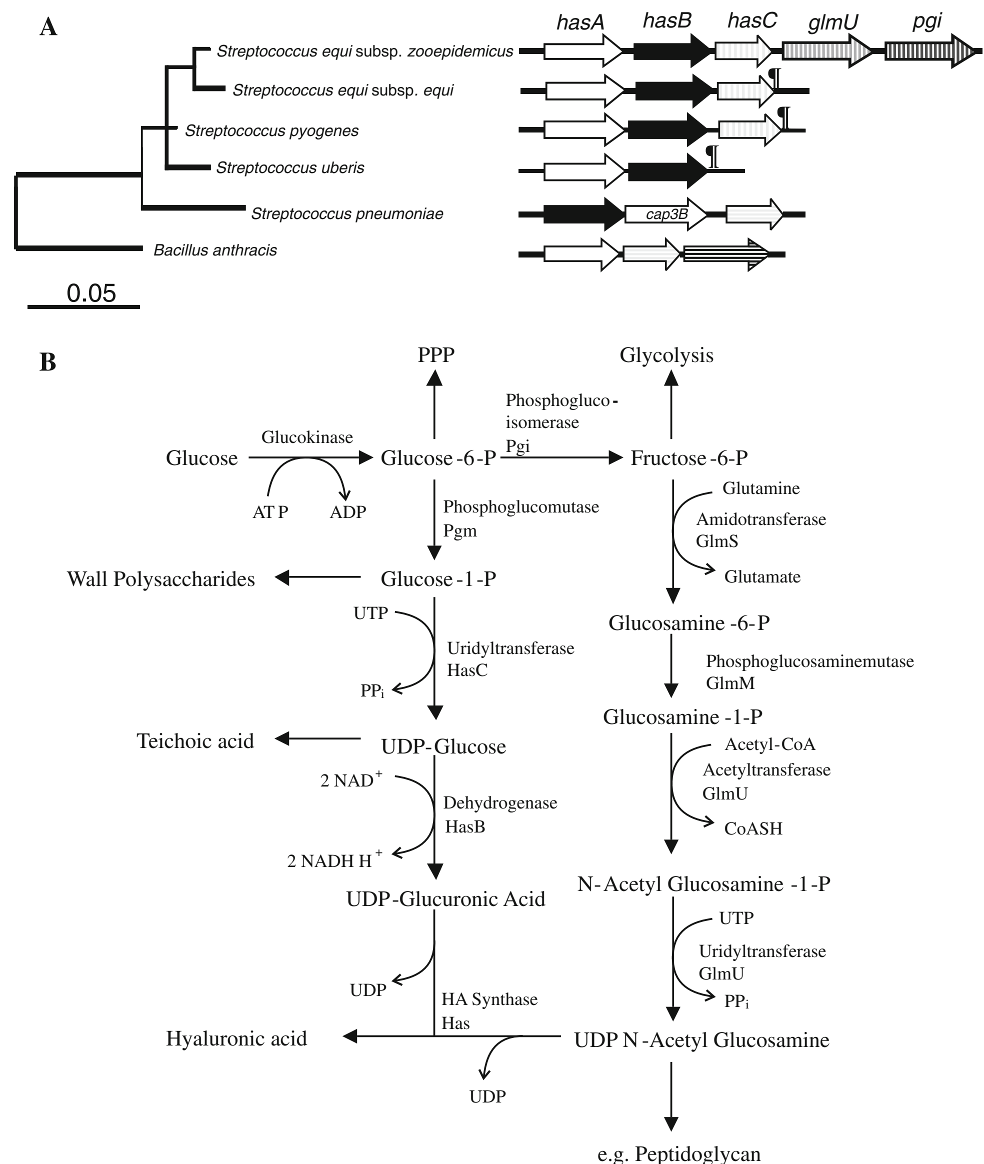


Fig. 1 (**A**) Gross operon structure of *Firmicutes* containing *hasA* homologues. The identity of individual genes is given at the top and indicated throughout by shading. Vertical lines indicate that the gene originated via intragenomic gene duplication; horizontal lines, that the gene was introduced via lateral gene transfer. Putative transcription terminator sequences are indicated by ¶. The relationships of the species are indicated by the evolutionary distance dendrogram (Kimura two-parameter model) at the left as determined by comparative 16S rRNA analysis. The scale bar indicates 0.05 change per nucleotide. (**B**) Biosynthetic pathways involved in HA production. Proteins encoded by the *has* operon from *S. zooepidemicus* are in boldface

Lateral gene transfer (LGT) has been extensively implicated in the assembly of the extracellular and capsular polysaccharide operons in streptococci (Bentley et al. 2006; Chaffin et al. 2002; Garcia et al. 2000; Mavroidi et al. 2004). Although the *has* operon has not specifically being investigated for LGT, the absence of nonvertebrate eucaryal HasA homologues led to the inclusion of HA synthase in a list of 223 putative LGTs from bacteria to vertebrates by the Human Genome Sequencing Consortium (Lander et al. 2001). We performed a rigorous phylogenetic analysis of genes involved in HA synthesis to determine the role of LGT in *has* operon evolution and propose a scenario for the evolutionary history of this operon in streptococci.

Materials and Methods

Primers

Oligonucleotide primers were synthesized by Gibco BRL and Sigma Genosys. Medium components were obtained from Difco and Oxoid.

