This is a repository copy of *Periscope Proteins are variable length regulators of bacterial cell surface interactions*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/173977/

Version: Accepted Version

Article:
Whelan, Fiona, Lafita, Aleix, Gilburt, James Alexander Hinett et al. (10 more authors) (2021) Periscope Proteins are variable length regulators of bacterial cell surface interactions. Proceedings of the National Academy of Sciences of the United States of America. e2101349118. ISSN 1091-6490

https://doi.org/10.1073/pnas.2101349118

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

PNAS
www.pnas.org

Main Manuscript for

Periscope Proteins are variable length regulators of bacterial cell surface interactions

Fiona Whelan[a,1,2], Aleix Lafita[b,1], James Gilburt[a,1], Clément Dégut[a,1], Samuel C. Griffiths[a,1], Huw T. Jenkins[c], Alexander N. St John[d], Emanuele Paci[d], James W.B. Moir[a], Michael J. Plevin[a], Christoph G. Baumann[a], Alex Bateman[b,3] and Jennifer R. Potts[a,3,4]

[a]Department of Biology, The University of York, York YO10 5DD, UK; [b]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, CB10 1SD, UK; [c]Department of Chemistry, University of York, York YO10 5DD, UK; [d]Astbury Centre for Structural Molecular Biology, The University of Leeds, Leeds, LS2 9JT, UK.

[1]F.W, A.L., J.G., C.D., and S.C.G. contributed equally to this work

[2]Current address: Department of Molecular and Biomedical Science, The University of Adelaide, Adelaide, South Australia, 5005, Australia.

[3] To whom correspondence may be addressed. **Email:** agb@ebi.ac.uk or jennifer.potts@sydney.edu.au

[4]Current address: School of Life and Environmental Sciences, University of Sydney, Camperdown, NSW 2006, Australia

Author Contributions: J.R.P. proposed the study; F.W., A.L., J.G., S.C.G., C.G.B., E.P., A.N.StJ., H.T.J., A.B. and C.D. performed experiments and analyzed data; J.R.P., A.B., C.G.B., M.J.P., J.W.B.M. and E.P. guided experimental design and data analysis. All authors contributed to the interpretation of the data and to the writing of the manuscript.

Competing Interest Statement: The authors declare no competing interests.

Classification: Biological Sciences, Biophysics and Computational Biology

Keywords: Structural biology; bacterial surface proteins; regulation of cell surface interactions.

This PDF file includes:

Main Text
Figures 1 to 6

Abstract

Changes at the cell surface enable bacteria to survive in dynamic environments, such as diverse niches of the human host. Here, we reveal “Periscope Proteins” as a widespread mechanism of bacterial surface alteration mediated through protein length variation. Tandem arrays of highly similar folded domains can form an elongated rod-like structure; thus, variation in the number of domains determines how far an N-terminal host ligand binding domain projects from the cell surface. Supported by newly-available long-read genome sequencing data, we propose this new class could contain over 50 distinct proteins, including those implicated in host colonization and biofilm formation by human pathogens. In large multi-domain proteins, sequence divergence

between adjacent domains appears to reduce inter-domain misfolding. Periscope Proteins break this “rule”, suggesting their length variability plays an important role in regulating bacterial interactions with host surfaces, other bacteria and the immune system.

Significance Statement

The structure of single and tandem SHIRT domains from the streptococcal surface protein Sgo0707 were determined. In conjunction with biophysics and molecular dynamics simulations the results show that observed gene length variation would result in differential projection of the host ligand binding domain on the bacterial cell surface. Analysis of long-read DNA sequence data reveals many other repetitive bacterial surface proteins that appear to undergo gene length variation. We propose these variable-length “Periscope Proteins” represent an important mechanism of bacterial cell surface modification with potential roles in infection and immune evasion.

Main Text

Introduction

Bacteria encounter complex and dynamic environments, including within human hosts, and have thus evolved various mechanisms that enable a rapid response for survival within, and exploitation of, new conditions. In addition to classical control by regulation of gene expression, bacteria exploit mechanisms that give rise to random variation to facilitate adaptation (e.g. phase and antigenic variation (1)). In Gram-positive and Gram-negative human pathogens, DNA inversions (2, (3), homologous recombination (4), DNA methylation (1) and promoter sequence polymorphisms (5) govern changes in bacterial surface components including capsular polysaccharide and protein adhesins, which can impact on bacterial survival and virulence in the host (1, (6). Many of these mechanisms are very well studied and widespread across bacteria.

A less well-studied mechanism is length variation in bacterial surface proteins. Variability in the number of sequence repeats in the Rib domain (7)-containing proteins on the surface of Group B

streptococci has been linked to pathogenicity and immune evasion (8). The repetitive regions of the *Staphylococcus aureus* surface protein G (SasG) (9) and *Staphylococcus epidermidis* SasG homologue, Aap (10) also demonstrate sequence repeat number variability. In SasG this variability regulates ligand binding by other bacterial proteins *in vitro* (11) in a process that has been proposed to enable bacterial dissemination in the host. Variations in repeat number have also been noted in the biofilm forming proteins Esp from *Enterococcus faecalis* (12) and, more recently, CdrA from *Pseudomonas aeruiginosa* (13). High DNA sequence identity in the genes that encode these proteins is likely to facilitate intragenic recombination events that would lead to repeat number variation (14) and, in turn, to protein sequence repetition. However, such sequence repetition is usually highly disfavoured in large multi-domain proteins (15), so its existence in these bacterial surface proteins suggests that protein length variation provides an evolutionary benefit. SasG, Aap and Rib contain N-terminal host ligand binding domains and C-terminal wall attachment motifs, thus our recent demonstration that the repetitive regions of both SasG (16) and Rib (17) form unusual highly elongated rods suggests that host-colonisation domains will be projected differing distances from the bacterial surface.

Here we show that repeat number variation in predicted bacterial surface proteins is more widespread and we characterise a third rod-like repetitive region in the *S. gordonii* protein (Sgo_0707) formed by tandem array of novel 'SHIRT' domains. Thus, we propose a new, and growing, class of "Periscope Proteins", in which long, highly similar DNA repeats facilitate expression of surface protein stalks of variable length. This mechanism could enable changes in response to selection pressures and confer key advantages to the organism that include evasion of the host immune system (8) and regulation of surface interactions (11) involved in biofilm formation and host colonisation.

Results

Defining the structural repeats of Sgo_0707 from *Streptococcus gordonii*

Having revealed the unusual repetitive, rod-like characteristics of both SasG (16) and Rib (17) in our previous studies, we used bioinformatic approaches to search for other cell-wall attached bacterial proteins with similar domain architectures. A8AW49 (herein Sgo_0707) encoded by the gene *Sgo_0707* from *S. gordonii* (Fig. 1*A*) has a C-terminal wall attachment motif, homologues with repeat number variation, and a structurally defined two-domain N-terminus (N1-N2; residues 36–458, PDB: 4igb), that is proposed to be involved in collagen binding (18). As the repeats had no Pfam definition we called the putative domain "SHIRT" (Streptococcal High Identity Repeats in Tandem; Fig. 1*B*). *S. gordonii* is a member of the *S. sanguinis* group of viridans streptococci (19) and is a common colonizer of the oral cavity. It is a pioneer organism in the establishment of dental plaque (20) and also implicated in infective endocarditis (21).

Defining the structural, rather than sequence, repeat boundaries in repetitive bacterial proteins is challenging. For Sgo_0707 the T-REKS server (22) predicts a repeat frame of 460–543 and 13 repeats of 84–90 residues. A construct based on the second repeat (residues 544–627; ΔN-Sgo_R2) was folded (*SI Appendix*, Fig. S1*A*) and solved to 0.95 Å resolution using X-ray crystallography (*SI Appendix*, Fig. S1*A* inset and Table 1) utilising *ab initio* molecular replacement (MR) with ideal fragments (23). Based on the significant truncation of the N-terminal β-strand (*SI Appendix*, S1*A* inset), we hypothesised that shifting the frame of the repeat by seven residues towards the N-terminus of Sgo_0707 would complete the fold. Sgo_R3 (residues 621–705) and Sgo_R10 (residues 1211–1299) based on this new definition were thus expressed and purified. They were found to have significantly higher melting temperatures (T_m) than the N-terminally truncated Sgo_R2 (ΔN-Sgo_R2, 55.9°C; Sgo_R3, 75.7°C; Sgo_R10, 75.9°C; *SI Appendix*, Fig. S1*B*).

The structure of Sgo_R10 (Fig. 1*B*) was solved at 0.82 Å resolution using MR with the structure of ΔN-Sgo_R2 for phasing; the data collection and refinement statistics are summarised in Table 1.

