Characterization of novel LPXTG-containing proteins of *Staphylococcus aureus* identified from genome sequences

Fiona M. Roche,1 Ruth Massey,2 Sharon J. Peacock,2 Nicholas P. J. Day,2 Livia Visai,3 Pietro Speziale,3 Alex Lam,4† Mark Pallen4† and Timothy J. Foster1

1Department of Microbiology, Moyne Institute of Preventive Medicine, Trinity College, Dublin 2, Ireland
2Nuffield Department of Clinical Laboratory Sciences, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK
3Department of Biochemistry, University of Pavia, 27100 Pavia, Italy
4Department of Microbiology and Immunobiology, The Queen’s University of Belfast, Grosvenor Road, Belfast BT12 6BN, UK

Twenty-one genes encoding surface proteins belonging to the LPXTG family have been identified by *in silico* analysis of six *Staphylococcus aureus* genome sequences. Eleven genes encode previously described proteins, while 10 have not yet been characterized. Of these, eight contain the cell-wall sorting signal LPXTG responsible for covalently anchoring proteins to the cell-wall peptidoglycan. The remaining two, SasF and SasD, harbour a single residue variation in the fourth position of the LPXTG motif (LPXAG). Western blotting of lysostaphin-solubilized *S. aureus* cell-wall proteins demonstrated the release of SasF in the cell-wall fraction, indicating that proteins carrying LPXAG are sorted normally. Analysis of primary sequences of the *Staphylococcus aureus* surface (Sas) proteins indicated that several share a similar structural organization and a common signal sequence with previously characterized LPXTG proteins of *S. aureus* and other Gram-positive cocci. Protein SasG has 128 residue B repeats that are almost identical at the DNA level. PCR analysis indicated that recombinants with repeat length variations are present in the bacterial population whereas they are not detectable in the B-repeat-encoding region of *sdrD*. The *sasG* and *sasH* genes are significantly associated with invasive disease isolates compared to nasal carriage isolates. Several IgG samples purified from patients recovering from *S. aureus* infections had higher titres against Sas proteins than control IgG, suggesting that expression occurred during infection in some patients.

**INTRODUCTION**

*Staphylococcus aureus* expresses cell-wall-associated surface proteins which play important roles in virulence. The ability to bind to host extracellular matrix and plasma components promotes adhesion to host tissues, evasion of host defence mechanisms and invasion of epithelial and endothelial cells (Foster & Hooûk, 1998; Nilsson *et al*., 1998; Peacock *et al*., 1999). More recently, surface proteins have also been shown to play roles in biofilm formation (Cucarella *et al*., 2001) and binding of host transferrin (Taylor & Heinrichs, 2002). Cell-wall-associated surface proteins share several common features that enable them to be covalently anchored to cell-wall peptidoglycan. They possess an N-terminal signal peptide required for Sec-dependent secretion and a conserved C-terminal cell-wall sorting signal which is essential for attachment of a protein to the cell wall by sortase (Mazmanian *et al*., 2001; Navarre & Schneewind, 1999). The C-terminal sorting signal comprises a conserved LPXTG motif followed by a hydrophobic stretch of amino acids and positively charged residues at the extreme C terminus. Sortase, encoded by *srtA*, is a membrane-bound transpeptidase that cleaves polypeptides between the threonine and glycine residues of the LPXTG motif and covalently links
them to the nascent pentaglycine crossbridge in peptidoglycan (Ton-That et al., 1999). Mutants lacking srtA are defective in the display of surface proteins and are attenuated in animal infection models (Mazmanian et al., 2000). Interrogation of the many microbial genome sequences now available showed that LPXTG proteins and homologues of the srtA gene exist in many Gram-positive bacteria (Janulczyk & Rasmussen, 2001; Pallen et al., 2001).

*S. aureus* is known to express 11 LPXTG proteins including protein A (Spa), the clumping factors ClfA and ClfB, the collagen-binding protein Cna, the serine aspartate repeat proteins SdrC, SdrD and SdrE, the fibronectin-binding proteins FnbA and FnbB (reviewed by Foster & Höök, 1999), the plasmin-sensitive protein Pls (Savolainen et al., 2001) andFmtB (Komatsuzawa et al., 2000). Many share a common domain organization with a similar sized N-terminal A region (ca 500 residues) which contains ligand binding activity. The ClfA and ClfB A regions are composed of independently folded subdomains consisting primarily of β-sheets and coils with a small amount of α-helix (Perkins et al., 2001; Deivanayagam et al., 2002). Repeat regions occur as dipeptides in SD repeats of Sdr proteins (Josefsson et al., 1998a) that act as a stalk to project the N-terminal A region away from the bacterial cell surface (Hartford et al., 1997) or larger tandem repeated domains which vary in size and identity between proteins. The B repeats of Sdr proteins bind Ca²⁺ and form rigid rod-like structures (Josefsson et al., 1998b) while the D repeats of the FnbA and FnbB proteins are flexible and bind fibronectin (McGavin et al., 1993; House-Pompeo et al., 1996).

The aims of this study were (i) to identify the diversity of LPXTG surface proteins from *S. aureus* genome sequences, (ii) to study the structural organization of each protein, (iii) to determine if any surface protein genes are associated with invasive disease, and (iv) to measure antibody titres in convalescent patients’ sera. This study has identified 10 novel LPXTG proteins from *S. aureus* genome sequences which are described using the nomenclature *Staphylococcus aureus* surface (Sas) protein (Mazmanian et al., 2001).

