The *sigH* gene sequence can subspeciate staphylococci.

Kazuya Morikawa[a]*, Ryosuke L. Ohniwa[a], Miyuki Kumano[a], Hideyuki Okamura[b], Shinji Saito[a], Toshiko Ohta[a]

[a]Institute of Basic Medical Sciences, Graduate School of Comprehensive Human Sciences, University of Tsukuba, Tsukuba, 305-8575 Japan

[b]Department of Biology, Osaka Dental University, Hirakata, 573-1121 Japan

* corresponding author. Institute of Basic Medical Sciences, Graduate School of Comprehensive Human Sciences, University of Tsukuba, Tsukuba, 305-8575 Japan. Tel & fax:+81-29-853-3928.

E-mail address: morikawa@sakura.cc.tsukuba.ac.jp (K. Morikawa)

running title: *Staphylococcus* species identification by *sigH*

Abstract

In an evolutionarily conserved gene organization (syntenic region), the *sigH* gene shares exceptionally low homology among staphylococcal species. We analyzed the 'positionally cloned' *sigH* sequences of 39 staphylococcal species. The topology of the SigH phylogenetic tree was consistent with that of 16S rRNA. Certain clinical isolates were successfully differentiated at the species level with the *sigH* sequence data set. We propose that the *sigH* gene is a promising molecular target in genotypic identification, since it is highly discriminative in differentiating closely related staphylococcal species.

1. Introduction

The *Staphylococcus* genus is currently classified into 39 species (Euzeby, 1997). The precise species identification of staphylococci is important in etiological and epidemiological studies. Phenotypic tests are broadly used in today's clinical laboratories, but precise discrimination of the species often requires additional information (Grant et al., 1994; Perl et al., 1994; Renneberg et al., 1995). Several gene sequences such as 16S rRNA (Takahashi et al., 1999; Becker et al., 2004; Fujita et al., 2005; Skow et al., 2005), *tuf* (Martineau et al., 2001), *rpoB* (Drancourt and Raoult, 2002; Mellmann et al., 2006), *hsp60* (Goh et al., 1997; Kwok and Chow, 2003), *gap* (Yugueros et al., 2000), *femA* (Vannuffel et al., 1999), and *sodA* (Poyart et al., 2001; Sivadon et al., 2004; Giammarinaro et al., 2005) have been used in the identification of staphylococcal species. However, highly conserved genes are not suitable for the discrimination of closely related species.

The *sigH* gene encodes a sigma factor that constitutes a large alternative sigma factor group in *Firmicutes*, and is located within an evolutionarily conserved gene organization (Morikawa et al., 2003). In sporulating bacteria such as *Bacillus subtilis*

and *Clostridium perfringens*, *SigH* regulates the initiation of sporulation and is highly

conserved even among the distinct genera. In contrast, *SigH* in non-sporulating bacteria

such as *Staphylococcus* and *Streptococcus* shares low homology among species.

In this study, we analyzed the *sigH* sequences of 39 staphylococcal species, aiming

to show that the *sigH* gene is a promising molecular target for discriminating between

closely related species. In addition, bioinformatic analysis revealed that there are similar

gene clusters with high local diversity, suggesting that such syntenic regions may be

promising molecular targets for species identification.

2. Materials and method

2.1. Bacterial strains

Staphylococcus strains used in this study are shown in Table 1. Clinical isolates of *S. saprophyticus* candidates were obtained as described previously (Higashide et al., 2006). Chromosomal DNA was extracted using conventional methods.

2.2. PCR amplification of the sigH locus and sequencing analysis

PCR primers (Table 1) were designed for the conserved region outside of the *sigH* coding sequence (Fig. 1). The *sigH* locus was amplified by PCR using Ex-taq polymerase (Takara). The PCR products were purified by UltraClean PCR Clean-up DNA Purification Kit (MO BIO) or by Wizard SV Gel and PCR Clean-Up System (Promega) and directly sequenced. Alternatively, the PCR products were cloned into pGEM-5Zf(+) using pGEM T-Vector System I (Promega), and multiple independent clones were subjected to sequencing analysis.