The model confirms that SHIRT has a novel α/β fold organized around a single α-helix and two distinct β-sheets (Fig. 1*B*). Fig. 1*A* shows a schematic of Sgo_0707 based on the structural boundaries of the repeats; *SI Appendix*, S1*C* shows the high level of protein sequence identity (82–100%) between adjacent SHIRT domains. Comparison of Sgo_0707 genes from different bacterial strains, including the homologous protein fibA from *S. sanguinis*, shows a high variability in the number of SHIRT domain repeats forming the stalk, ranging from 3 to 14 copies (Fig. 1*C*). SHIRT domains are found in many other proteins, often in tandem array (*SI Appendix*, Fig. S2).

Tandemly-arrayed Sgo_0707 SHIRT domains form an extended rod-like structure

A tandem domain construct (Sgo_R3-4; residues 621—789) was crystallised and the structure solved via MR using the ΔN-Sgo_R2 model; data collection and refinement statistics are summarised in Table 1. The structure (Fig. 2*A*) reveals two complete domains with a very short (Pro-Ala-Pro) linker (*SI Appendix*, Fig. S1*C, D*); the structure of Sgo_R3-4 is ordered throughout (residues 623–789). Each domain adopts the SHIRT fold and the inter-domain interface is limited; this was confirmed by comparing the T_m of Sgo_R3 (75.7°C) and Sgo_R3-4 (76.6°C; *SI Appendix*, Fig. S3). The similarity of unfolding curves for single and double SHIRT constructs suggests that the two domains in the tandem construct unfold independently. Small angle X-ray scattering (SAXS) analysis substantiates the anisotropic head-to-tail domain arrangement in solution (*SI Appendix*, Fig. S4*A*). Notably, there is a significant twist between domains when viewed along the long axis of the molecule.

Molecular dynamics (MD) simulations of the Sgo_R3-4 construct show that individual domains are particularly stable (RMSD <1.5 Å for Cα atoms) over the length of the trajectory (0.8 μs); their individual length is conserved during the simulation and the length of Sgo_R3-4 fluctuates only moderately around 97 Å (Fig. 2*B*). The distributions of α, β, γ inter-domain (17) angles (*SI Appendix*, Fig. S5*A*) observed in the simulations of the Sgo_R3-4 construct (Fig. 2*C*) were used

to generate models of longer constructs (Fig. 2*D*). The radius of gyration (R_g) of the simulated constructs increases following the relation $R_g \propto N^{\nu}$, where ν is the Flory exponent and describes the increase in size of a polymer (protein) made of N monomers (amino acids). Such an exponent is ~0.6 for denatured proteins and ~0.4 for folded ones (24). Polymers formed of sequential SHIRT domains are highly extended; R_g scales with the number of domains (or equivalently, amino acids) with an exponent $\nu \sim 0.8$ (Fig. 2*E*), which is remarkable given the width of the distribution of the angles describing the mutual orientation of adjacent domains (Fig. 2*C*).

To assess the elongation of Sgo_0707 in solution, we collected SAXS data for constructs comprising two (Sgo_R3-4) and seven (Sgo_R2-8) tandemly-arrayed SHIRT domains (Fig. 3 and *SI Appendix*, Fig. S4). Both Sgo_R3-4 and Sgo_R2-8 are monomeric in solution, eluting as monodisperse peaks from size exclusion chromatography (SEC) columns (*SI Appendix*, Fig. S4, *A* and *B*, inset). Both the crystal structure of Sgo_R3-4 (Fig. 2*A*) and an elongated model for Sgo_R2-8 (*SI Appendix*, Fig. S4*B*) are consistent with the SAXS data measured in solution (model:data fits of χ^2=1.1 and χ^2=1.2, respectively; *SI Appendix*, Fig. S4, *A* and *B* and Materials and Methods). Analysis of the data for Sgo_R3-4 and Sgo_R2-8 results in equal Porod exponents (1.2, Fig. 3*A*), as well as similar radii of gyration of a cross-section (R_g^c) values (R3-4 = 6.4±0.0 Å; R2-8 = 7.0±0.0 Å; Fig. 3, *B* and *C*). Consistent with the models (Fig. 2*D*), the D_{max} determined using SAXS scales with the number of domains; Sgo_R3-4 exhibits a D_{max} of 107 Å and Sgo_R2-8 has a D_{max} of 371 Å (*SI Appendix*, Fig. S4, *A* and *B*). Therefore, whilst displaying a much larger intra-particle maximum dimension, the Sgo_R2-8 construct has a comparable shape and R_g^c to Sgo_R3-4.

In our previous study (17), we observed elongation in 2-domain Rib constructs with rotation of angle α, whilst angles β and γ were smaller. We conducted the same analysis for Sgo_0707

constructs, based on fitting to our experimental SAXS data (*SI Appendix*, Fig. S5). As with Rib, MD simulations of the 2-domain Sgo_R3-4 construct show that a range of α angles give good fits to the observed SAXS data, whilst β and γ are restricted to a narrow range, consistent with an elongated conformation (*SI Appendix*, Fig. S5*B*). For fitting of MD simulations of the 7-domain Sgo_R2-8 construct to SAXS data, we observed that longer end-to-end distances improved the quality of the fit (*SI Appendix*, Fig. S5*C*). Taken together, these data are consistent with multiple tandemly-arrayed SHIRT domains from Sgo_0707 behaving as an elongated, rod-like particle.

To further assess the elongation of multi-domain SHIRT constructs, we used a high-resolution single-molecule technique (SHRImP (16)) to measure the intramolecular distance between two Alexa Fluor 488 (AF488) dyes covalently attached to cysteine residues engineered at specific sites in Sgo_R2-8 (S666C and S1086C). If the extended inter-domain topology observed in the two-domain construct is maintained, a distance of 24.1 nm between the mean dye positions (*SI Appendix*, Fig. S6*A*) is predicted by using the distance between the two cysteine residues (Fig. 3*D* inset) and simulating the increased volume accessible to each dye due to the chemical linker (25). AF488-labelled Sgo_R2-8$^{S666C/S1086C}$ was imaged using total internal reflection fluorescence microscopy (TIRFM) using two different poly-D-lysine concentrations to coat the slides. Both surface treatments effectively immobilised AF488-labelled Sgo_R2-8$^{S666C/S1086C}$ and each SHRImP-TIRFM histogram had a single peak consistent with monodispersity. The measured inter-dye distances were 25.5 nm and 27.8 nm (Fig. 3*D*) with 2 µg/mL and 20 µg/mL poly-D-lysine concentration, respectively, and increased to 29.8 nm and 33.0 nm, (*SI Appendix*, Fig. S6*B*), respectively, when measurements were performed in the presence of 100-fold molar excess of unlabelled Sgo_R2-8 (termed 'blocking' protein). This implies the rod-like conformation of Sgo_R2-8 protein is malleable, adopting, compared to solution, a slightly more elongated state when on a surface which is accentuated (Δx = 4-5 nm) at high protein concentrations that presumably favour lateral protein-protein interactions.

Large-scale identification of potential Periscope Proteins in bacterial genomes

We call SasG, Rib and Sgo_0707 'Periscope Proteins' due to their having variable length stalk-like regions that define the distance that the N-terminal functional domain projects from the bacterial surface. We conducted an unbiased identification of other potential Periscope Proteins from the NCTC3000 bacterial genomes: an ongoing project lead by the Wellcome Sanger Institute (UK) that provides high quality annotated genome assemblies for 3,000 bacterial strains from Public Health England's National Collection of Type Cultures (NCTC) using the long-read PacBio technology (https://www.sanger.ac.uk/resources/downloads/bacteria/nctc/). The use of long-reads spanning whole genes overcomes the challenges in the assembly of repetitive genes in bacterial genomes (26), providing reliable numbers of repeats in the assembled genes and associated proteins, and allowing the identification of length variability in Periscope Proteins. We identified 1,576 proteins containing long and highly identical repeats (i.e. similar to those forming stalks in Periscope Proteins), out of a total of 2.5 million proteins in the NCTC3000 dataset, from different species including both Gram-positive and Gram-negative bacteria. Clustering of the repeating sequences reveals most of them are classified into existing Pfam families (Fig. 4, Supplementary Dataset S1) that correspond to globular domains, as in known Periscope Proteins, and from a wide range of secondary structure and fold types.