**METHODS**

**Prediction of LPXTG proteins.** Six finished and unfinished *S. aureus* genome sequences were searched for open reading frames (ORFs) with potential to encode sortase substrates [strains N315, Mu50 (Kuroda et al., 2001); COL (http://www.tigr.org/); EMRSA-16, MSSA (http://www.sanger.ac.uk/Projects/S_aureus/)] (Enright et al., 2000; NCCTC 8325 (http://dna1.chem.ou.edu/)). A variety of approaches was used to find novel sortase substrates, including (1) pattern searches with the motif [L/R][P/T][S/A][G/N][A/T][X(0,10)][DE][Q/K/N/R][DE][Q/K/N/R][DE][Q/K/N/R][DE][Q/K/N/R][DE][Q/K/N/R][DE][Q/K/N/R][DE][Q/K/N/R][DE][Q/K/N/R] (written according to the Prosite conventions: http://expasy.cbr.nrc.ca/tools/scnpsiti3.html), which finds an LPXTG-like motif followed by at least 10 non-polar residues (capable of spanning the membrane) and at least one polar or charged residue within the last five residues; and (2) BLAST and PSI-BLAST searches on our ViruloGenome site (http://www.vgc.ac.uk/) looking for regions similar to the C termini of known LPXTG-linked proteins (using a higher than usual e value threshold of 1·0 for inclusion of new sequences in subsequent iterations). New LPXTG-like motifs were accepted only if they were adjacent to a transmembrane domain followed by a charged C-terminal tail (Mazmanian et al., 2001).

**Primary structure analysis.** On-line bioinformatic tools were used to characterize protein sequences identified from the *in silico* analysis. Signal peptides and repeat domains were identified using the SIGNALP (http://www.cbs.dtu.dk/services/SignalP/) and REPRO (http://mathbio.nimr.mrc.ac.uk/~tgeorge/repro/) algorithms. Similarity searches were carried out using NCBI BLAST (Altschul et al., 1997) and protein sequence alignments used CLUSTAL W (EMBL) (Thompson et al., 1994). ARTEMIS, a genome visualization program available from the Sanger Centre (http://www.sanger.ac.uk/Software/Artemis/), was used to examine gene localization. Protein secondary structure was predicted from a consensus of algorithms offered at http://npa-pbl.ibcb.fr/cgi-bin/npsa_automat.pl?page=npsa_spред.html and http://www.expasy.ch/ sites.

**PCR analysis of variable repeat regions.** The entire ORFs of *sasG* and *sasD* were PCR-amplified from strain 8325-4 (Novick, 1967) genomic DNA using primers detailed in Table 1.

**Western blotting of the lysostaphin-solubilized cell-wall fraction.** *S. aureus* 8325-4 spa (McDevitt et al., 1995) cultures were grown to exponential phase in tryptic soy broth (Difco) with aeration (OD₆0₀ value of 0·8). Cells were harvested by centrifugation at 7000 g at 4 °C for 15 min and washed with PBS. Cells were concentrated by resuspension in lysis buffer (50 mM Tris/HCl, 20 mM MgCl₂, pH 7·5) supplemented with 30% raffinose at an OD₆0₀ value of 40. Cell-wall proteins were solubilized by incubation with lysostaphin (200 μg ml⁻¹) at 37 °C for 20 min in the presence of protease inhibitors (mini Complete; Roche Molecular Biochemicals). Pretoplasma were recovered by centrifugation at 6000 g for 20 min, and the supernatant was taken as the wall fraction. Samples were prepared for electrophoresis by boiling for 5 min in final sample buffer [0·125 M Tris/HCl, 4% (w/v) SDS, 20% (v/v) glycerol, 10% (v/v) β-mercaptoethanol, 0·002% (w/v) bromophenol blue] and analysed at 10% (w/v) acrylamide gels. Gels were transferred to PVDF Western blotting membranes (Boehringer) by the semidy method (Bio-Rad). Membranes were blocked for 15 h at 4 °C with 5% (w/v) skim milk. Rabbit polyclonal anti-SasF antibodies were used at a 1:1000 dilution. Bound antibody was recognized by incubation with protein A-peroxidase (Sigma, 1:2000) and visualized using the Chemiluminescence Detection System (Roche Molecular Biochemicals).

**Construction and purification of histidine-tagged fusion proteins.** Recombinant N-terminal domains of Sas proteins were expressed with hexahistidine affinity tags at their N termini using the expression vector pQE30 (Qiagen). Each gene fragment was amplified by PCR using 8325-4 genomic DNA as the template, except for *sasK*. Out of the six genomes studied *sasK* is only present in Mu50 and N315. In the absence of these strains, a strain of the same clonal type, EMRSA-3 (B. Cookson, personal communication), was used (Enright et al., 2002). The forward and reverse primers are listed in Table 1. The PCR fragments were cloned into pQE30 at the BarnHI and HindIII sites and the resulting plasmids transformed into *Escherichia coli* TOPP-3 (Strategene) for protein expression. The recombinant proteins were purified using Ni²⁺-chelate chromatography (O’Connell et al., 1998). The SasF40-211 construct was used to immunize a rabbit to generate anti-SasF antibodies.

**Isolation of IgG from convalescent patients’ sera.** Antisera from 33 individuals with *S. aureus* infections were obtained from the Ospedale di Circolo di Varese, Italy. The patients were diagnosed with a variety of staphylococcal infections, including sepsis,
Table 1. Primer sequences used to construct recombinant proteins and to amplify repeat regions