Sequencing was performed with DYEnamic ET Terminator Cycle Sequence Kit (Amersham Biosciences) or BigDye Terminator Cycle Sequencing Ready Reaction Kit

(Applied Biosystems) and analyzed using the ABI 310 genetic Analyzer (Applied

Biosystems). The *sigH* gene sequences were deposited in the GenBank/EMBL database.

The 16S rDNA was amplified with primers 16S-F and 16S-R (Table 1) and directly

sequenced.

2.3. Phylogenetic analysis

The phylogenetic tree of *SigH* was constructed using the DNADIST and

NEIGHBOR programs of the PHYLIP 3.6 package (Felsenstein, 1989). Bootstrap

analysis was performed with the SEQBOOT and CONSENSE programs of the same

package (100 iterations; bootstrap values of higher than 50% are shown). The resultant

phylogenetic tree was visualized using the TREEVIEW program (Page, 1996).

3. Results

3.1. The sigH gene is located between conserved genes.

The *sigH* gene has low homology among species but is embedded in a conserved gene cluster (Fig. 1; (Morikawa et al., 2003)). The genes adjacent to *sigH* in *S. epidermidis*, *S. haemolyticus* and *S. saprophyticus* share over 70% identity with those in *S. aureus*, whereas the *sigH* gene shares less than 40 % identity. We supposed that such a local diversity would be useful for identifying species; the high divergence of the target gene is necessary to discriminate between closely related species, whilst the conserved regions encompassing it would allow us to amplify the target gene.

3.2. All staphylococcal species carry the sigH gene at the expected position.

Degenerate primers were designed for the positions that were expected to encompass the *sigH* gene (Fig 1). In all of the 39 species, the PCR fragments with the expected length (0.8~1.8 kbp) were obtained using a set of primers. Sequencing analysis showed that all of the 39 species had the *sigH* gene at the expected position. Mutations generating a stop-codon or frameshift were found only in *S. auricularis*.

3.3. SigH exhibits high divergence.

   SigH exhibited high sequence diversity except in two moderately conserved regions.

These two regions are known to be well conserved among σ^{70}-family sigma factors and

correspond to 'region 2.4' and 'region 4.2', which are responsible for the recognition of

the -10 and –35 promoter sequences, respectively (Wösten, 1998). The lowest identity

among species was between *S. lentus* and *S. arlettae* at 15.5 % (Table 2). Regarding *S.*

pulvereri and *S. vitulinus* as the same species (see below), the highest value was 95.3%

between *S. delphini* and *S. pseudintermedius,* followed by 91.1% between *S. delphini*

and *S. intermedius* and 88.9% between *S. intermedius* and *S. pseudintermedius*. These

values are significantly lower than those of other genes that have been used for species

identification. According to the public database, *SigH* divergence within a single species

seems to be limited; the lowest identity among 9 *S. aureus* isolates was 98.6% (data not

shown).

3.4. Phylogeny of SigH exhibits concordant evolution with the species.

   Our previous phylogenetic analysis for 12 staphylococcal species and other

Firmicutes bacteria suggested that SigH seemed to have diversified along with the evolution of the species without horizontal transfer (Morikawa et al., 2003). The present study substantiates this finding (Fig. 2): the topology of the SigH tree exhibited concordance with the results based on 16S rRNA and DNA-DNA hybridization.

In addition, the *SigH* phylogenetic analysis suggests or supports the following points concerning the phylogeny of closely related species: 1) *S. nepalensis* is closer to *S. cohnii* than to *S. saprophyticus* The relationship of these species was obscure according to the analysis of 16S rRNA (98.8% ~ 99% identity) (Spergser et al., 2003). 2) *S. pseudintermedius* is closer to *S. delphini* than to *S. intermedius,* supporting the result of DNA-DNA hybridization (Devriese et al., 2005). 3) *S. pulvereri* and *S. vitulinus* constitute a single species, because the *SigH* identity was 99.5%. This was previously proposed by others (Petras, 1998; Svec et al., 2004).