Most of the domain families we find correspond to Immunoglobulin-like beta sandwich folds in the E-set clan (Pfam: CL0159), but we also find domain families with beta grasp folds in the Ubiquitin clan (Pfam: CL0072) and three-helix bundles in the bacterial immunoglobulin/albumin-binding clan (Pfam:CL0598). As expected, the distance between the N- and C-termini of domain repeats is large and their termini are oriented close to 180 degrees (3.14 radians; *SI Appendix* Table S3), enabling linear arrangements of tandem domain repeats. In addition, most inter-domain linkers are short (smaller than 5 residues) and some of them are rich in prolines, conferring additional

rigidity to the linear domain arrangements. Some exceptions with short inter-termini distance and small angles, such as LysM and SH3_3 domains, contain longer and more flexible linkers.

We further clustered full-length repetitive proteins and identified a total of 180 unique groups, 84 of which exhibit repeat number variation (Supplementary Data 1). Manual inspection confirmed that 56 of these groups are examples of Periscope Proteins, 30 of which could be assigned to recognisable gene names. We retrieved known Periscope Proteins, such as SasG, Rib and Sgo_0707 (clusters 4, 25 and 26 respectively), as well as novel ones, for example CdrA from *Pseudomonas aeruginosa* (27) which has been reported recently to exhibit variability in the size of the repeat region (13), CshA from *S. gordonii (28)*, and SAP077A_019 (UniProt D2JAN8) from *S. aureus* (Fig. 5). The majority of confirmed Periscope Proteins (40/56) have at least one Gene Ontology (GO) location term assigned, all of them associating to the cell-wall or bacterial membrane.

We further find that the number of repeats per gene can be large (>40 in one case (29); Fig. 6), spanning thousands of amino acids, and that variation in the number of stalk repeats can be extreme, increasing the length of the protein by an order of magnitude in some cases. Importantly, we found the most extreme repeat numbers and length variation in proteins with the highest repeat identity (Fig. 6). Furthermore, as shown in Fig. 5, the number of repeats changes drastically even between similar bacterial strains, suggesting a rapid evolutionary rate of repeat number change in Periscope Proteins.

Discussion

The work presented in this study identifies a new class of bacterial surface proteins; through length variation these proteins are likely to modulate surface interactions. Sgo0707 is the third example of this new class for which we have characterized both isolated domains and tandem arrays. SasG, the first example (16), is comprised of arrays of E and G5 domains. E domains are

unstable in the absence of G5 domains. The interdomain interfaces (E-G5 and G5-E) make very significant contributions to the overall stability of the tandem array of E-G5 domains. G5 domains also have an unusual flat structure rather than a “typical” hydrophobic core. The second example, Rib, is different. We reported the first 3D structure of a Rib domain (17), which appears to be an interesting example of domain atrophy from the common Ig-fold. In Rib, there is no evidence for a significant interdomain interface. SHIRT (presented here) is different again. We report the first 3D structure of a SHIRT domain which, whilst also containing β-sheets, has a novel topology with no obvious relationship to the Ig-fold. Like Rib, there is no evidence for significant interdomain interfaces in tandem arrays of SHIRT domains. Rib and Sgo_0707 have relatively short interdomain linkers between Rib and SHIRT domains, respectively, and/or linkers that contain proline residues; both features are likely to contribute to the observed elongation of the tandemly-arrayed domains. In addition, due to requirements of forming an elongated array, we predict that, as observed for SasG (PDB 3TIP; (30)), Rib (PDB 6S5X; (17)) and Sgo_0707 (Fig.1), Periscope Proteins will often contain domain folds that locate the N- and C-termini of the protein sequence at either end of the fold (*SI Appendix* Table 3). Thus, in Periscope Proteins, we find that different types of domains, when arrayed in tandem, can form an elongated rod that typically serves to project a functional domain distal to the cell surface.

To classify this diverse family of proteins with a memorable analogy, we call them ‘Periscope Proteins’. We propose that high sequence similarity at the DNA level enables intragenic recombination events ((14); and thus loss or gain of repeats) with selection pressures resulting in enrichment of bacteria with shorter or longer proteins. Periscope Proteins include proteins implicated in biofilm formation in both Gram-positive and Gram-negative bacteria, for example SAP077A_019 from *S. aureus* and related proteins in *S. epidermidis*, *S. xylosus*, *S. capitis, S. simiae* and *S. warneri*, the *Enterococcus faecalis* protein Esp (12) and related proteins, and CdrA from *Pseudomonas aeruginosa* (13*).* We expect the total number of Periscope Proteins across bacterial species to be much larger and more widespread than reported in this study. Our knowledge of Periscope Proteins will increase as more long-read sequence data for a diverse set

of strains not included in the NCTC3000 collection become available. We suggest that length variation is the main function of the tandemly-arrayed regions of Periscope Proteins; however, as some of the domains found in such arrays (such as LysM domains) have had a ligand function assigned they could play an additional functional role in some contexts.

Large multi-domain proteins usually have <50% sequence identity between adjacent domains (15) which has been proposed as an evolutionary strategy to minimise inter-domain misfolding (31, (32, (33). The existence of Periscope Proteins, which contain tandemly-arrayed, structured domains with high sequence similarity, confounds this observation. Considering this potential "misfolding problem", the existence of Periscope Proteins suggests that length variation (driven by the requirement for high identity at the DNA level, and thus high protein similarity) is functionally important and confers a significant advantage to the organism. Simply expressing a Periscope Protein (an extreme form of length variation) results in changes in ligand binding by other bacterial surface proteins through spatial competition. For example, expression of Pls (34) and SasG significantly inhibits binding of *S. aureus* to the human plasma proteins fibronectin and fibrinogen, respectively; the latter resulting in altered bacterial colony morphology (35). In both cases it is suggested that the elongated Periscope Protein is blocking interactions of the host protein with bacterial proteins closer to the cell surface. There is also more direct evidence for the role of length *variation* in Periscope Proteins under selection pressure. For example, in mice immunised with anti-serum raised against the nine-repeat alpha C, GBS expressing alpha C with one Rib repeat were 100-fold more pathogenic than GBS expressing alpha C with nine Rib repeats (36). It was proposed that the shorter protein is not recognised because it is less exposed. When SasG was expressed on *S. aureus* with differing numbers of repeats, only the longer variants blocked binding of other bacterial surface proteins to their target (11). These authors noted that the ability of a bacterium to detach from a ligand could be important for dissemination in the host. Finally, deletion of CdrA had the most deleterious effect on biofilm formation for *P. aeruginosa* strains expressing the longest CdrA variants (13). Rib repeat number variation within a single strain can be observed within 24 hours following inoculation of a mouse

with group B streptococci (37), suggesting the potential for dynamic regulation by Periscope Proteins on a physiologically relevant timescale. In summary, surface variation through the Periscope Protein class appears to be a highly distinctive way to alter a variety of interactions linked to colonisation and infection.

Materials and Methods

Cloning.

The *E. coli* codon-optimised coding sequence for *S. gordonii* strain Challis (substrain DL1) Sgo_0707 residues 544–795 (UniProt: A8AW49) was synthesized (Genewiz; *SI Appendix,* Table S2) and the truncated single repeat ΔN-Sgo_R2 (amino acids 544–627), single repeats Sgo_R3 (aa 621–705) and Sgo_R10 (aa 1211–1299) (synthesized by Eurofins Genomics), and tandem repeat Sgo_R3-4 (aa 621–789) were cloned downstream of a hexahistidine tag and 3C protease specific linker by the In-fusion method (Clontech; primer sequences listed in *SI Appendix* Table S1). DNA coding for the 7-repeat construct (Sgo_R2-8) comprising the 2nd to 8th SHIRT repeats (aa 537–1125) was synthesised (Eurofins Genomics), and inserted by homologous recombination into a modified pBAD vector (pBADcLIC2005) generating a coding sequence comprising GGGFA-Sgo_R2-8-His_{10}. Site-directed mutagenesis was used to introduce cysteine mutations in the Sgo_R2-8 construct for fluorescent dye modification at positions S666 and S1086 (Sgo_R2-$8^{S666C/S1086C}$; mutagenesis primer sequences listed in *SI Appendix,* Table S1).

Protein Expression and Purification.