Bacterial Strains and Plasmids

The mucoid Group C *Streptococcus equi* subsp. *zooepidemicus* ATCC 35246 was obtained from the American Type Culture Collection (Rockville, MD, USA). A nonmucoid (HA-negative) strain was isolated on sheep blood agar plates. The *E. coli* host strain STBL4™ was obtained from Life Technologies. Streptococci were grown in M17G, and *E. coli* in LB medium. The pGEM-T (Promega) cloning vector kit was used in this study.

DNA Purification and Manipulations

High molecular mass genomic DNA from *S. zooepidemicus* was isolated by the method of Anderson and McKay (1984) with an additional hyaluronidase (final concentration, 200 U/ml) digest to remove capsular material prior to alkaline lysis. DNA was then purified with the Tissue spin column kit (Qiagen). Plasmid DNA was purified from *E. coli* cultures (3 ml) with Wizard Mini-Preps (Promega) or Mini kit (Qiagen) and resuspended in a final volume of 50 µl in water.

Amplification and Sequencing of the *has* Operon

To amplify the 5′ end of *S. zooepidemicus hasA* including the promoter region, chromosome walking was performed (Siebert et al. 1995). Five restriction enzymes (*Dra*I, *Eco*RV, *Pvu*II, *Ssp*I, *Xba*I) were used for the generation of *S. zooepidemicus* chromosomal DNA fragment libraries. Two primers were designed based on the partial *hasA* sequence obtained via degenerate PCR according to Kumari and Weigel (1997) (P3, 5′-TCA AGG GCT GTA GGA CAA ACA AAT G-3′; P4, 5′-GCA GTA CCC CAT AAA ACG GAG ATA AC-3′). An ~1.4-kb DNA fragment was amplified by nested PCR using primer combinations P3+Ap1 and P4+Ap2 (Ap1 and Ap2 according to Siebert et al. [1995]) with fragment library *Dra*I as template. This fragment was cloned into pGEM-T and sequenced. The sequence corresponded to the 5′ end of the *has* operon including ca. 100 bp upstream of the transcription start site.

The full *S. zooepidemicus has* operon was PCR amplified using primer P64 (5′-GGT TAG CGT TTA AAG GCA C-3′) and primer P99 (5′-GAA ACA CAG CAC AAT CAG AG-3′). P64 was designed to anneal upstream of *hasA*, while P99 was designed to anneal downstream of the *S. equi has* operon (*S. equi* Sequencing Group at the Sanger Center; can be obtained at ftp://ftp.sanger.ac.uk/pub/pathogens/se/). The remotely located paralogues of *hasC* and *glmU* (*N*-acetyl-glucosamine-1-phosphate uridyltransferase) were amplified using primers designed against available sequence data from the *S. equi* genome project. We amplified *hasC* using primer combination P104 (5′-AGC TGG TGC TGC TCT TGG-3′) and P80 (5′-CGG TTT TGC TGC CGT TGC-3′) and *glmU* using P106 (5′-TGG AGC ACA GCA AGG CAG-3′) and P107 (5′-TAC AGC CAA TAT AGC GAC TGC-3′).

Direct Genome Sequencing

Genome sequencing was performed using 3 µg of template DNA and 16 µl of ABI Prism Big Dye Sequencing kit (Applied Biosystems) to gather additional sequence data upstream of the *has* operon. The thermocycler profile used was 99 cycles of 96°C for 30 s, 50°C for 15 s, and 60°C for 4 min. A spin column step ensured the removal of unincorporated nucleotides (Centri spin columns; Princeton Separations). A sequencing reaction with primer P48 (5′-GCA AAC AAG GTC ATT AAT TTA CC-3′) resulted in ca. 500 bp of additional upstream sequence.

Phylogenetic Analyses

HA synthase gene (*hasA*) homologues were identified in the public databases as follows. First, the amino acid sequence of *S. zooepidemicus hasA* and its PFAM motif (Glycos_transf_2; http://www.pfam.sanger.ac.uk/) were used as starting points for profile searches on the NCBI server (PSI-BLAST [Altschul et al. 1997]). The same strategy was used on unfinished genome databases, not yet included in the central database, available from NCBI (http://www.ncbi.nlm.nih.gov/BLAST/), JGI (http://www.jgi.doe.gov/), and TIGR (http://www.tigr.org/). An initial dataset of hasA homologues (ca. 1000 sequences) was

reduced by eliminating partial and redundant entries. Polymer extrusion is a unifying property of the HasA family members, hence all glycosyl transferases lacking transmembrane domains also were excluded. The amino acid sequences were aligned using ClustalW (version 1.8 [Thompson et al. 1994]) or T-COFFEE (Notredame et al. 2000), and the alignment refined manually in the sequence database and analysis software package ARB (Ludwig et al. 2004) using published motifs (Kamst and Spaink 1999). Amino acid datasets for *hasB, hasC, glmU*, and *pgi* (glucose-6-phosphate isomerase) were generated using the same approach.

Masks were created in ARB using the base frequency filter tool (25% minimal similarity) to remove regions of ambiguous positional homology from all amino acid alignments. This produced masked alignments of 185, 305, 251, 442, and 304 amino acids for *hasA, hasB, hasC, glum*, and *pgi*, respectively, available upon request from the authors. Evolutionary distance (ED) dendrograms were constructed from the masked amino acid alignments in ARB using a Dayhoff PAM correction and the Fitch Margoliash additive-tree method (Supplementary Figs. S2–S6). Support for interior nodes of the dendrograms was determined using 200 bootstrap resamplings of maximum parsimony (MP) trees calculated in PAUP* version 4.0b2a (D. L. Swofford, Sinauer Associates, Sunderland, MA) and Bayesian posterior probabilities using the general time-reversible model and gamma correction (Huelsenbeck and Ronquist 2001). In all cases, posterior probabilities were calculated from 1000 trees representing 100,000 generations after a 100,000 generation "burn-in." The ED tree (Fig. 2) serving as a proxy for organismal evolution (vertical descent) was constructed in ARB using a Lane masked small-subunit rRNA gene alignment comprising 1287 characters and the Kimura two-parameter model correction and neighbor joining.