<table>
<thead>
<tr>
<th>Recombinant protein</th>
<th>Primer sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>rSasA91-575</td>
<td>F5'-CGGGGATCCCGCTTTGATGCACCATTAC-3'</td>
</tr>
<tr>
<td></td>
<td>R5'-CCCCAAGCTTATCATGAAAAATATGTC-3'</td>
</tr>
<tr>
<td>rSasG50-428</td>
<td>F5'-CGGGGATCCCGCACTGAAAACATATTG-3'</td>
</tr>
<tr>
<td></td>
<td>R5'-CCCCAAGCTTATCATGAAAAATATGTC-3'</td>
</tr>
<tr>
<td>rSasI40-321</td>
<td>F5'-CGGGGATCCGCACTGAAAACATATTG-3'</td>
</tr>
<tr>
<td></td>
<td>R5'-CCCCAAGCTTATCATGAAAAATATGTC-3'</td>
</tr>
<tr>
<td>rSasF40-211</td>
<td>F5'-CGGGGATCCGCACTGAAAACATATTG-3'</td>
</tr>
<tr>
<td></td>
<td>R5'-CCCCAAGCTTATCATGAAAAATATGTC-3'</td>
</tr>
<tr>
<td>rSasH39-253</td>
<td>F5'-CGGGGATCCGCACTGAAAACATATTG-3'</td>
</tr>
<tr>
<td></td>
<td>R5'-CCCCAAGCTTATCATGAAAAATATGTC-3'</td>
</tr>
<tr>
<td>rSasJ48-477</td>
<td>F5'-CGGGGATCCGCACTGAAAACATATTG-3'</td>
</tr>
<tr>
<td></td>
<td>R5'-CCCCAAGCTTATCATGAAAAATATGTC-3'</td>
</tr>
<tr>
<td>rSasE36-292</td>
<td>F5'-CGGGGATCCGCACTGAAAACATATTG-3'</td>
</tr>
<tr>
<td></td>
<td>R5'-CCCCAAGCTTATCATGAAAAATATGTC-3'</td>
</tr>
<tr>
<td>rSasC38-432</td>
<td>F5'-CGGGGATCCGCACTGAAAACATATTG-3'</td>
</tr>
<tr>
<td></td>
<td>R5'-CCCCAAGCTTATCATGAAAAATATGTC-3'</td>
</tr>
<tr>
<td>rSasK27-175</td>
<td>F5'-CGGGGATCCGCACTGAAAACATATTG-3'</td>
</tr>
<tr>
<td></td>
<td>R5'-CCCCAAGCTTATCATGAAAAATATGTC-3'</td>
</tr>
<tr>
<td>sdrD gene</td>
<td>F5'-CAATAAGAAGGAAAAAGGCTAG-3'</td>
</tr>
<tr>
<td></td>
<td>R5'-ACCGAATACAAATAATGAC-3'</td>
</tr>
<tr>
<td>sasG gene</td>
<td>F5'-GAATAAGAAGGAAAAAGGCTAG-3'</td>
</tr>
<tr>
<td></td>
<td>R5'-TTAATCTTCTTCTTACG-3'</td>
</tr>
</tbody>
</table>

Statistical analysis. Patients (n = 33) were divided into three groups according to age, group I (20–45 years), group II (45–65 years) and group III (65–90 years). One-way ANOVA was used to compare variability in patient IgG response within each age group to each protein.

ELISA assay. IgG from sera was tested for reactivity with recombinant SasG50-428, SasA91-575, SasI40-321, SasF40-211, SasK27-175, SasC38-432, SasH39-253, SasJ48-477 and SasE36-292. Microtitre wells were incubated overnight with 100 μl of 50 mM sodium carbonate, pH 9.5, containing 10 μg ml−1 of each recombinant protein. To block additional protein-binding sites, the wells were treated for 1 h at 22 °C with 200 μl PBS containing 2 % (w/v) BSA. The wells were then washed five times with PBS (0-1% Tween 20 in PBS) and incubated with 2 μg of antibody present in 100 μl of 2 % BSA in PBS at 22 °C. Unbound antibody was removed by washing the wells five times with PBS. Bound antibody was detected by incubation (1 h at 37 °C) of the plates with a rabbit anti-human IgG conjugated to horseradish peroxidase (Dako, Gostrup, Denmark). After washing, binding was quantified using the substrate 4-chloro-1-naphthol (Sigma) and measuring the absorbance at 492 nm in a microplate reader (Bio-Rad). An ELISA reading of greater than two times the mean of the controls was deemed to represent a response to recent exposure to the antigen.

Bacterial strains. The presence of nine genes encoding putative cell-wall-anchored proteins was examined in 155 isolates recovered from patients with invasive S. aureus disease (94 hospital-acquired and 61 community-acquired) and 179 isolates recovered from healthy blood donors. These isolates were collected within Oxfordshire, UK, between 1997 and 1998 using a prospective case control design (Peacock et al., 2002). They include EMRSA-16 and MSSA strains being sequenced by the Sanger Centre. Invasive isolates were recovered from a normally sterile body site in the Diagnostic Microbiology Laboratory, John Radcliffe Hospital, Oxford. Of these, 86 % were blood culture isolates and 14 % were from cerebrospinal fluid or tissue recovered at operation or biopsy from sites of deep infection such as osteomyelitis and deep abscesses. They include EMRSA-16 and MSSA strains being sequenced by the Sanger Centre.

PCR screening of sas genes in clinical isolates. The presence of sas genes was analysed by PCR amplification using primers designed to non-variable regions within each gene (Table 2). Genomic DNA from each bacterial strain was isolated using the Wizard Genomic DNA Purification kit (Promega) with the addition of lysostaphin (30 μg ml−1) at the cell lysis step. The PCR mixtures for all nine sets of primers were as follows: 1 × reaction buffer, 1-5 mM MgCl2, 100 pmol forward and reverse primers, 100 ng template DNA, 200 μM dNTP mix and 2-5 U Taq polymerase. The PCR cycling conditions were as follows: 94 °C for 3 min, followed by 30 cycles of 94 °C for 1 min, 50 °C for 1 min and 72 °C for 1 min, which was suitable for amplifying the same size PCR product from each gene. Aliquots of the reaction mixtures were analysed by 1 % agarose gel electrophoresis. Positive controls, EMRSA-16 and MSSA, and a negative control (reaction mixture minus DNA) were included in each PCR run.