For ΔN-Sgo_R2, Sgo_R3 and Sgo_R3-4, expression was induced in *E. coli* BL21 (DE3) cells in log phase growth with the addition of 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and subsequently incubated with shaking at 20°C for 20 h. Cells were harvested by centrifugation, resuspended in 20 mM Tris.HCl, 150 mM NaCl, 20 mM imidazole, pH 7.5, and lysed by sonication. Soluble protein was purified by standard nickel affinity chromatography methods, 3C

protease was added at a ratio of 1:100 (w/w) to remove the hexahistidine tag, and proteolysed material isolated by a second nickel affinity chromatography purification. Sgo_R10 and Sgo_R3-4 were further purified by preparative SEC using a Superdex 75 16/600 column (GE Healthcare) equilibrated in 20 mM Tris.HCl, 150 mM NaCl, pH 7.5 (Sgo_R10) and pH 8 (Sgo_R3-4). Sgo_R2-8 constructs were transformed into Rosetta-2 *E. coli* (DE3) cells and grown in LB-Miller media containing 100 µg/mL ampicillin at 37°C to an OD_{600} of 0.5-0.6. Gene expression was induced by addition of 0.1% (w/v) L-arabinose, followed by incubation at 20°C. Soluble protein was purified as described for Sgo_R10. Protein containing cysteine mutations were purified by affinity chromatography and SEC with the addition of 5 mM β-mercaptoethanol at all steps. Protein samples prepared for SEC-SAXS were dialysed into 20 mM Tris.HCl, 150 mM NaCl, 1 mM EDTA, pH 7.5.

Protein crystallization.

Proteins were concentrated (ΔN-Sgo_R2 to 48 mg/mL; Sgo_R10 to 30 mg/mL and Sgo_R3-4 to 35.2 mg/mL) by centrifugal filtration with 3 kDa molecular weight cutoff (MWCO) PES (Vivaspin). ΔN-Sgo_R2 crystallized by sitting drop vapour diffusion within 9 months in conditions comprising 0.1 M HEPES pH 7, 2.4 M ammonium sulphate at 291 K. This crystal was passed through 4 M sodium malonate prior to flash cooling. Sgo_R10 crystallized in 5 weeks at 277 K in conditions comprising 2.2 M ammonium sulphate and 150 mM potassium thiocyanate. The crystal was harvested under mineral oil and flash cooled in liquid nitrogen prior to data collection. Sgo_R3-4 crystallized in 4 days in conditions comprising 65% (v/v) 2-methyl-2,4-pentanediol and 0.1 M Tris.HCl pH 8 at 277 K and flash cooled in liquid nitrogen prior to data collection.

Structure determination.

X-ray data were collected at 100 K on the I03 beamline at Diamond Light Source (Didcot, UK). using a Pilatus 3 6M detector. Data were indexed and integrated by XDS (38) and scaled and

merged by Aimless (39). The structure of ΔN-Sgo_R2 was solved by MR with 2 antiparallel 5 residue ideal β-strands using Fragon (23) and the model built automatically with ARP/wARP (40). Sgo_R3-4 and Sgo_R10 were phased by molecular replacement with Phaser (41) using search model ΔN-Sgo_R2. Models were manually-built using Coot (42) and refined to completion with REFMAC5 (43) for Sgo_R3-4 and PHENIX (44) for Sgo_R10 (Table 1). The coordinates and structure factors have been deposited in the protein data bank with accession codes (ΔN-Sgo_R2, 7AVJ; Sgo_R10, 7AVK ; Sgo_R3-4, 7AVH). The structures were aligned by secondary structure matching with Superpose (45) and cartoons rendered with CCP4mg (46) with secondary structure defined by the database of protein secondary structure assignment (DSSP) (47).

Determination of T_m.

Differential scanning fluorimetry (DSF) was peformed using a Nanotemper Prometheus NT.48 instrument. Protein concentrations were 1 mM and solution conditions were 20 mM Tris.HCl pH 7.5, 150 mM NaCl.

Small Angle X-ray Scattering (SAXS).

SAXS experiments were performed at beamline B21, Diamond Light Source (Didcot, UK) over a momentum transfer range (q) of 0.01 $Å^{-1} < q < 0.4$ $Å^{-1}$. Scattering intensity (I vs q, where $q = 4\pi\sin\theta/\lambda$ and 2θ is the scattering angle) was collected using a Pilatus 2M detector, with a beam-to-detector distance of 4014 mm and an incident beam energy of 12.4 keV. Sgo_R3-4 and Sgo_R2-8$^{S666C/S1086C}$ were injected on an inline Shodex KW-203.5 column equilibrated in 20 mM Tris-HCl, 150 mM NaCl, 3 mM KNO_3, 5mM β-mercaptoethanol pH 7.5, each at 7.5 mg/mL, and data processing and reduction was performed with Chromixs (48). R_g^c (Ln[$I(q)$ vs q^2], and Distance Distribution ($P(r)$) plots were calculated with Primus (49). The range of useful scattering angles was assessed using Shanum (50). SWISS-MODEL (51) was used for the generation of a structural model of Sgo_R2-8, using the Sgo_R3-4 crystal structure to generate iterative tandem

domains for R3-5, R5-6 and R7-8. Sgo_R2 of the Sgo_R2-8 model was generated based on Sgo_R4 from the Sgo_R3-4 structure. For validation of the Sgo_R3-4 crystal structure, tag residues unresolved in the crystal structure were added using Modeller and all-atom ensembles generated using Allosmod (51). In each case, 50 independent pools of 100 models were created, and calculation and fitting of theoretical scattering curves to experimental data was performed using FoXS (52). This process was automated using Allosmod-FoXS (53). Plots were generated using OriginPro v9.5.5.409 (OriginLab), as were the gradients of the linear regions of double logarithmic plots for the calculation of Porod exponents (54).

Production of fluorescently-labelled protein.

As described above, cysteine residues were engineered into the 3rd and 8th repeats of the Sgo_R2-8 construct at positions (S666, S1086), where fluorescence quenching by nearby amino acid residues is minimised. Sgo_R2-8$^{S666C/S1086C}$ (27 µM) was dialysed into 150 mM NaCl, 20 mM Tris.HCl, 1 mM EDTA, pH 7.5, followed by dialysis into 150 mM NaCl, 20 mM Tris.HCl, 1.35 mM tris(2-carboxyethyl)phosphine (TCEP), pH 7.5. The protein was then reacted at 20˚C in low light conditions with a 20x molar excess of Alexa Fluor 488 C_5 maleimide (ThermoFisher, 540 µM), which was added step-wise over a period of 1.5 h using a 10 mM stock in anhydrous DMSO. The labelling reaction was quenched by adding dithiothreitol (DTT) at 10x molar excess to the maleimide. The protein solution was dialysed into 150 mM NaCl, 20 mM Tris.HCl, 1 mM DTT, pH 7.5 prior to purification by SEC on a S200 30/10 column (Amersham) equilibrated in the same buffer to remove remaining free dye. Purified proteins were stored at -80°C. The labelling efficiency (~2 fluorophores/protein) was estimated from the spectrophotometrically determined concentrations of fluorophore ($\varepsilon_{495\ nm}$ = 72000 M^{-1} cm^{-1}) and protein ($\varepsilon_{280\ nm}$ = 99475 M^{-1} cm^{-1}) after correction for absorption at 280 nm by the fluorophore.

Sample preparation for SHRImP-TIRF microscopy.

A 100 mM Trolox solution was freshly prepared by solubilising 25 mg of Trolox powder (Fluka) in 50 µL methanol, followed by dilution with 850 µL of 0.31 M NaOH solution. All stock buffer solutions were passed through a 0.22 µm pore filter. Adsorption buffer contained 10 mM HEPES, 10 mM NaCl, pH 7.0, 1 mM Trolox, 0.02% (w/v) 5 µm diameter silica beads (Bangs Labs), 8 pM Alexa Fluor 488 (AF488)-labelled Sgo_R2-8$^{S666C/S1086C}$ protein and, in the 'blocked' samples only, 800 pM unlabelled Sgo_R2-8 protein. Imaging buffer contained 10 mM HEPES, 10 mM NaCl, pH 7.0, 1 mM Trolox. Poly-D-lysine-coated quartz slides were prepared as described previously (16). 1 µM AF488-Sgo_R2-8$^{S666C/S1086C}$ and 200 nM Sgo_R2-8 stock solutions were thawed from -80°C storage and diluted with 20 mM Tris.HCl, 150 mM NaCl (pH 7.5) buffer to the desired concentration before addition to the adsorption buffer at a 25x or 50x dilution. In 'blocked' samples, the molar concentration of unlabelled Sgo_R2-8 in the adsorption buffer was maintained at 100x the molar concentration of AF488-Sgo_R2-8$^{S666C/S1086C}$.

In low light conditions, 50 µL of adsorption buffer was distributed along the centre-line of a 2 µg/mL or 20 µg/mL poly-D-lysine-coated quartz slide and then covered with a clean coverslip (No. 1, 22 mm x 64 mm, Menzel-Gläser). A flow chamber was created by sealing the two opposite, short sides of the coverslip with nail varnish. A small amount of imaging buffer was added along the unsealed sides of the flow chamber to prevent it drying out. After 10 min of incubation at room temperature (20-22˚C), ~500 µL of imaging buffer was flowed through the chamber created by the slide-silica bead-coverslip sandwich to wash away unbound protein, and the chamber was sealed with nail varnish.