Nucleotide Sequences

The nucleotide sequence of the *has* operon from *S. zooepidemicus* has been deposited in the GenBank database under accession number AF347022, the nucleotide sequence of the remote *hasC* under AF468691, and the nucleotide sequence of the remote *glmU* under AF468690.

Results

Characterization of *has* Operons in *Streptococcus equi* Subspecies

The complete coding region of the *has* operon from *S. equi* subsp. *zooepidemicus* ATCC 35246 (referred to as *S. zooepidemicus*) was determined using a combination of direct genome sequencing and PCR with primers designed

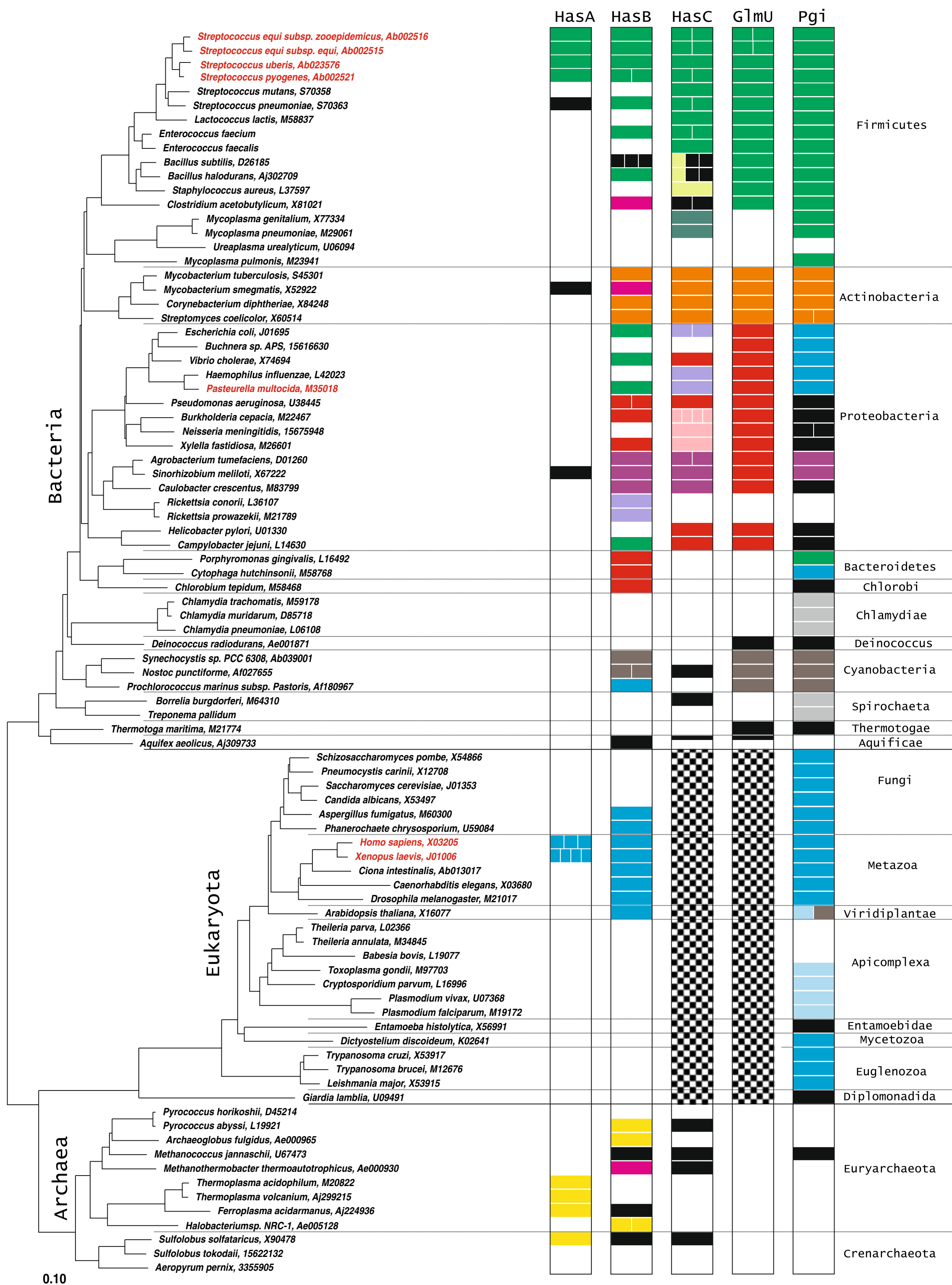
Fig. 2 An evolutionary distance dendrogram (Kimura two-parameter model) of representatives of the three domains of life with available genome sequences, based on comparative analysis of 16S/18S rRNA genes. Major lineages (phyla) within each domain are distinguished by horizontal lines and named at the extreme right. 16S/18S rDNA sequence accession numbers follow the species names, and those in red indicate documented HA producers. The scale bar indicates 0.1 change per nucleotide. The colored columns at the right each represent the phylogenetic affiliation of a *has* operon gene relative to the 16S/18S rRNA phylogeny. Reproducibly monophyletic groups within each *has* operon gene tree (see Materials and Methods for details of tree construction) were assigned a color (Supplementary Figs. S2–S6), and this color was then transferred to the appropriate boxes within a column. Sequences not significantly affiliated with other homologues within their respective trees are shown in black. Checkered areas within the *hasC* and *glmU* columns indicate that alternative enzymes and pathways exist for these functions in eukaryotes. Multiple copies of any given *has* operon gene within a genome are indicated by partitioning of a box within a column