Statistical analysis. The PCR results were initially analysed according to origin (nasal carriage isolate or invasive), and further
subdivided depending on whether the disease was hospital- or community-acquired. Contingency tables were used to compare proportions between groups. To adjust for any differences in the clonal structures of the disease and carriage isolate populations, analyses were stratified by MLST-defined lineage using the Mantel–Haenszel method. To check that any associations were independent, conditional logistic regression was used to adjust for linkage disequilibrium between genes. A lineage was defined as a group of organisms that had the same sequence as at least one other member of the group at five or more of the seven MLST loci examined. Strains that were unrelated to any other at five of seven loci were arbitrarily grouped as a ‘lineage’ for the purpose of this stratification. No corrections were made for multiple comparisons, though statistical significance was set conservatively at \( P = 0.01 \) and 99% confidence intervals were used throughout.

### RESULTS

#### Identification of LPXTG proteins

Six finished and unfinished \textit{S. aureus} genome sequences were searched for ORFs with potential to encode surface proteins genes belonging to the LPXTG family. Proteins were identified that had an LPXTG motif followed by a putative membrane-spanning hydrophobic domain and several charged residues at the C terminus. Twenty-one LPXTG proteins were identified, 10 of which had not been previously described. Nine of the novel LPXTG proteins were reported in a review (Mazmanian \textit{et al}., 2001) but they were not characterized. We have adopted the terminology proposed by Mazmanian \textit{et al}. (2001), \textit{Staphylococcus aureus} surface (Sas) protein. SasK studied here for the first time is only present in strains Mu50 (GenBank accession no. BAB58757) and N315 (GenBank accession no. BAB43686). We did not study SasD.

#### A conserved motif in the signal sequences

A feature of proteins belonging to the LPXTG family is the presence of an N-terminal secretory signal sequence. Signal sequences were identified using the \textsc{signalp} prediction algorithm. When the signal sequences from all 21 \textit{S. aureus} LPXTG proteins were aligned, 15 were found to contain a conserved sequence, (Y/F)SIRK, or variants thereof (data not shown). Subsequent \textsc{blast} searches confirmed the observations by Tettelin \textit{et al}. (2001) that this motif is common among sortase substrates from Gram-positive cocci, although apparently neither exclusive to nor universal among these proteins. SasA is encoded by a gene within an accessory \textit{secA2/secY2} cluster. As noted by Bensing & Sullam (2002) for the GspA LPXTG protein from \textit{Streptococcus gordonii}, which is encoded in a similar \textit{secA2/secY2} cluster, the signal sequence of SasA is predicted by \textsc{signalp} to be unusually long (90 residues). This suggests that the accessory \textit{SecA2/SecY2} system might be required for the export of SasA in \textit{Staphylococcus aureus}, as is the case for GspB in \textit{Streptococcus gordonii} (Bensing & Sullam, 2002). All of the other uncharacterized LPXTG proteins had conventional signal peptides.

#### Primary characterization of the Sas proteins

\textbf{SasG.} The primary structure of SasG resembles that of proteins PIs of \textit{Staphylococcus aureus} and Aap of \textit{Staphylococcus epidermidis} (GenBank accession nos P80544 and CAB77251, respectively; Fig. 1). A distinct subdomain organization of the A regions of these proteins is evident upon primary and secondary structure analysis. PIs and Aap both contain a repeat domain at the extreme N terminus of the A region which is lacking in SasG. All three proteins contain a previously undescribed conserved domain at the C terminus of the A region which is either 212 residues (SasG and Aap) or 217 residues (PIs) in length (Fig. 1). In SasG this domain lies between residues 208–420 and is 52% identical to that of PIs and 59% identical to that of Aap. Interestingly, the N-terminal 181 residues of this domain also occur within the A region of SasA (residues 296–476; Fig. 1) and within the N terminus of a non-LPXTG protein of \textit{Bacillus thuringiensis} (GenBank accession no. P56957, Cry22A) and one of \textit{Lactococcus lactis} (GenBank accession no. AAK06279, YwH), showing 32, 36 and 40% amino acid identity with the SasG domain, respectively. Fig. 2 shows an alignment of the conserved part of this domain from all six proteins (56–60 residues in length). The A regions of SasG, PIs and Aap appear to be composed of subdomains, with those of PIs and Aap comprising a repeat domain and a non-repeat domain followed by a 212/217 residue conserved domain. The A domain of SasG is smaller and appears to comprise two subdomains, a unique N-terminal domain and a domain that is related to the conserved domain of PIs and Aap.

Another common feature of these three proteins is the 128 residue B repeats (followed by a partial B repeat of 71 aa in the case of SasG). Individual repeats in SasG show a high