SHRImP-TIRF microscopy.

Fluorescence excitation and detection of AF488 dye emission was achieved using a custom, prism-coupled TIRF microscope as described previously (16) with the following modification and increased optical magnification. Quantum dots were not routinely added to the adsorption buffer

as an image-focusing aid. Video data (100 frames) were collected using an Evolve 512 (Photometrics) electron-multiplying CCD camera (500 ms exposure) and the pixel size was equivalent to 96 nm in the magnified image.

Detection and localisation of single fluorophores.

Fluorescent spot detection and calculation of inter-AF488 dye distances for each spot that photobleached in two-steps were performed as previously described (16). In addition, an eccentricity ratio was calculated for each spot (using the intensity profile in the *x* and *y* directions) to remove events that included weakly surface-adsorbed AF488-labelled protein or partially photobleached clusters of AF488-labelled protein. The intensity profile was calculated for a central region (2 x 10 pixels^2) along the *x* and *y* axes of each spot image. This image was the sum of the first 5 video frames for each 10x10 pixel^2 image stack (100 frames). The *x* and *y* intensity profiles were fit with a one-dimensional Gaussian function in MATLAB (MathWorks, Cambridge, UK) and a ratio of the widths for the Gaussian fits (= $\sigma_{y\text{-axis}} / \sigma_{x\text{-axis}}$) was used to obtain the spot eccentricity. Fluorescent spots with an eccentricity ratio between 0.9 and 1.1 were retained (this filter removed 15–30% of the spots in an experiment). Bin size for the inter-AF488 dye distance histograms was calculated using the Freedman-Diaconis rule (55), and a single Gaussian distribution was fit to each histogram in KaleidaGraph (Synergy Software).

Molecular dynamics simulations.

Molecular dynamics simulations were performed starting from the X-ray crystal structure model of Sgo_R3-4. Simulations have been performed using the CHARMM36m (56) force field and NAMD (57). The protein was energy minimized and solvated in a periodic rectangular box 123 Å x 46 Å x 40 Å needed to guarantee a layer of at least 12 Å solvent around the elongated protein. After a 1 ns equilibration, the systems were simulated at 303K for 0.8 μs. Simulations were performed in the isothermal-isobaric ensemble, where the temperature was kept constant on average through

a Langevin thermostat and the pressure was set to 1 atm through an isotropic Langevin piston manostat. For each saved frame of the simulation (one every 2 ns) the positions of the Ca atoms of Sgo_R4 were least square superposed to those of Sgo_R3, which implies a translation and a rotation around the each of the three principal axes of Sgo_R3 (hence applying the reverse transformation to the coordinates of Sgo_R4 the original conformation of that frame is recovered). Constructs with N domains were obtained by taking the original X-ray crystal structure, duplicating the coordinates of Sgo_R4 and applying to them the reverse transformation (using translation and rotations from random frames of the trajectory) N times.

Periscope protein identification in NCTC3000 collection.

A total of 2,579,577 proteins were extracted from 734 annotated bacterial genomes downloaded from the NCTC3000 project website (October 2019). Tandem repeats longer than 50 residues and with at least 80% sequence identity were detected with the T-REKS tool. Repeat sequences were clustered using BLASTp with a bit score threshold of 30 and extracting connected components of the resulting sequence similarity network (SSN). Similarly, proteins containing the long repeats were clustered using BLASTp with a bit score threshold of 100 and additional minimum sequence identity of 90% and coverage of 50% thresholds, by extracting the connected components of the SSN. Proteins were mapped to UniProt (58) identifiers (version 2020_04), with their associated Gene Ontology (GO) terms, and to Pfam (59) families (version 32.0) using PHMMER. Several new Pfam families were created from repeat sequences during the course of this study, including SHIRT (PF18655), YDG (PF18657), MBG_2 (PF18676), SSSPR-51 (PF18877), CshA_repeat (PF19076), and Big_13 (PF19077), among others. Phylogenetic trees of bacterial genomes were generated by first selecting all pairs of homologous genes for each genome pair using BLASTn, then computing a genomic sequence identity matrix and finally creating a dendrogram of strains by hierarchical clustering.

Analysis of inter-domain linkers and domain termini orientation

A subset of eleven proteins with diverse domain repeats were selected from the Periscope proteins table. Inter-domain linkers were extracted from the protein sequences according to the Pfam domain boundaries for each family, with minor adjustments to cover the structural boundaries for incomplete Pfam models (missing terminal residues). Structures of the domain repeats, or close homologs from the same family or families within the same Pfam clan, were selected for each protein. The distance between the terminal residues and their relative orientation were calculated using the TADOSS software (60). The distance is measured as the Euclidean distance between the C-alpha atoms of the N-terminal and C-terminal residues, while the orientation is measured as the angle between the vectors formed by the four N- and four C-terminal residues (C-alpha atoms).

Data Availability.

Atomic models of ΔN-Sgo_R2, Sgo_R10 and Sgo_R3-4 have been deposited in the Protein Data Bank with PDB IDs 7AVJ, 7AVK and 7AVH, respectively. All other study data are included in the main text and SI Appendix.

Acknowledgments

We acknowledge Johan Turkenburg and Sam Hart for assistance with crystal testing and data collection, Judith Hawkhead for protein expression and purification, and Andrew Leech, Rachael Cooper and Michael Hodgkinson for T_m measurements and figure preparation. The authors also thank the Diamond Light Source for access to beamline I03 (proposal number mx-9948) that contributed to the results presented here. This work used the Bioscience Technology Facility at the University of York. JG was funded by an MRC Discovery Award (MC_PC_15073). AB and AL are funded by the European Molecular Biology Laboratory. JRP, FW and SCG were funded by the British Heart Foundation (FS/12/36/29588 and PG/17/19/32862).

References

1. M. van der Woude, B. Braaten, D. Low. *Epigenetic phase variation of the pap operon in Escherichia coli*, *Trends Microbiol.* **4**, 5-9 (1996).
2. J. Li *et al. Epigenetic Switch Driven by DNA Inversions Dictates Phase Variation in Streptococcus pneumoniae*, *PLOS Pathogens* **12**, e1005762 (2016).
3. J. M. Abraham, C. S. Freitag, J. R. Clements, B. I. Eisenstein. *An invertible element of DNA controls phase variation of type 1 fimbriae of Escherichia coli*, *Proc. Natl. Acad. Sci. USA* **82**, 5724-5727 (1985).
4. C. Vink, G. Rudenko, H. S. Seifert. *Microbial antigenic variation mediated by homologous DNA recombination*, *FEMS Microbiol. Rev.* **36**, 917-948 (2012).
5. Z. Wen, Y. Liu, F. Qu, J.-R. Zhang. *Allelic Variation of the Capsule Promoter Diversifies Encapsulation and Virulence In Streptococcus pneumoniae*, *Sci. Rep.* **6**, 30176 (2016).
6. J. O. Kim, J. N. Weiser. *Association of intrastrain phase variation in quantity of capsular polysaccharide and teichoic acid with the virulence of Streptococcus pneumoniae*, *J. Infect. Dis.* **177**, 368-377 (1998).
7. M. Stålhammar-Carlemalm, L. Stenberg, G. Lindahl. *Protein rib: a novel group B streptococcal cell surface protein that confers protective immunity and is expressed by most strains causing invasive infections*, *J. Exp. Med.* **177**, 1593-1603 (1993).
8. C. Gravekamp, D. S. Horensky, J. L. Michel, L. C. Madoff. *Variation in repeat number within the alpha C protein of group B streptococci alters antigenicity and protective epitopes*, *Infect. Immun.* **64**, 3576-3583 (1996).
9. F. M. Roche *et al. Characterization of novel LPXTG-containing proteins of Staphylococcus aureus identified from genome sequences*, *Microbiology* **149**, 643-654 (2003).
10. H. Rohde *et al. Polysaccharide intercellular adhesin or protein factors in biofilm accumulation of Staphylococcus epidermidis and Staphylococcus aureus isolated from prosthetic hip and knee joint infections*, *Biomaterials* **28**, 1711-1720 (2007).
11. R. M. Corrigan, D. Rigby, P. Handley, T. J. Foster. *The role of Staphylococcus aureus surface protein SasG in adherence and biofilm formation*, *Microbiology* **153**, 2435-2446 (2007).
12. A. Toledo-Arana *et al. The Enterococcal Surface Protein, Esp, Is Involved in Enterococcus faecalis Biofilm Formation*, *Appl. Environ. Microbiol.* **67**, 4538 (2001).
13. C. Reichhardt *et al. The Versatile Pseudomonas aeruginosa Biofilm Matrix Protein CdrA Promotes Aggregation through Different Extracellular Exopolysaccharide Interactions*, *J. Bacteriol.* **202**, e00216-00220 (2020).
14. C. S. Lachenauer, R. Creti, J. L. Michel, L. C. Madoff. *Mosaicism in the alpha-like protein genes of group B streptococci*, *Proc. Natl. Acad. Sci. USA* **97**, 9630-9635 (2000).
15. C. F. Wright, S. A. Teichmann, J. Clarke, C. M. Dobson. *The importance of sequence diversity in the aggregation and evolution of proteins*, *Nature* **438**, 878-881 (2005).
16. D. T. Gruszka *et al. Cooperative folding of intrinsically disordered domains drives assembly of a strong elongated protein*, *Nat. Commun.* **6**, 7271 (2015).
17. F. Whelan *et al. Defining the remarkable structural malleability of a bacterial surface protein Rib domain implicated in infection*, *Proc. Natl Acad. Sci. USA* **116**, 26540-26548 (2019).
18. A. Nylander *et al. Structural and functional analysis of the N-terminal domain of the Streptococcus gordonii adhesin Sgo0707*, *PLoS One* **8**, e63768 (2013).