according to database entries as outlined under Materials and Methods. The *has* operon of *S. zooepidemicus*, with a size of 6.8 kb, contains the most number of genes (five) of any *has* operon characterized to date (Fig. 1). A draft genome sequence of *S. zooepidemicus* strain H70 has since become available (www.sanger.ac.uk/Projects/S_zooepidemicus) and the *has* operon therein was 98% identical to our sequence. In order, the operon contains *hasA*, *hasB*, *hasC*, *glmU*, and *pgi*, encoding hyaluronate synthase, UDP-glucose dehydrogenase, glucose-1-phosphate uridyltransferase, a putative *N*-acetyl-glucosamine-1-phosphate uridyltransferase, and a putative glucose-6-phosphate isomerase, respectively (GenBank accession number AF347022). HasB and HasC catalyze the synthesis of the UDP-glucuronic acid precursor from glucose-1-phosphate, and Pgi and GlmU are two of the enzymes involved in the synthesis of the second HA precursor, UDP-*N*-acetyl-glucosamine, from glucose-6-phosphate (Fig. 1). Primer extension analysis identified a transcription start site 50 bp upstream of *S. zooepidemicus hasA* and a σ^{70} promoter Pribnow Schaller box 6 bp upstream from the transcription start (Supplementary Fig. S1). From the apparent presence of a conserved sorbitol operon immediately downstream of *pgi*, we conclude that the operon is complete.

We found a similar gene order in the recently finished *S. equi* subsp. *equi* genome (www.sanger.ac.uk/Projects/S_equi; referred to as *S. equi*; Fig. 1), however, the genes are organized in two operons (*hasA*, *hasB*, *hasC* and *hasC*, *glmU*, *pgi*) as described below.

Role of Lateral Gene Transfer in Streptococcal *has* Operon Formation

The role of LGT between different bacterial species in operon formation was investigated by establishing the phylogeny for homologues of each of the five genes in the *S. zooepidemicus has* operon (see Materials and Methods).



Completed genome sequences were used (with the exception of a few basal eukaryote genomes) to provide a complete picture of the *has* operon genes in these organisms. Phylogenetic trees were inferred from the gene datasets using evolutionary distance, MP, and Baysian inference to minimize potential tree reconstruction artifacts. Branching points resolved by all three methods with MP bootstrap or Baysian posterior probability values ≥75% were considered to be reliable. In a few instances, branchpoints with Baysian posterior probabilities of 100% but MP bootstraps <75% were also used to define monophyletic groups.

Monophyletic groups within each gene tree (Supplementary Figs. S2–S6) were color-coded and overlaid on a small-subunit (16S/18S) ribosomal RNA gene (rDNA) tree (Fig. 2). Sequences not significantly affiliated with other homologues within their respective trees are shown in black. Ribosomal RNAs are highly conserved functional RNAs integral to cellular protein production and are thought to be good indicators of organismal phylogeny (Woese 1987). Incongruencies between rDNA and other gene phylogenies may provide evidence of LGT (Doolittle 1999) and appear in Fig. 2 as a mixing of colors. For example, the *hasB* gene in *Prochlorococcus marinus* (a cyanobacterium) is closely related to eucaryal *hasB* genes shown in blue in Fig. 2, and is not closely related to the *hasB* genes of other cyanobacteria shown in brown in Fig. 2. Therefore, we hypothesize that *Prochlorococcus* received its *hasB* via a lateral transfer from a eucaryal donor. More generally, proteobacterial HasB sequences are intermingled with *Firmicutes* and *Bacteroidetes*, and proteobacterial Pgi sequences with eukaryota. By contrast, the gene coding for GlmU appears to have been vertically transmitted since its topology closely follows that of 16S rRNA as evidenced by the consistency of color groupings (Fig. 2). The color-coding of HasC sequences tends to suggest that the gene has been subject to lateral transfer, but this may rather be an artifact of poor resolution in the HasC tree due to limited comparative information.

Figure 2 confirms the extremely limited representation of HasA gene homologues in the tree of life observed by others (Salzberg et al. 2001). Homologues of HasA with a putative glycosyl transferase function were identified in *Sulfolobus solfataricus* and *Mycobacterium smegmatis* and are the first described *hasA* homologue representatives of the *Crenarchaeota* and *Actinobacteria* phyla, respectively (see Supplementary Fig. S2). No additional *hasA* homologues were identified in nonvertebrate eukaryote genomes, suggesting that the highly restricted distribution of *hasA* homologues in eukaryotes is real and not a sampling artifact as suggested for some metabolic genes (Rogers et al. 2007). Notably, however, the tree provides no direct evidence for lateral transfer of *hasA* between vertebrates and streptococci, as both groups are monophyletic and consistent with the 16S rRNA tree despite very low overall representation in the tree of life (Fig. 2). Unfortunately, the eukaryotic genome sequences are generally less complete than their bacterial and archaeal counterparts, and we therefore cannot entirely rule out the presence of *hasA* genes in the nonvertebrate eukaryotes presented in Fig. 2. The absence of *hasB* in most basal eukaryotes as well, however, supports the view that the role of HA in embryonic development is a relatively late invention in eukaryotes (Spicer and McDonald 1998).