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>sasA F</td>
<td>5'-GCATGACCAGCAAGCTTTTG-3'</td>
</tr>
<tr>
<td>sasA R</td>
<td>5'-GTTGGAAGCAGATTTGATG-3'</td>
</tr>
<tr>
<td>sasI F</td>
<td>5'-CGATACCAAGAAGCTAAACG-3'</td>
</tr>
<tr>
<td>sasI R</td>
<td>5'-CCTTTTTGTTTCCATCACTAC-3'</td>
</tr>
<tr>
<td>sasF F</td>
<td>5'-CGAGGCTAAAGTTTGAGATC-3'</td>
</tr>
<tr>
<td>sasF R</td>
<td>5'-GCTGGATCTTCCGATGTATC-3'</td>
</tr>
<tr>
<td>sasJ F</td>
<td>5'-GTCAAGGTCCTAAAGGAAA-3'</td>
</tr>
<tr>
<td>sasJ R</td>
<td>5'-GGCAAACCTTTTGTCACTTC-3'</td>
</tr>
<tr>
<td>sasG F</td>
<td>5'-GGGAACTCAACAGAGCCG-3'</td>
</tr>
<tr>
<td>sasG R</td>
<td>5'-GAGAAAGGAGCTTTTCTAACG-3'</td>
</tr>
<tr>
<td>sasH F</td>
<td>5'-GTGTAATGGAGATTGCAA-3'</td>
</tr>
<tr>
<td>sasH R</td>
<td>5'-CGTTGCTGTGAGGTTG-3'</td>
</tr>
<tr>
<td>sasE F</td>
<td>5'-GCAACAGAAGCTCAAGG-3'</td>
</tr>
<tr>
<td>sasE R</td>
<td>5'-GATTGTTCTAGTTGATCCG-3'</td>
</tr>
<tr>
<td>sasC F</td>
<td>5'-GCGAGTCTGGAAAGG-3'</td>
</tr>
<tr>
<td>sasC R</td>
<td>5'-CGCATCTTTCTACACACTC-3'</td>
</tr>
<tr>
<td>sasK F</td>
<td>5'-CGTTGAGCACAAGCCGATG-3'</td>
</tr>
<tr>
<td>sasK R</td>
<td>5'-ATAAAGACGCACTATTGCTG-3'</td>
</tr>
</tbody>
</table>
Fig. 1. Schematic representation of novel surface proteins of *S. aureus* showing the positions and sizes of signal sequences (S), A regions (denoted by a dashed line), B repeat regions (Bn) and wall-/membrane-spanning regions (WM). The vertical thin lines represent the SD or SX dipeptide repeat domains. Thick black lines denote smaller amino acid repeat regions. The conserved domain within the Pls, Aap, SasG and SasA A regions is marked by a dotted box. The site of proteolytic cleavage within the Pls and SasA A domains is indicated by a vertical arrow.

Fig. 2. CLUSTAL W multiple-sequence alignment of the N-terminal 56–60 residues of a conserved domain found within six proteins from *S. aureus*, *S. epidermidis*, *B. thuringiensis* and *L. lactis*. Residues marked in bold and with an asterisk are highly conserved among all six proteins. Residues marked with a colon (:) indicate functionally conserved residues and those marked with a single dot (.) indicate functionally similar residues.
yielded a 4 kb band, corresponding to full-length sasG, followed by a set of minor fragments that differed from one another by approximately 400 bp, which corresponds to the length of one repeat (Fig. 3a). DNA from the fainter bands was gel-purified and used as a template for PCR with primers designed to amplify DNA encoding the sasG A region. In addition, DNA from the 4-8 kb band representing the full-length sasG gene was gel-purified and used as a template for PCR with primers designed to re-amplify the entire sasG gene. In each case a single band was detected (Fig. 3b). These experiments demonstrate that the minor bands in Fig. 3(a) were sasG-specific products and were not due to mispriming of the repeat DNA during PCR amplification. These data suggest that frequent recombination occurs between repeat DNA of sasG consistent with the sequence being 98–100 % identical at both the amino acid and DNA level.

In contrast, PCR amplification of the sdrD gene from strain 8325-4 yielded only one PCR product (Fig. 3c). The DNA encoding the sdrD B repeats are 68–78 % identical between repeats. Since repeat length variants cannot be detected by PCR it appears that this lower level of identity between B repeats is not sufficient for frequent recombination to occur.

The sasG gene resides on a 5 kb fragment which is absent from strain EMRSA-16. This region bears two sar homologues, sarH2 and sarH3, and a gene of unknown function. There is a high level of amino acid sequence identity (87–99 %) between the SasG proteins from 8325, COL, N315 and Mu50, whereas SasG from MSSA is only 55–69 % identical to these. Residues 50–206 of MSSA SasG are more conserved (96 % identical) than residues 207–410, which correspond to the domain conserved in Pls and Aap. This protein has a similar domain organization to proteins of the Sdr family (Josefsson et al., 1998a). It contains a large N-terminal A region which we propose to be composed of three separately folded subdomains, N1 (residues 91–244), N2 (residues 245–476) and N3 (residues 477–575). Residues 91–244 of SasA contain a high proportion of hydrophilic residues similar to domain N1 of ClfA and ClfB (28 % serine compared to 9 % serine in the rest of the A region; Perkins et al., 2001; Deivanayagam et al., 2002). This is probably responsible for the aberrant migration of recombinant proteins on SDS-PAGE gels. Furthermore, treatment of full-length recombinant SasA A region with purified metalloprotease, a treatment which removes domain N1 of ClfA and ClfB (McAleese et al., 2001), resulted in loss of residues 91–244 from the full-length molecule (data not shown). Domain N2 carries the 181 residue conserved domain, the C-terminal end of which possibly represents the border between N2 and N3.

### Table 3. Variable numbers of repeats in SasA and SasG

<table>
<thead>
<tr>
<th>Protein</th>
<th>Numbers of variable repeats within each genome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8325</td>
</tr>
<tr>
<td>SasA SX dipeptide repeats</td>
<td>731</td>
</tr>
<tr>
<td>SasG B repeats</td>
<td>8-2</td>
</tr>
</tbody>
</table>

**Fig. 3.** Analysis of length variants of sasG and sdrD using PCR. (a) PCR amplification of the full-length sasG gene from genomic DNA isolated from strain 8325-4 (lane 2). (b) Lane 2, PCR amplification of the full-length sasG gene from gel-purified full-length sasG gene PCR product from (a) lane 2; lane 3, PCR amplification of the sasG A region from gel-purified smaller sasG variants from (a) lane 2. (c) PCR amplification of the sdrD gene from genomic DNA of strain 8325-4 (lane 2). Lane 1 in (a, b, c) contains molecular size markers.
Two 88 residue B repeats which are 42% identical occur after region A. Unlike SasG, there is no variation in the number of these repeats in SasA from all six sequenced strains. These repeats are 45–57% identical to repeats that occur in the Bhp protein from S. epidermidis (GenBank accession no. AAK29746). The A and B domains of SasA are linked to the cell-wall-anchoring domain by SX dipeptide repeats, which resemble the SD dipeptide repeats of the Clf–Sdr protein family. The length of the SX repeat varies from strain to strain (Table 3).