19. C. D. Doern, C.-A. D. Burnham. *It's Not Easy Being Green: the Viridans Group Streptococci, with a Focus on Pediatric Clinical Manifestations*, *J. Clin. Microbiol.* **48**, 3829 (2010).
20. N. S. Jakubovics, S. A. Yassin, A. H. Rickard. *Community Interactions of Oral Streptococci*, *Adv. Appl. Microbiol.* **87**, 43-110 (2014).
21. I. M. Tleyjeh *et al. Temporal trends in infective endocarditis: a population-based study in Olmsted County, Minnesota*, *JAMA* **293**, 3022-3028 (2005).
22. J. Jorda, A. V. Kajava. *T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm*, *Bioinformatics* **25**, 2632-2638 (2009).
23. H. T. Jenkins. *Fragon: rapid high-resolution structure determination from ideal protein fragments*, *Acta Crystallogr. D* **74**, 205-214 (2018).
24. H. Hofmann *et al. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy*, *Proc. Natl Acad. Sci. USA* **109**, 16155 (2012).
25. S. Kalinin *et al. A toolkit and benchmark study for FRET-restrained high-precision structural modeling*, *Nat. Methods* **9**, 1218-1225 (2012).
26. O. K. Tørresen *et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases*, *Nucleic Acids Res.* **47**, 10994-11006 (2019).
27. B. R. Borlee *et al. Pseudomonas aeruginosa uses a cyclic-di-GMP-regulated adhesin to reinforce the biofilm extracellular matrix*, *Mol. Microbiol.* **75**, 827-842 (2010).
28. C. R. Back *et al. The streptococcal multidomain fibrillar adhesin CshA has an elongated polymeric architecture*, *J. Biol. Chem.* (2020).
29. J. Lebeurre *et al. Comparative Genome Analysis of Staphylococcus lugdunensis Shows Clonal Complex-Dependent Diversity of the Putative Virulence Factor, ess/Type VII Locus*, *Front. Microbiol.* **10**, 2479 (2019).
30. D. T. Gruszka *et al. Staphylococcal biofilm-forming protein has a contiguous rod-like structure*, *Proc Natl Acad Sci U S A.* **109** (2012).
31. J. H. Han, S. Batey, A. A. Nickson, S. A. Teichmann, J. Clarke. *The folding and evolution of multidomain proteins*, *Nat Rev: Mol Cell Biol* **8**, 319-330 (2007).
32. A. Borgia *et al. Transient misfolding dominates multidomain protein folding*, *Nat. Commun.* **6**, 8861 (2015).
33. M. B. Borgia *et al. Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins*, *Nature* **474**, 662-665 (2011).
34. K. Savolainen *et al. Expression of pls, a Gene Closely Associated with the mecA Gene of Methicillin-Resistant Staphylococcus aureus, Prevents Bacterial Adhesion In Vitro*, *Infect. Immun.* **69**, 3013 (2001).
35. H. A. Crosby *et al. The Staphylococcus aureus Global Regulator MgrA Modulates Clumping and Virulence by Controlling Surface Protein Expression*, *PLOS Pathogens* **12**, e1005604 (2016).
36. C. Gravekamp, B. Rosner, L. C. Madoff. *Deletion of Repeats in the Alpha C Protein Enhances the Pathogenicity of Group B Streptococci in Immune Mice*, *Infect. Immun.* **66**, 4347-4354 (1998).
37. L. C. Madoff, J. L. Michel, E. W. Gong, D. E. Kling, D. L. Kasper. *Group B streptococci escape host immunity by deletion of tandem repeat elements of the alpha C protein*, *Proc. Natl. Acad. Sci. USA* **93**, 4131-4136 (1996).

38. W. Kabsch. *XDS*, *Acta Crystallogr. D* **66**, 125-132 (2010).
39. P. R. Evans, G. N. Murshudov. *How good are my data and what is the resolution?*, *Acta Crystallogr. D* **69**, 1204-1214 (2013).
40. G. Langer, S. X. Cohen, V. S. Lamzin, A. Perrakis. *Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7*, *Nat. Protoc.* **3**, 1171-1179 (2008).
41. A. J. McCoy *et al. Phaser crystallographic software*, *J. Appl. Crystallogr.* **40**, 658-674 (2007).
42. P. Emsley, K. Cowtan. *Coot: model-building tools for molecular graphics*, *Acta Crystallogr. D* **60**, 2126-2132 (2004).
43. M. D. Winn *et al. Overview of the CCP4 suite and current developments*, *Acta Crystallogr. D* **67**, 235-242 (2011).
44. D. Liebschner *et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix*, *Acta Crystallogr. D* **75**, 861-877 (2019).
45. R. Maiti, G. H. Van Domselaar, H. Zhang, D. S. Wishart. *SuperPose: a simple server for sophisticated structural superposition*, *Nucleic Acids Res.* **32**, W590-W594 (2004).
46. S. McNicholas, E. Potterton, K. S. Wilson, M. E. M. Noble. *Presenting your structures: the CCP4mg molecular-graphics software*, *Acta Crystallogr. D* **67**, 386-394 (2011).
47. W. Kabsch, C. Sander. *Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features*, *Biopolymers* **22**, 2577-2637 (1983).
48. A. Panjkovich, D. I. Svergun. *CHROMIXS: automatic and interactive analysis of chromatography-coupled small-angle X-ray scattering data*, *Bioinformatics* **34**, 1944-1946 (2018).
49. P. V. Konarev, V. V. Volkov, A. V. Sokolova, M. H. J. Koch, D. I. Svergun. *PRIMUS: a Windows PC-based system for small-angle scattering data analysis*, *J. Appl. Crystallogr.* **36**, 1277-1282 (2003).
50. P. V. Konarev, D. I. Svergun. *A posteriori determination of the useful data range for small-angle scattering experiments on dilute monodisperse systems*, *IUCrJ* **2**, 352-360 (2015).
51. A. Waterhouse *et al. SWISS-MODEL: homology modelling of protein structures and complexes*, *Nucleic Acids Res.* **46**, W296-W303 (2018).
52. D. Schneidman-Duhovny, M. Hammel, John A. Tainer, A. Sali. *Accurate SAXS Profile Computation and its Assessment by Contrast Variation Experiments*, *Biophys. J.* **105**, 962-974 (2013).
53. M. Guttman, P. Weinkam, A. Sali, Kelly K. Lee. *All-Atom Ensemble Modeling to Analyze Small-Angle X-Ray Scattering of Glycosylated Proteins*, *Structure* **21**, 321-331 (2013).
54. B. Hammouda. *A new Guinier-Porod model*, *J. Appl. Crystallogr.* **43**, 716-719 (2010).
55. D. Freedman, P. Diaconis. *On the Histogram as a Density Estimator: L 2 Theory*, *Z. Wahrscheinlichkeitstheorie* **57**, 453-476 (1981).
56. J. Huang *et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins*, *Nat. Methods* **14**, 71-73 (2017).
57. J. C. Phillips *et al. Scalable molecular dynamics with NAMD*, *J. Comput. Chem.* **26**, 1781-1802 (2005).
58. T. U. Consortium. *UniProt: a worldwide hub of protein knowledge*, *Nucleic Acids Res.* **47**, D506-D515 (2018).

59. S. El-Gebali *et al. The Pfam protein families database in 2019*, *Nucleic Acids Res.* **47**, D427-D432 (2018).
60. A. Lafita, P. Tian, R. B. Best, A. Bateman. *TADOSS: computational estimation of tandem domain swap stability*, *Bioinformatics* **35**, 2507-2508 (2019).
61. S. R. Eddy. *Accelerated Profile HMM Searches*, *PLOS Computational Biology* **7**, e1002195 (2011).