3.5. Species identification by sigH sequencing.

In a previous epidemiological study in our laboratory, *S. saprophyticus* was collected from urine and vagina (Higashide et al., 2006). The 108 candidates for *S.*

saprophyticus screened by conventional methods (e.g., Gram staining, production of enzymes, novobiosin resistance) were subjected to final phenotypic testing using API staph ID 32 (BioMeruiux). It failed to identify 6 isolates as *S. saprophyticus*. Using these 6 isolates, we sequenced their *sigH* and 16S rRNA genes, expecting that the *sigH* sequencing would help the species identification. Three of the isolates were judged to be *S. saprophyticus* because their SigH and 16S rRNA sequences were 100% identical to those of *S. saprophyticus*. On the other hand, the SigH sequences of the other 3 strains, which were almost identical to each other, were most closely related to *S. cohnii* with an identity value of 79%. Therefore, we judged them to be a species close to *S. cohnii* or a subspecies of *S. cohnii*. Their 16S rDNA sequence was identical to that of *S. cohnii*.

3.6. Other gene clusters with high local diversity.

We explored additional gene cluster(s) carrying embedded diverse genes (termed a 'similarity gulf') by comparing four genome sequences (*S. aureus, S. epidermidis, S. haemolyticus, and S. saprophyticus*) (Fig. 3). SAS026 is a small hypothetical protein embedded between *murE* and *prfC,* which encode

10

UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase and peptide chain release factor 3, respectively. SA1712 encodes a hypothetical transmembrane protein and is located between SA1713 and SA1711. As expected, among the four species, the phylogenetic relationships of SAS026 and SA1711 exhibited no conflict with those of others including *SigH*.

4. Discussion

We established the *sigH* sequence data set of all 39 staphylococcal species. Its high divergence is useful for discriminating between closely related species that cannot be assigned by conventional phenotypic testing or by using more highly conserved genes. Among the genes tested so far, the one most successfully used for discriminating between closely related species is *sodA* (Poyart et al., 2001; Sivadon et al., 2004; Sivadon et al., 2005). It was used to design hybridization probes and enabled the identification of all staphylococcal species tested, but unfortunately, even the divergence of *sodA* was not sufficient to discriminate between certain subspecies (Giammarinaro et al., 2005). Since the divergence of the *sigH* gene is higher than that of any molecular

target tested so far, it may serve as a better gene to identify such a subspecies. The extraordinary divergent regions within the *sigH* gene (outside of region 2 and region 4) may be promising targets for designing hybridization probes.

In the analysis of unknown clinical isolates, the SigH sequence did not match any of those in our data set, although their 16S rDNA sequence was identical to that of *S. cohnii*. Here, it should be noted that *S. cohnii* is further divided into two subspecies: *S. cohnii* subsp. *cohnii* and *S. cohnii* subsp. *urealyticus*, and the strain in our data set (GTC248) is *S. cohnii cohnii*. The 16S rDNA sequence cannot discriminate, since it shows little difference with those of *S. cohnii cohnii* (0/464bp), *S. cohnii urealyticus* (1/464bp), and *S. nepalensis* (4/464b). Further SigH sequence information of other subspecies should be included in the SigH data set, if the SigH divergence is used to differentiate*Staphylococcus* at the subspecies level.

The *sigH* gene diverged along with the evolution of the species (Fig. 2). This would be consistent with its being embedded in the syntenic region that has no trace of horizontal transfer. We could detect only two other 'similarity gulfs' under the employed search conditions, but more might be found with less limiting conditions, e.g.,

conditions that allow gaps. These 'similarity gulfs' would be useful for precise species identification.

ACKNOWLEDGMENT

REFERENCES

Becker K, Harmsen D, Mellmann A, Meier C, Schumann P, Peters G, von Eiff C (2004) Development and evaluation of a quality-controlled ribosomal sequence database for 16S ribosomal DNA-based identification of *Staphylococcus* species. *J Clin Microbiol* 42:4988-4995.

Devriese LA, Vancanneyt M, Baele M, Vaneechoutte M, De Graef E, Snauwaert C, Cleenwerck I, Dawyndt P, Swings J, Decostere A, Haesebrouck F (2005)

Staphylococcus pseudintermedius sp. nov., a coagulase-positive species from animals. *Int J Syst Evol Microbiol* 55:1569-1573.