Although Fig. 2 suggests that interspecies, interphylum, and even interdomain LGT has occurred for HasB and Pgi genes, HA-producing streptococci are monophyletic for all of the *has* operon genes. Moreover, in all trees *S. equi* and *S. zooepidemicus* are reproducibly nearest neighbors immediately flanked by *S. pyogenes* (Supplementary Figs. S2–S6), consistent with the 16S rRNA tree and indicative of vertical descent. Phylogenetic inference therefore suggests that interspecies gene transfer has not played a role in *has* operon formation, but intraspecies (strain-level) LGT cannot be discounted.

Origin of *has* Operons in Streptococci

The identification of two new *has* operons in the *S. equi* subspecies brings to four the number of known functional *has* operons (Fig. 1). There is also a phylogenetically related operon (*cap*) of *Streptococcus pneumoniae* type 3 (Arrecubieta et al. 1994), which has a different function—the synthesis of a cellubiuronic acid polymer (Cartee et al. 2000). In addition, there is a putative *has* operon of unverified function on the mega-plasmid pXO1 of *Bacillus anthracis* (Okinaka et al. 1999), a bacterium which lacks an HA capsule. The HasA homologue from *B. anthracis* groups with streptococcal *hasA*, while the Cap3B for *S. pneumoniae* is an orphan homologue (Supplementary Fig. S2). Conceivably, all six HasA homologues could have originated in a common *Firmicutes* ancestor, with Cap3B having undergone rapid evolution and functional reassignment due to selective pressure as a human pathogen.

Neither HasA nor Cap3B is functional in the absence of HasB (Fig. 1B) and we first considered the possibility that the *hasA*/*hasB* pairing originated in a common *Firmicutes* ancestor. HasB's coded by the pXO1 operon, however, map with other *Bacillus* strains, while the five streptococcal HasB's map in a distant group together with a number of Proteobacteria (Supplementary Fig. S3). The HasC gene on pXO1 is also of *Bacillus* origin, suggesting that *hasA* alone was transferred between pXO1 and a streptococcal ancestor (direction unknown). Furthermore, gene rearrangement (BA rather than AB; Fig. 2) and rapid divergence of the *S. pneumoniae* operon genes cast doubts about a common streptococcal ancestor for the *cap* and *has* operons. Thus, a common streptococcal ancestor can only be inferred for the four *has* operons.

The branching order of the streptococci cannot be fully resolved for the four species on the basis of available 16S rRNA sequences (Fig. 1A). As expected, the two *S. equi* subspecies group together but their specific relationship to *S. pyogenes* and *S. uberis* is unclear. Phylogenetic inference from individual *has* operon genes and a concatenated HasA/HasB dataset also failed to resolve the branching

order (data not shown). However, the ribosomal binding site of *hasA* in *S. uberis* and *S. pyogenes*, but not in *S. zooepidemicus*, contains a 22-bp fragment with 13 conserved positions and identical spacing. Also, the intergenic region is small in *S. uberis* and *S. pyogenes* (24 and 35 bp, respectively), while fairly long (251 bp) in the two *S. equi* subspecies. Finally, of the 24 bp of the *S. uberis* intergenic region, 15 bp is conserved in the *S. pyogenes* sequence, while no similarity could be detected in the intergenic region between *S. uberis* and the two *S. equi* subspecies. These noncoding sequence data suggest a closer relationship of *S. uberis* and *S. pyogenes* to each other than to the *S. equi* subspecies and argue against a simple linear addition of *hasC* to the operons.

Assembly of Streptococcal *has* Operons

Genes involved in HA precursor synthesis (*hasC*, *glmU*, and *pgi*) have been added to the basic HasA/HasB operon over the course of evolution (Fig. 1A). These additional *has* operon genes all have functional paralogues (Fitch 1970) remotely located in their host genomes, with the exception of *pgi* (www.sanger.ac.uk/Projects/S_zooepidemicus, www.sanger.ac.uk/Projects/S_equi), and it has been shown in *S. pyogenes* that the operon *hasC* can be knocked out without loss of the HA capsule (Ashbaugh et al. 1998). This raises the possibility that intragenomic gene duplication (paralogue formation) is the principal mechanism involved in *has* operon assembly.

We identified two nearly identical copies of *hasC* in the *S. zooepidemicus* genome, one in the *has* operon (*hasC*) and the second in a two gene operon containing *gpsA*, which codes for an NAD(P)$^+$ dependent glycerol-3-phosphate dehydrogenase (EC 1.1.1.94) and *hasC* (*galU*) (Fig. 3). Comparison of the sequences flanking the *hasC* genes indicate that *galU* and 24 bp of *gpsA* were duplicated and incorporated into the *has* operon (Fig. 3). Traces of a possible mechanism involved in this gene duplication are an 8-bp palindrome on the 5′ insertion border with *hasB* and a 6-bp repeat at the insertion borders (Fig. 3).