Other Sas proteins. Many of the remaining uncharacterized Sas proteins also seem to comprise a modular organization. SasC has a 585 residue N-terminal A region with a similar predicted secondary structure to A regions of Clf–Sdr proteins. Therefore, it is likely also to have a subunit organization. Region A is followed by two 300 residue repeat domains with 38% identity (Fig. 1). At the carboxy terminus of the second repeat lies a 12 aa motif [AT(D/T)EEKQ(A/V)/A(L/V)NQ] which is repeated twice. The N terminus of the protein (residues 38–1305) is 49% identical to the FmtB protein involved in expression of methicillin resistance (Komatsuzawa et al., 2000).

SasI lacks a typical A region but contains two 140 residue repeat domains at the N terminus which are 38% identical (Fig. 1). Hydropathy analysis predicts a 60 residue hydrophilic domain in the centre of the protein (residues 480–540).

SasF is predicted to be composed of α-helices with a potential coiled-coil domain at the N terminus. The LPXTG motif of this protein varies by a single residue from the canonical sequence (LPKAG). To determine if this variation of the classical LPXTG motif enables SasF to be sorted and covalently anchored to the cell-wall peptidoglycan, SasF-specific antisera were used to probe the cell-wall fraction of exponentially grown 8325-4 cells by Western blotting. A single protein band of 65 kDa was detected (Fig. 4) suggesting that the SasF protein is cell-wall-associated and has undergone sorting despite containing a variation in the LPXTG motif.

The gene encoding SasK occurs only in N315 and Mu50. It lies in a 2 kb region which is absent from the other S. aureus genomes studied. An ORF encoding a protein with 40% identity to a glutamyl endopeptidase from Bacillus intermedius lies immediately upstream from sasK. There is no intergenic sequence between the ORFs, suggesting they are co-transcribed. SasK is a 204 aa protein bearing two 13 residue repeats.

SasE, SasD, SasH and SasI do not carry repeat domains and are present in all six genome sequences. The sasE and sasI genes encode proteins that have been implicated in transferrin binding (Taylor & Heinrichs, 2002). They lie adjacent to and divergent from an iron-regulated operon containing the gene for sortase B (Mazmanian et al., 2002).

SasH has 32% identity to a 5’ nucleotidase from Bacillus halodurans.

ELISA-based screen using convalescent patients’ sera

To determine if the newly identified surface proteins are expressed during growth of the bacteria in vivo, recombinant N-terminal truncates from SasG, SasA, SasI, SasF, SasK, SasC, SasH, SasJ and SasE were purified and used in ELISA-based assays to screen for antibodies in sera from patients who had recovered from S. aureus infections. IgG isolated from sera of 33 patients was tested by ELISA for reactivity with the recombinant proteins (Fig. 5). We observed higher titres of antibodies to SasG, SasA, SasI, SasH and SasJ compared to the control IgG, suggesting that these proteins are expressed during S. aureus infections. Sera from individual patients varied substantially in titre towards particular proteins. For example, IgG from patients 3, 12, 16 and 17 did not react with SasG. However, IgG from patients 1, 10, 13 and 30 had higher titres against SasG. Likewise, patients 4, 5, 10, 11 and 13 had high titres against SasA compared to that of IgG purified from patients 20, 23 and 32. Similar variability was observed when the same panel of IgG was tested for the reactivity with fibronectin-binding protein (Casolini et al., 1998).

Variation in IgG reactivity for each protein across the patient group was also observed. To determine if this variability was affected by patient age the one-way ANOVA test was used. The patient group (n = 33) was divided into three age groups and for each group a P value related to all the proteins was calculated. Results clearly demonstrated that the variability of IgG reactivity is a function of the patient’s age, as patients belonging to the younger age group (20–45 years) showed less variability (P = 0.00001) in
recognizing each protein compared to patients belonging to older age groups (45–65 years, \( P = 2.44 \times 10^{-3} \), 65–90 years, \( P = 1.11 \times 10^{-16} \)).

**The **sasG** and **sasH** genes are positively associated with disease isolates**

To determine if any of the novel surface protein genes are more commonly associated with disease, the presence or absence of the nine genes was determined for 155 isolates associated with invasive disease and 179 carriage isolates obtained from healthy blood donors. Genomic DNA was analysed by PCR with primers specific for each gene. On univariate analysis sasG and sasH were significantly more common in the invasive group (Table 4). To adjust for any effects of clonality, the prevalence of each gene in the disease and carriage groups was compared within but not between clonal complexes using the Mantel–Haenszel method. The clonal structure of this collection of isolates has been defined by multilocus sequence typing (http://www.mlst.net). The presence of sasG and sasH remained significantly associated with disease isolates (Table 4). When isolates from nosocomial infections were removed from the analysis and community-acquired disease isolates were compared with carriage strains in a stratified analysis, both sasG and sasH remained associated with disease [Odds ratios 3·1 (\( P = 0.003 \)) and 3·4 (\( P = 0.005 \)), respectively]. We examined the possibility that identification of these virulence-associated genes was actually the result of linkage disequilibrium. Conditional logistic regression modelling of the relationship of the two factors both to disease and to each other suggested that sasG and sasH are both independent of each other and are independently associated with disease.

**DISCUSSION**

With the availability of many completed microbial genome sequences it is possible by *in silico* analysis to identify groups or families of proteins by screening the translation products of whole genomes for conserved domains or consensus motifs. In this study, we have analysed the genome sequences of six strains of *S. aureus* that are in the public domain for ORFs specifying proteins with potential to be anchored to the cell wall by sortase. Ten previously uncharacterized sequences were identified, bringing the total number of LPXTG proteins of *S. aureus* to 21. One of these, sasK, is only present in strains Mu50 and N315, while sasG is present in all strains except EMRSA-16. The other eight genes are present in all six strains.