Drancourt M, Raoult D (2002) *rpoB* gene sequence-based identification of *Staphylococcus* species. *J Clin Microbiol* 40:1333-1338.

Euzeby JP (1997) List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet (Last full update July 10, 2006. URL: http://www.bacterio.net). *Int J Syst Bacteriol* 47:590-592.

Felsenstein J (1989) PHYLIP-Phylogeny inference package version 3.2. *Cladistics* 5:164-166.

Fujita S, Senda Y, Iwagami T, Hashimoto T (2005) Rapid identification of staphylococcal strains from positive-testing blood culture bottles by internal transcribed spacer PCR followed by microchip gel electrophoresis. *J Clin Microbiol* 43:1149-1157.

Giammarinaro P, Leroy S, Chacornac JP, Delmas J, Talon R (2005) Development of a new oligonucleotide array to identify staphylococcal strains at species level. *J Clin Microbiol* 43:3673-3680.

Gill SR, Fouts DE, Archer GL, Mongodin EF, Deboy RT, Ravel J, Paulsen IT, Kolonay JF, Brinkac L, Beanan M, Dodson RJ, Daugherty SC, Madupu R, Angiuoli SV, Durkin AS, Haft DH, Vamathevan J, Khouri H, Utterback T, Lee C, Dimitrov G, Jiang L, Qin H, Weidman J, Tran K, Kang K, Hance IR, Nelson KE, Fraser CM (2005) Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. *J Bacteriol* 187:2426-2438.

Goh SH, Santucci Z, Kloos WE, Faltyn M, George CG, Driedger D, Hemmingsen SM (1997) Identification of *Staphylococcus* species and subspecies by the chaperonin 60 gene identification method and reverse checkerboard hybridization. *J Clin Microbiol* 35:3116-3121.

Grant CE, Sewell DL, Pfaller M, Bumgardner RV, Williams JA (1994) Evaluation of two commercial systems for identification of coagulase-negative staphylococci to species level. *Diagn Microbiol Infect Dis* 18:1-5.

Higashide M, Kuroda M, Ohkawa S, Ohta T (2006) Evaluation of a cefoxitin disk

diffusion test for the detection of *mecA*-positive methicillin-resistant *Staphylococcus saprophyticus*. *Int J Antimicrob Agents* 27:500-504.

Kwok AY, Chow AW (2003) Phylogenetic study of *Staphylococcus* and *Macrococcus* species based on partial *hsp60* gene sequences. *Int J Syst Evol Microbiol* 53:87-92.

Martineau F, Picard FJ, Ke D, Paradis S, Roy PH, Ouellette M, Bergeron MG (2001) Development of a PCR assay for identification of staphylococci at genus and species levels. *J Clin Microbiol* 39:2541-2547.

Mellmann A, Becker K, von Eiff C, Keckevoet U, Schumann P, Harmsen D (2006) Sequencing and staphylococci identification. *Emerg Infect Dis* 12:333-336.

Morikawa K, Inose Y, Okamura H, Maruyama A, Hayashi H, Takeyasu K, Ohta T (2003) A new staphylococcal sigma factor in the conserved gene cassette: functional significance and implication for the evolutionary processes. *Genes Cells* 8:699-712.

Page RD (1996) TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* 12:357-358.

Perl TM, Rhomberg PR, Bale MJ, Fuchs PC, Jones RN, Koontz FP, Pfaller MA (1994) Comparison of identification systems for *Staphylococcus epidermidis* and other

coagulase-negative *Staphylococcus* species. *Diagn Microbiol Infect Dis* 18:151-155.

Petras P (1998) *Staphylococcus pulvereri = Staphylococcus vitulus*? *Int J Syst Bacteriol* 48 Pt 2:617-618.

Poyart C, Quesne G, Boumaila C, Trieu-Cuot P (2001) Rapid and accurate species-level identification of coagulase-negative staphylococci by using the sodA gene as a target. *J Clin Microbiol* 39:4296-4301.