The major part of the *S. zooepidemicus hasC* paralogues is identical (862 of 903 bp; hatched region in Fig. 3), suggesting that the gene duplication was a recent event. However, inspection of the *S. equi* genome revealed a nearly identical *galU* duplication including the 3′ end of *gpsA* (Fig. 3), with the exception that the *S. equi galU* is part of the *has* operon and the *hasC* homologous to the *hasC* of *S. zooepidemicus* is part of the second operon including *glmU* and *pgi*. As it would be highly improbable that identical duplication events occurred independently in *S. zooepidemicus* and *S. equi*, *hasC* duplication must have occurred prior to subspeciation. Therefore, the near-identity of *galU* and *hasC* in *S. zooepidemicus* likely is due to frequent homologous recombination of these paralogues between base 1 and base 862 (Fig. 3). This also appears to be the case for the *S. equi hasC* paralogues, which are identical in the 1- to 862-bp region but distinct from the *S. zooepidemicus* paralogues (97% identical). The 3′ end of the *hasC* paralogues does not appear to be involved in homologous recombination since *hasC*'s in the two subspecies are more closely related to each other than to their respective *galU* paralogues (Fig. 3). Why homologous recombination of *hasC* paralogues is selected for in these *S. equi* subspecies is unknown, but it is not ubiquitous in HA-producing streptococci. A gene duplication of *hasC* occurred in the *S. pyogenes* genome, however, the paralogues are only 79% identical on the DNA level, suggesting no homologous recombination of the paralogues has occurred since their duplication. As with the *S. equi* subspecies duplication, the 3′ end of *S. pyogenes gpsA* was transferred with *hasC* to the *has* operon. However, it must have been an independent event from the duplication in the *S. equi* subspecies because a greater

Fig. 3 Schematic alignment of the two *hasC* (*hasC*, *galU*) paralogues in *S. zooepidemicus* and *S. equi*. The genome location of the paralogues is indicated at the right. Gene names and start and stop codons for the *hasCs* are shown at the top. Putative homologue recombination sites are magnified at the bottom. Palindromes within these sites are in boldface, repeated sequences are underlined, and stop codons are italicized. Shading indicates sequence similarities (see text for details)

portion of *gpsA* was transferred (45 vs. 24 bp; data not shown).

The duplication of *glmU* into the *has* operon appears to follow a similar scenario as described for *hasC* in both *S. zooepidemicus* and *S. equi*. The insertion borders are identical for the two strains, again suggesting that the gene duplication occurred prior to subspeciation. However, the direction of *glmU* transfer could not be determined from the sequence data because no flanking genes were involved, and indeed the insertion borders exactly coincide with the open reading frame of *glmU*. Indirect evidence that the transfer of duplication was into the *has* operon comes from the single copy of *glmU* found in *S. pyogenes*, *S. mutans*, and *S. pneumoniae* (Supplementary Fig. S5). The *glmU*'s of these organisms have the same local gene environment as the non-*has* operon *glmU* of the *S. equi* subspecies (Supplementary Fig. S7).

In contrast to *hasC* and *glmU*, *pgi* was not found to be duplicated in the *S. equi* subspecies. As the Pgi is, however, essential for growth of streptococci on sugars as carbon source, but not HA production, *has* operon independent transcription is required. Indeed, we identified a highly conserved σ^{70} promoter upstream of *pgi* using BPROM (http://www.softberry.com/all.htm), suggesting that *pgi* expression is under the control of at least two promoters, the *has* operon and the *pgi* promoter.

Finally, the four *has* operons have moved within the respective streptococci genomes as they currently reside in conserved stretches of genes (Supplementary Figs. S8 and S9). Specifically, the *has* operons of the *S. equi* subspecies reside downstream of a conserved hypothetical RNA encoding gene. Although the *has* operons of these two species vary in gene content, downstream of *galU* and *pgi* the same genes are located. These gene arrangements suggest that the *has* operons in the two *S. equi* subspecies can change gene content (three or five genes) via homologue recombination of the highly conserved *hasC*/*galU* genes. Thus the two different *has* operon structures in *S. equi* and *S. zooepidemicus* (Fig. 1) are two alternatives of the same gene arrangement. The *pgi* of the *S. zooepidemicus has* operon was acquired by insertion of the four-gene *has* operon (*hasA*, *hasB*, *hasC*, *glmU*) upstream of the *pgi* gene (Supplementary Fig. S8). In contrast, the operon of *S. uberis* is located downstream of the *pgi* and sorbitol operon and upstream of a putative hippurate hydrolase gene and the *galU* (Supplementary Fig. S9). The *S. pyogenes* operon resides in a highly conserved stretch of genes that are present in all streptococci (Supplementary Fig. S9). In summary, despite evolving without recent LGT, the streptococcal *has* operons appear to be highly dynamic within their host genomes, having been constructed via a number of independent gene duplications and entire operon relocations.

Discussion

We have characterized two new streptococcal *has* operons. A five-gene operon was cloned from *S. zooepidemicus*, while a three-gene operon was identified in the draft genome sequence of the closely related *S. equi*. The *S. equi* subspecies *S. zooepidemicus* has significantly greater HA production than *S. uberis* and *S. pyogenes*, which have a *has* operons containing only HA synthase and one or two genes involved in HA precursor synthesis respectively. The inclusion of additional genes involved in precursor formation in the *S. equi* subspecies is most likely driven by a selective advantage of enhanced HA production and hence coat thickness. Indeed, *S. zooepidemicus* can direct more than 10% of glucose carbon toward HA (Blank et al. 2005), which is 10 to 100 times higher than the glucose usage for extracellular or capsular polysaccharides in other streptococci such as *S. thermophilus* (De Vuyst et al. 2001).

HA synthase was previously identified as one of 223 putative candidates for LGT from bacteria to human (Lander et al. 2001) based on the absence of nonvertebrate *hasA* homologues. While our results confirm this absence in nonvertebrate genomes, there is no unequivocal phylogenetic evidence pointing toward LGT despite the fact that HasA genes are clearly capable of being mobilized by their presence on a *Bacillus anthracis* plasmid and in a viral genome (Supplementary Fig. S2). *hasA* homologues have a very limited representation in all domains, with only five monophyletic clades: vertebrate HasA, streptococcal HasA, rhizobial NodC, an archaeal glycosyl transferase cluster, and a putative glycosyl transferase group represented by *Mycobacterium smegmatis* in Supplementary Fig. S2. No group branches within the radiation of any other group and the vertebrate HA synthases are no more closely related to streptococcal HA synthases than to rhizobial NodC (producing oligochitins) or putative archaeal glycosyl transferases (Supplementary Fig. S2). Thus, we concur with Salzberg et al. (2001) that other possible explanations for this unusual gene distribution, such as massive gene loss, are equally (un)likely.