Figures and Tables

Fig. 1.

Close homologs of A8AW49 contain variable numbers of repeats that each form the novel 'SHIRT' fold. (*A*) Schematic of Sgo_0707 showing the N-terminal adhesin domains N1 and N2, tandem repeat (SHIRT) domains (red) and C-terminal LPNTG cell-wall crosslinking motif (black box 'L'). (*B*) Structure of Sgo_R10 and topology; sheets S1 (red) and S2 (blue) are highlighted in boxes**.** (*C*) A phylogenetic tree of Sgo_0707 homologs (>70% identity) identified using PHMMER (61) from ENSEMBL bacterial genomes using the N-terminal domain sequence (residue range 1-450 for Sgo_0707) and containing the LPXTG cell wall anchor motif at the C-termini. The number of SHIRT domain repeats in the stalk of each protein is shown as a bar plot with scale bar below.

Fig. 2.

The tandem SHIRT repeat Sgo_R3-4 adopts an anisotropic (head-to-tail) structure with limited inter-domain "bend" but significant "twist". (*A*) The structure of Sgo_R3-4 (left) and connecting short, well-ordered Pro-Ala-Pro inter-domain linker; alternate conformers of P704 are included (electron density $2mF_o$-DF_c map contoured at 0.1182 electrons/$Å^3$ blue chickenwire, right). (*B*) Frequency of domain lengths and (*C*) inter-domain angles (*SI Appendix*, Fig. S5*A*) over a 0.8 μs all-atom, fully solvated molecular dynamics simulation of Sgo_R3-4 at 303 K. The ends of the domains Sgo_R3, Sgo_R4 and Sgo_R3-4 and linker are identified by the Cα atoms of residues T623–P704, T707–A789 and T627–A789, respectively; arrows show the value of the distances in the crystal structure. (*D*) Models of 7 tandemly-arrayed domains based on the α, β, γ angles in c.

(*E*) Scaling of R_g of simulated model SHIRT constructs with increasing number of domains (or, equivalently, amino acids) compared to denatured and native proteins (approximated by blue and red lines, respectively). Error bars are standard deviations over 100 models generated.

Fig. 3.

Tandemly-arrayed SHIRT domains form a highly elongated rod. (*A*) Porod plots from SAXS analysis of Sgo_R3-4 (top) and Sgo_R2-8 (bottom). Porod exponents were calculated from the negative slope of the linear region of the data (R3-4, blue, R^2 = 1.0; R2-8, red, R^2 = 1.0). (*B*) Modified Guinier region for Sgo_R3-4 with cross-sectional radius of gyration (R_g^c, annotated) calculated from fitted region (blue line, R^2 = 0.98, q^*R_g limits = 0.3–1.1). (*C*) Modified Guinier region for Sgo_R2-8 with R_g^c calculated from fitted region (red line, R^2 = 0.99, q^*R_g limits = 0.3–1.0. (*D*) SHRImP-TIRFM determination of the inter-dye distances for AF488-labelled Sgo_R2-8$^{S666C/S1086C}$ (schematic inset) immobilised on 2 µg/mL (blue) or 20 µg/mL (red) poly-D-lysine treated quartz surfaces. Colour-coded, dashed lines indicate Gaussian fits to the histograms for the 2 µg/mL (n = 90, R^2 = 0.86) and 20 µg/mL (n = 85, R^2 = 0.90) poly-D-lysine-treated quartz surfaces (mean ± standard error = 25.5 ± 2.39 nm and 27.8 ± 2.21 nm, respectively).

Fig. 4.

Sequence clustering of long tandem highly identical repeats identified across proteins of the NCTC3000 genomes. The largest clusters are annotated using Pfam (Pfam IDs are found in Supplementary Dataset S1). Clusters with unknown Pfam classification are marked with “?”. New protein domain families built from sequence clusters into Pfam are highlighted in blue.

Fig. 5.

Variation of stalk region repeat numbers in Periscope Proteins. Phylogenetic trees of (*A*) *Staphylococcus aureus* (*B*) *Pseudomonas aeruginosa* (*C*) various streptococcal species including *S. agalactiae* and *S. pyogenes* and (*D*) *S. gordonii* genomes in the NCTC3000 collection, mapped to the number of repeats in stalk regions in Periscope genes; respectively, SasG with G5/E repeats, SAP077A_019 with Big_6 domain repeats, CdrA with MBG_2 domain repeats, surface protein Rib with Rib domain repeats, Sgo_0707 homolog containing SHIRT domain repeats, and surface adhesin CshA with ~100-residue globular domain repeats of a new Pfam family (CshA_repeat).

Fig. 6.

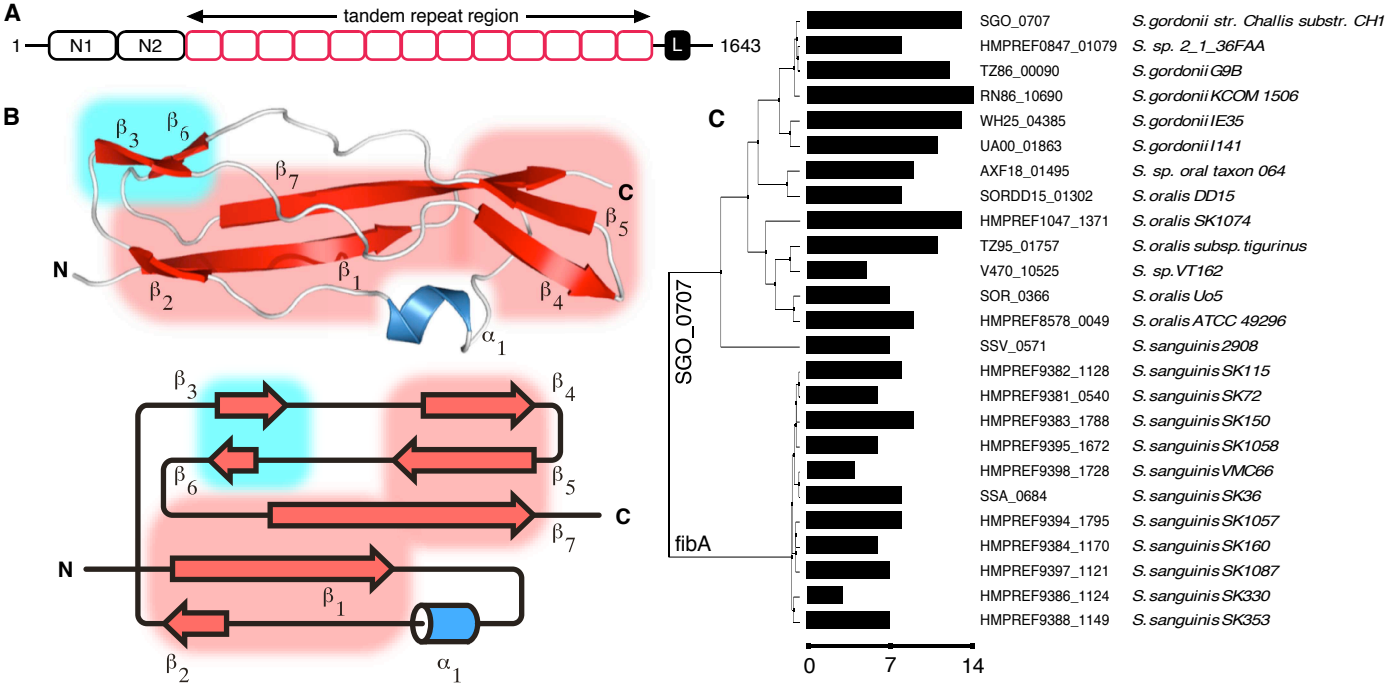
Repeat number variation in Periscope Proteins as a function of repeat sequence identity. The sequence identity of tandem repeats in the genome plotted against the variation in repeat number observed for each Periscope Protein cluster. The repeat number variation is calculated as the difference between the maximum and minimum observed repeat numbers in proteins within each cluster. The maximum number of repeats is shown as a viridis color scale. The repeat DNA identity is calculated as the maximum repeat sequence identity across proteins in each cluster. The number of proteins in each cluster is shown as point size. Names for the most relevant Periscope Protein clusters are shown as labels.