Renneberg J, Rieneck K, Gutschik E (1995) Evaluation of Staph ID 32 system and Staph-Zym system for identification of coagulase-negative staphylococci. *J Clin Microbiol* 33:1150-1153.

Sivadon V, Rottman M, Chaverot S, Quincampoix JC, Avettand V, de Mazancourt P, Bernard L, Trieu-Cuot P, Feron JM, Lortat-Jacob A, Piriou P, Judet T, Gaillard JL (2005) Use of genotypic identification by *sodA* sequencing in a prospective study to examine the distribution of coagulase-negative *Staphylococcus* species among strains recovered during septic orthopedic surgery and evaluate their significance. *J Clin Microbiol* 43:2952-2954.

Sivadon V, Rottman M, Quincampoix JC, Avettand V, Chaverot S, de Mazancourt P,

Trieu-Cuot P, Gaillard JL (2004) Use of *sodA* sequencing for the identification of clinical isolates of coagulase-negative staphylococci. *Clin Microbiol Infect* 10:939-942.

Skow A, Mangold KA, Tajuddin M, Huntington A, Fritz B, Thomson RB, Jr., Kaul KL (2005) Species-level identification of staphylococcal isolates by real-time PCR and melt curve analysis. *J Clin Microbiol* 43:2876-2880.

Spergser J, Wieser M, Taubel M, Rossello-Mora RA, Rosengarten R, Busse HJ (2003) *Staphylococcus nepalensis* sp. nov., isolated from goats of the Himalayan region. *Int J Syst Evol Microbiol* 53:2007-2011.

Svec P, Vancanneyt M, Sedlacek I, Engelbeen K, Stetina V, Swings J, Petras P (2004) Reclassification of *Staphylococcus pulvereri* Zakrzewska-Czerwinska et al. 1995 as a later synonym of *Staphylococcus vitulinus* Webster et al. 1994. *Int J Syst Evol Microbiol* 54:2213-2215.

Takahashi T, Satoh I, Kikuchi N (1999) Phylogenetic relationships of 38 taxa of the genus *Staphylococcus* based on 16S rRNA gene sequence analysis. *Int J Syst Bacteriol* 49 Pt 2:725-728.

Uchiyama I (2007) MBGD: a platform for microbial comparative genomics based on

the automated construction of orthologous groups. *Nucleic Acids Res* 35:D343-346.

Vannuffel P, Heusterspreute M, Bouyer M, Vandercam B, Philippe M, Gala JL (1999)

Molecular characterization of *femA* from *Staphylococcus hominis* and

Staphylococcus saprophyticus, and femA-based discrimination of staphylococcal

species. *Res Microbiol* 150:129-141.

Wösten MMSM (1998) Eubacterial sigma-factors. *FEMS Microbiol Rev* 22:127-150.

Yugueros J, Temprano A, Berzal B, Sanchez M, Hernanz C, Luengo JM, Naharro G

(2000) Glyceraldehyde-3-phosphate dehydrogenase-encoding gene as a useful

taxonomic tool for *Staphylococcus* spp. *J Clin Microbiol* 38:4351-4355.

Figure captions

Fig. 1

Gene organization around the *sigH* gene. Number below the gene name is the percent identity of each amino acid (DNA) sequence compared with *S. aureus*. Arrows shows the locations of the degenerate primers.

Fig. 2

Unrooted phylogenetic tree based on SigH. The numbers above and below the branches represents the bootstrap values of more than 50 (100 replications). Cluster groups shown on the right are based on 16S rRNA (Takahashi et al., 1999), which shows well concordance to the SigH phylogenetic tree. The scale bar shows the Dayhoff distance among the SigH molecules. In this analysis, the frame shift mutation in *S. auricularis* was eliminated.

Fig. 3

'Similarity gulf' in syntenic region in genus *Staphylococcus*. We collected the gene

clusters carrying diversified genes (< 40% identity) sandwiched by conserved genes (>

70% identity) by Perl programming with the dataset provided by KEGG (Kanehisa et al.,

2002). These gene organizations seem to be restricted in genus *Staphylococcus*

(searched by MBGD, Uchiyama, 2007), in contrast to the *sigH* cluster that is broadly

conserved in *Firmicutes*.