It is well established that proteins bearing an N-terminal signal peptide and C-terminal cell-wall sorting signal with a conserved LPXTG motif are anchored to the cell-wall envelope by a transpeptidation mechanism involving sortase
Most of the newly characterized LPXTG proteins examined in this study bear the typical features of this family of proteins and hence are predicted to be cell-wall-associated. Exceptions include SasD and SasF, which contain an LPXAG motif within the cell-wall sorting signal. Using SasF-specific antibodies, Western blotting of the cell-wall fraction of 8325-4 cells detected a single 65 kDa protein corresponding to the predicted molecular mass of the SasF protein. Since SasF can be solubilized by lysozyme during the generation of stable protoplasts, it suggests that SasF is associated with the cell-wall fraction and infers that sorting has occurred. These data therefore indicate that a substitution of threonine with an alanine residue within the LPXTG motif does not affect the sorting reaction. We propose that SasD is also cell-wall-associated.

The domain architecture of proteins is well documented (Doolittle, 1995). In this study, primary structural analysis strongly suggests that proteins of the LPXTG family are modular in architecture and have evolved through the acquisition of distinct domain-sized polypeptides, some of which have expanded by duplication and homologous recombination. Many of these domains are found in other species, such as the B repeats of SasA and SasG, the 212 residue conserved domain of SasG, the SD and SX repeats of the Sdr proteins and the (Y/F)SIRK motif containing signal peptide, suggesting they have been acquired by recent horizontal transfer.

The conserved (Y/F)SIRK motif is present in the signal peptides of many S. aureus and other Gram-positive LPXTG proteins. The possibility that this sequence targets the protein to sortase is unlikely, since it is also found in non-cell-wall-anchored proteins and more importantly, some S. aureus LPXTG proteins (Cna, SasF, SasK, SasH and SasA) do not carry the motif. Therefore, the significance of this motif and its function are not clear. SasA bears an unusually long signal peptide that probably targets this protein for secretion by the accessory SecA2/SecY2 system (Bensing & Sullam, 2002), although why this should require an accessory secretion system is also unclear. Specific amino acid patterns have previously been identified within signal peptides of proteins of similar function or of proteins destined for different cellular compartments (Edman et al., 1999). Expression of hybrid proteins with different signal peptide types within wild-type and SecA2 mutants could address the effects of the (F/Y)SIRK motif of and the SasA extended signal sequence on efficiency and mechanism of secretion.

Another typical domain arrangement of LPXTG proteins is the modular A region. Most LPXTG proteins of S. aureus contain a ~500 residue A region, which has recently been proposed to be composed of three independently folded subdomains in the case of ClfA and ClfB (Perkins et al., 2001; Deivanayagam et al., 2002). N1 is extremely hydrophilic with an elongated secondary structure, whereas N2 and N3 are globular proteins bearing the ligand binding activity. As bacterial cells enter stationary phase, domain N1 of both proteins is cleaved from the bacterial cell surface by a staphylococcal metalloprotease and in the case of ClfB, ligand binding activity is lost (McAleece et al., 2001; J. Higgins & T. J. Foster, unpublished data). The site of cleavage is a predicted α-helix which separates domains N1 and N23. A similar phenomenon occurs with P1s expression on the bacterial cell surface. As cells progress to stationary phase, full-length P1s (230 kDa) is proteolytically cleaved to a 175 kDa truncated form (Savolainen et al., 2001). The cleavage site (between residues 387 and 388) may represent the border between subdomain N2 (residues 49–387) and subdomain N3 (residues 388–655) bearing the 217 residue conserved domain (Fig. 1). The SasA A region is also predicted to have three subdomains. The architecture of this region is very similar to ClfA and ClfB. Like these proteins, SasA has an extremely hydrophilic N1 subdomain that is proteolytically cleaved off from full-length recombinant SasA region A by metalloprotease. The site of cleavage is also a predicted α-helix. It is therefore possible that similar processing occurs in SasA as with ClfA and ClfB.

### Table 4. PCR results for LPXTG genes in carriage and disease isolates – univariate analysis and analysis stratified by clonal complex

<table>
<thead>
<tr>
<th>LPXTG gene</th>
<th>Carriage (n=179)</th>
<th>Invasive (n=155)</th>
<th>Odds ratio (99% CI)</th>
<th>P value</th>
<th>Stratifed analysis</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>sasA</td>
<td>179 (100%)</td>
<td>155 (100%)</td>
<td>0-6 (0-22 to 2-6)</td>
<td>0-43</td>
<td>0-68 (0-17 to 2-8)</td>
<td>0-49</td>
</tr>
<tr>
<td>sasI</td>
<td>173 (97%)</td>
<td>147 (95%)</td>
<td>2-1 (0-34 to 13)</td>
<td>0-35</td>
<td>1-6 (0-55 to 4-7)</td>
<td>0-25</td>
</tr>
<tr>
<td>sasE</td>
<td>159 (89%)</td>
<td>142 (92%)</td>
<td>2-8 (1-5 to 5-3)</td>
<td>&lt;0-0001</td>
<td>2-1 (0-30 to 14)</td>
<td>0-32</td>
</tr>
<tr>
<td>sasH</td>
<td>172 (96%)</td>
<td>152 (98%)</td>
<td>2-5 (1-2 to 5-1)</td>
<td>0-0006</td>
<td>3-3 (1-6 to 6-7)</td>
<td>0-0001</td>
</tr>
<tr>
<td>sasG</td>
<td>121 (68%)</td>
<td>130 (84%)</td>
<td>1-4 (0-76 to 2-5)</td>
<td>0-18</td>
<td>1-4 (0-29 to 6-8)</td>
<td>0-58</td>
</tr>
<tr>
<td>sasF</td>
<td>63 (35%)</td>
<td>66 (43%)</td>
<td>1-6 (0-56 to 4-7)</td>
<td>0-32</td>
<td>2-0 (0-56 to 7-0)</td>
<td>0-15</td>
</tr>
</tbody>
</table>