Similarly, there is no evidence that any of the other genes in the streptococcal *has* operons were obtained by interspecies LGT. In this regard, the *has* operons differ from most polysaccharide operons found in bacteria, including other streptococci, which appear to be chimeric assemblies of indirect glycosyl transferases obtained from at least two different bacterial genera (Bourgoin et al. 1999). We attribute this difference to different selective pressures. Where HA is a nonimmunogenic coat, most bacterial polysaccharides are immunogenic and diversification of polysaccharide composition is a means of bypassing the immune system. Therefore, LGT is likely to be favored under these conditions. For example, 90

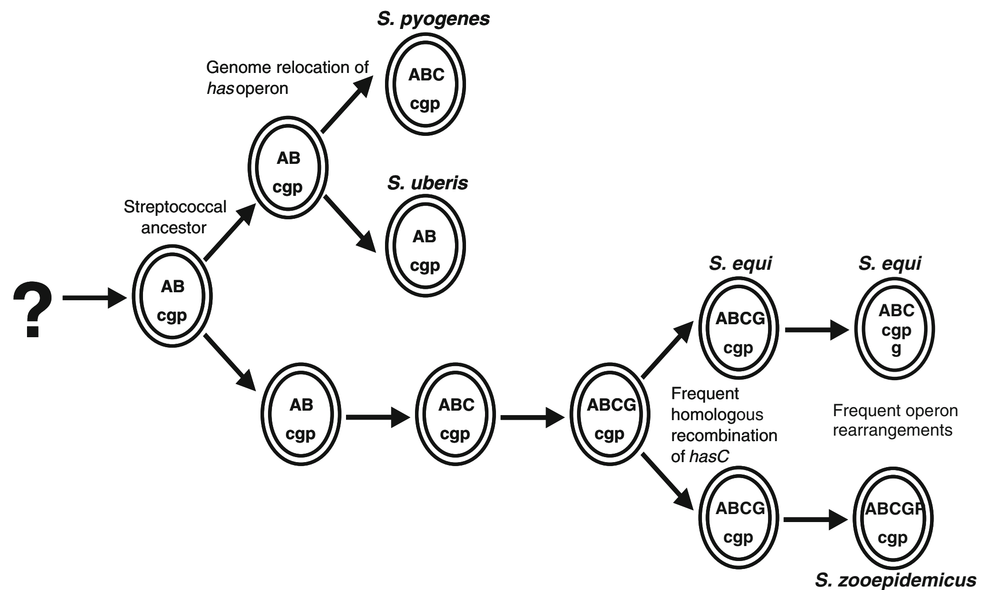


Fig. 4 Proposed scenario for the evolution of *has* operons in HA-encapsulated streptococci. Letters represent genes as follows: A, *hasA*; B, *hasB*; C, *hasC*; G, *glmU*; P, *pgi*. Capital letters indicate genes located in the *has* operon, and lowercase letters indicate remotely located genes

recognized serotypes of *S. pneumoniae* are distinguished principally by their capsule composition, which is a result of frequent lateral exchange of genes between species (Bentley et al. 2006; Chaffin et al. 2002; Garcia et al. 2000; Mavroidi et al. 2004). Incorporation of a single, new glycosyl transferase activity is sufficient to change the polysaccharide repeating unit and hence the capsule composition (Bentley et al. 2006).

By contrast, we propose that the increase from the two essential genes in the *has* operon (*hasA* and *hasB*) to three and five genes occurred by intragenomic gene duplication, although strain-level LGTs cannot be entirely ruled out. Sequence similarity and spacing in the ribosomal binding site and intergenic sequence suggest that *S. uberis* and *S. pyogenes* had a common ancestor, which diverged from the common *S. equi* subspecies ancestor (Fig. 4). This is consistent with the finding that *hasC* duplication occurred separately in *S. pyogenes* and the two *S. equi* subspecies, as indicated by the insertion border with *gpsA*. In the two *S. equi* subspecies, *hasC* and *glmU* duplication appear to have occurred prior to subspeciation (Fig. 4) judging by the sequence identity in the flanking regions for *hasC* and the identical insertion length and position for both *hasC* and *glmU*. In addition to duplication, there is evidence of frequent homologous recombination for the two *hasC* paralogues in both *S. equi* subspecies (Fig. 3), partial operon rearrangement in *S. equi*, as well as full operon relocation in *S. equi*, *S. pyogenes*, *S. uberis*, and *S. zooepidemicus*.

A detailed understanding of the evolution of operons that produce industrially important polymers, such as the *has* operon, may assist in systematic efforts to engineer strains for improved production rates and yields.