*Odds ratio and P value after Mantel–Haenszel stratification for MLST-defined lineage.*
Repeated sequences are also a common feature of the newly characterized LPXTG proteins. The SX dipeptide repeats of SasA resemble the SD dipeptide repeats of the Clf–Sdr family. The length of the SX repeats varies considerably from strain to strain, as do the SD repeats (McDevitt & Foster, 1995), and it seems reasonable to think that they have the same function, namely to project the ligand-binding region from the cell surface (Hartford et al., 1997). Despite the strain-to-strain length variation of the SD and SX repeats, the underlying frequency of recombination between the DNA repeats is actually quite low. The SD-repeat-encoding region of clfA of strain Newman is composed of an array of 18 bp DNA repeats which are saturated with base substitution mutations such that very few are identical (Shields et al., 1995). PCR analysis of the SD repeat region yielded a single PCR product indicative of a low frequency of recombination (McDevitt & Foster, 1995). Similar results were obtained by PCR across the SX-repeat-encoding region of SasA (data not shown).

In contrast, the B repeats of SasG are between 98 and 100% identical at the amino acid and nucleotide levels. Such a high level of DNA sequence identity is likely to result in a high frequency of unequal recombination and a detectable level of length variants in any population. This explains the observation that PCR of sasG from genomic DNA isolated from a bacterial population generated a number of different sized fragments varying by the length of one repeat. The SdrD protein also contains a B repeat region which varies in the number of repeats between different strains (Josefsson et al., 1998a). PCR amplification of the sdrD gene from strain 8325–4, which bears five B repeat domains, yielded only a single PCR product. The DNA encoding B repeats is 68–78% identical so frequency of unequal recombination would be lower than for sasG, explaining the failure to detect variants by PCR. Length variation of sasG is reminiscent of the Bca protein from Group B streptococci, where it is believed to be a mechanism for evading the immune response during infection, a shorter protein being selected in neonates when compared with the corresponding isogenic maternal strain (Madoff et al., 1996). Similar variants were detected when the B repeat region of pls was analysed by PCR but this was attributed to homology between one of the primers and the repeats (Savolainen et al., 2001). The Pls B repeats did not vary in length after 200 generations of growth in broth, suggesting that shorter variants, for example, did not have a growth advantage in vitro (Savolainen et al., 2001).

The success of S. aureus as a human pathogen is likely to involve many virulence determinants. However, it has been shown that certain individual virulence factors are linked to specific diseases. For example, the collagen-binding adhesin (Cna) is present in only 32% of carriage isolates but is found in 52% of isolates from invasive disease (Peacock et al., 2002). An earlier study suggested an association between cna and septic arthritis (Switalski et al., 1993). In this study, the sasG and sasH genes were found to be virulence-associated genes; genes encoding the surface proteins FnbpA, Cna and SdrE were previously found to be virulence-associated in the same panel of invasive strains, indicating the important role played by some of the surface-expressed proteins in disease pathogenesis (Peacock et al., 2002). These associations help prioritize future studies of LPXTG proteins and indicate that sasG and sasH warrant early, in-depth characterization.

To determine if the new collection of Sas proteins is expressed during growth in infected human patients, the N-terminal region of each was expressed and used in ELISA tests with IgG purified from sera of convalescent patients. Elevated titres occurred in five of nine Sas proteins compared to control IgG, suggesting that exposure to antigens had occurred during previous infection(s). Variability in patients’ IgG response to each protein was observed. ANOVA indicated that this variability in antibody response is dependent on the age of the patients, with older patients showing higher variability compared to younger groups. Another factor that may contribute to this variability is the presence or absence of the particular gene from the infecting strain. For example, sasG and sasH occur at low frequency, 53 and 68%, respectively, in the invasive isolate and only 4/33 and 7/33 patients had elevated titres to the SasG and SasH proteins, respectively. In contrast, sasA, sasl and sasf are present in 100, 96 and 97% of the strains, and a greater number of patients had elevated titres to the SasA, Sasl and Sasf (11/33, 10/33 and 18/33, respectively).

In the cases of sasK and sase, which are present in fewer than 44% of the clinical isolates tested, the low titres might be due to the infecting strains lacking these genes. Sasl was also shown to be expressed during infection because sera from convalescent patients reacted with Sasl peptides displayed on the surface of E. coli (Etz et al., 2002), while SdrD was identified by serological proteome analysis (Vytvytska et al., 2002).

It is now apparent that S. aureus can express up to 21 different LPXTG-anchored surface proteins. The domain architecture, repeat regions and common signal peptide motifs suggest that the evolution of LPXTG-anchored surface proteins may have arisen in a modular fashion. Some of these proteins were recognized by sera from infected patients and were also found to be associated with clinically invasive disease but their functions remain to be elucidated. One or more of these proteins could be suitable for development as vaccine candidates, alongside the current contenders CIFA (Josefsson et al., 2001) and Cna (Nilsson et al., 1998; Visai et al., 2000).

ACKNOWLEDGEMENTS

The Health Research Board of Ireland and the Northern Ireland Health and Personal Social Services Research and Development Office are acknowledged for the support of a North–South Co-operation Grant to M. P. and T. J. F. The Wellcome Trust is acknowledged for its support.
of N. P. J. D., R. M. and S. P., and the BBSRC for its support for the ViruloGenome project. The Italian Ministry of University and Education [MIUR, PRIN (2001)/2003] for the support to P. S. is also acknowledged. We thank the Wellcome Sanger Research Institute, the Wellcome Trust, the University of Oklahoma and the Institute for Genomic Research for making incomplete genome sequences publicly available.

REFERENCES