Acknowledgment This work was supported by Grant A89937110 from the Australian Research Council.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Anderson DG, McKay LL (1984) Genetic and physical characterization of recombinant plasmids associated with cell aggregation and high-frequency conjugal transfer in *Streptococcus lactis* ML3. J Bacteriol 158:954–962

Arrecubieta C, Lopez R, Garcia E (1994) Molecular characterization of cap3A, a gene from the operon required for the synthesis of the capsule of *Streptococcus pneumoniae* type 3: sequencing of mutations responsible for the unencapsulated phenotype and localization of the capsular cluster on the pneumococcal chromosome. J Bacteriol 176:6375–6383

Ashbaugh CD, Alberti S, Wessels MR (1998) Molecular analysis of the capsule gene region of group A streptococcus: the *hasAB* genes are sufficient for capsule expression. J Bacteriol 180:4955–4959

Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M, Donohoe K, Harris D, Murphy L, Quail MA, Samuel G, Skovsted IC, Kaltoft MS, Barrell B, Reeves PR, Parkhill J, Spratt BG (2006) Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet 2:e31

Blank LM, McLaughlin RL, Nielsen LK (2005) Stable production of hyaluronic acid in *Streptococcus zooepidemicus* chemostats operated at high dilution rate. Biotechnol Bioeng 90:685–693

Bourgoin F, Pluvinet A, Gintz B, Decaris B, Guedon G (1999) Are horizontal transfers involved in the evolution of the *Streptococcus thermophilus* exopolysaccharide synthesis loci? Gene 233:151–161

Cartee RT, Forsee WT, Schutzbach JS, Yother J (2000) Mechanism of type 3 capsular polysaccharide synthesis in *Streptococcus pneumoniae*. J Biol Chem 275:3907–3914

Chaffin DO, McKinnon K, Rubens CE (2002) CpsK of *Streptococcus agalactiae* exhibits alpha2,3-sialyltransferase activity in *Haemophilus ducreyi*. Mol Microbiol 45:109–122

Chong BF, Blank LM, McLaughlin R, Nielsen LK (2005) Microbial hyaluronic acid production. Appl Microbiol Biotechnol 66: 341–351

DeAngelis PL, Papaconstantinou J, Weigel PH (1993) Molecular cloning, identification, and sequence of the hyaluronan synthase gene from group A *Streptococcus pyogenes*. J Biol Chem 268:19181–19184

De Vuyst L, De Vin F, Vaningelgem F, Degeest B (2001) Recent developments in the biosynthesis and applications of heteropolysaccharides from lactic acid bacteria. Int Dairy J 11:687–707

Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284:2124–2129

Dougherty BA, van de Rijn I (1994) Molecular characterization of *hasA* from an operon required for hyaluronic acid synthesis in group A streptococci. J Biol Chem 269:169–175

Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–113

Garcia E, Llull D, Munoz R, Mollerach M, Lopez R (2000) Current trends in capsular polysaccharide biosynthesis of *Streptococcus pneumoniae*. Res Microbiol 151:429–435

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

Kamst E, Spaink PH (1999) Functional domains in the chitin oligosaccharide synthase NodC and related beta-polysaccharide synthases. Trends Glycosci Glycotechnol 11:187–199

Kumari K, Weigel PH (1997) Molecular cloning, expression, and characterization of the authentic hyaluronan synthase from group C *Streptococcus equisimilis*. J Biol Chem 272:32539–32546

Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lee JY, Spicer AP (2000) Hyaluronan: a multifunctional, megaDalton, stealth molecule. Curr Opin Cell Biol 12:581–586

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G, Forster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, Konig A, Liss T, Lussmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A, Schleifer KH (2004) ARB: a software environment for sequence data. Nucleic Acids Res 32:1363–1371

Mavroidi A, Godoy D, Aanensen DM, Robinson DA, Hollingshead SK, Spratt BG (2004) Evolutionary genetics of the capsular locus of serogroup 6 pneumococci. J Bacteriol 186:8181–8192

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 302:205–217

Okinaka RT, Cloud K, Hampton O, Hoffmaster AR, Hill KK, Keim P, Koehler TM, Lamke G, Kumano S, Mahillon J, Manter D, Martinez Y, Ricke D, Svensson R, Jackson PJ (1999) Sequence and organization of pXO1, the large *Bacillus anthracis* plasmid harboring the anthrax toxin genes. J Bacteriol 181:6509–6515

Rogers MB, Watkins RF, Harper JT, Durnford DG, Gray MW, Keeling PJ (2007) A complex and punctate distribution of three eukaryotic genes derived by lateral gene transfer. BMC Evol Biol 7:89

Salzberg SL, White O, Peterson J, Eisen JA (2001) Microbial genes in the human genome: lateral transfer or gene loss? Science 292:1903–1906

Siebert PD, Chenchik A, Kellogg DE, Lukyanov KA, Lukyanov SA (1995) An improved PCR method for walking in uncloned genomic DNA. Nucleic Acids Res 23:1087–1088

Spicer AP, McDonald JA (1998) Characterization and molecular evolution of a vertebrate hyaluronan synthase gene family. J Biol Chem 273:1923–1932

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. Nucleic Acids Res 22: 4673–4680

Tlapak-Simmons VL, Baggenstoss BA, Clyne T, Weigel PH (1999) Purification and lipid dependence of the recombinant hyaluronan synthases from *Streptococcus pyogenes* and *Streptococcus equisimilis*. J Biol Chem 274:4239–4245

Ward PN, Field TR, Ditcham WG, Maguin E, Leigh JA (2001) Identification and disruption of two discrete loci encoding hyaluronic acid capsule biosynthesis genes *hasA*, *hasB*, and *hasC* in *Streptococcus uberis*. Infect Immun 69:392–399

Wibawan IW, Pasaribu FH, Utama IH, Abdulmawjood A, Lammler C (1999) The role of hyaluronic acid capsular material of *Streptococcus equi* subsp. *zooepidemicus* in mediating adherence to HeLa cells and in resisting phagocytosis. Res Vet Sci 67:131–135

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271