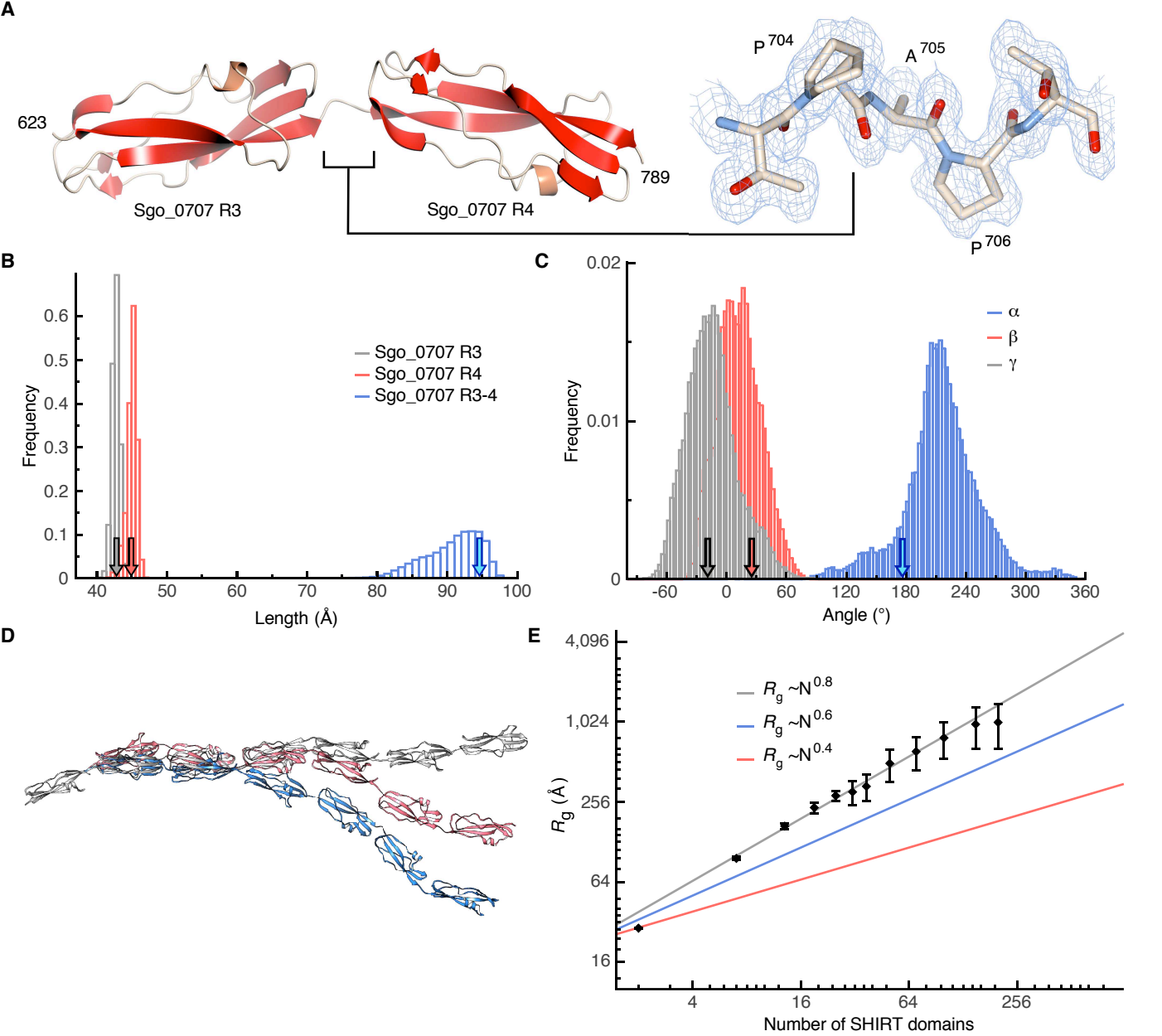
Table 1. Crystallographic data collection and refinement statistics

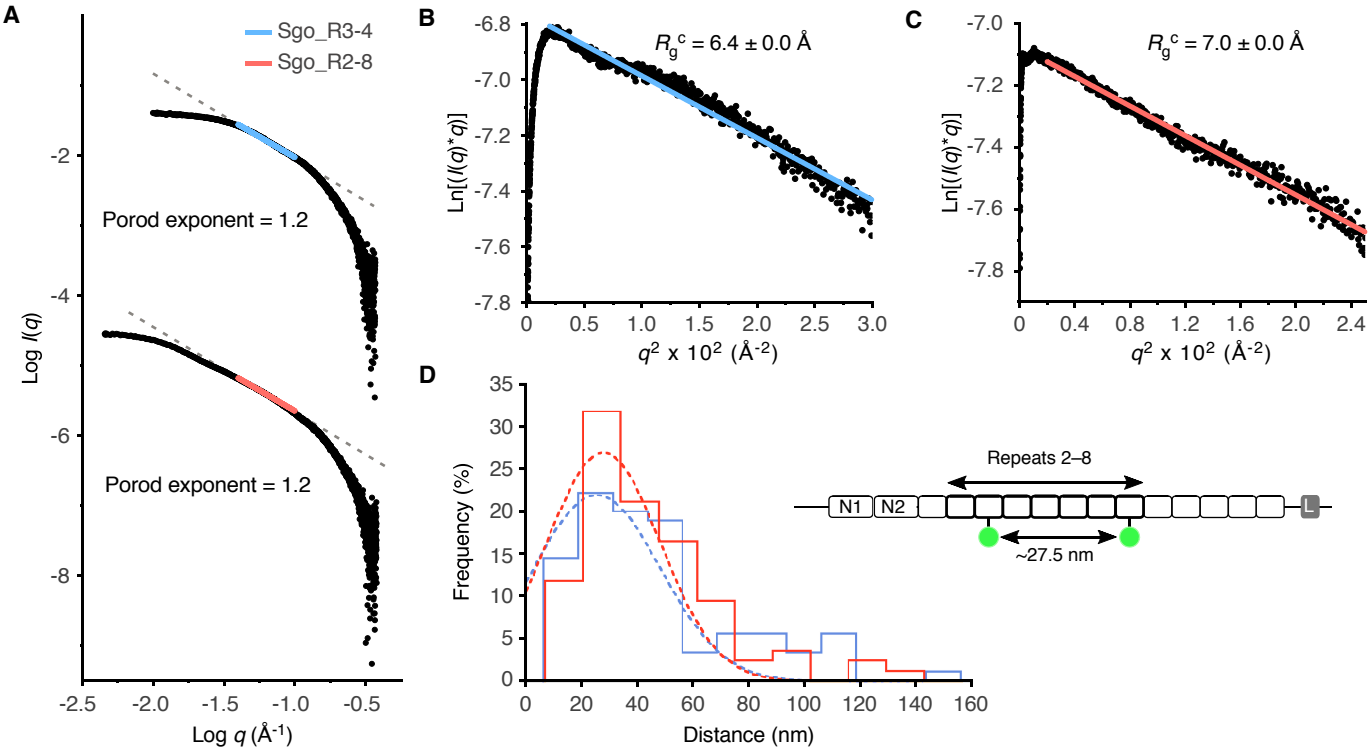
Data Collection Statistics	**ΔN-Sgo_R2**	**Sgo_R10**	**Sgo_R3-4**
Wavelength (Å)	0.85	0.78	0.976
Space group	$P2_12_12_1$	$P2_12_12$	$P2_1$
Cell dimensions a, b, c (Å) β (°)	 21.4, 40.8, 82.5 90.0	 65.4 48.0 48.4 90.0	 24.1, 37.2, 101.1 95.8
Resolution limits (Å)[1]	82.5—0.95 (0.96—0.95)	38.9—0.82 (0.85 —0.82)	37.2—1.35 (1.37—1.35)
No. reflections Total[1] Unique[1]	 281952 (12914) 44546 (2604)	 292386 (24481) 147184 (11281)	 157956 (7439) 39416 (1911)
R_{merge}[1,2]	0.034 (0.681)	0.045 (2.057)	0.066 (0.960)
Mean $I/\sigma I$[1]	24.8 (2.3)	28.3 (0.3)	8.9 (1.4)
Half-set correlation CC (1/2)[1]	0.999 (0.839)	1.000 (0.239)	0.998 (0.562)
Wilson B-factor (Å^2)	8.9	10.8	10.8
Completeness (%)[1]	95.2 (75.8)	96.8(76.5)	99.9 (99.6)
Redundancy[1]	6.3 (5.0)	2.0(1.9)	4.0 (3.9)
Refinement Statistics			
Resolution (Å)[1]	43.2 (0.95)	38.9 (0.82)	34.9 (1.35)
No. of reflections Working Free	 42240 2193	 144297 1454	 37511 1892
R_{work}/R_{free} (%)	12.5/13.3	14.7/16.2	13.4/17.0
Rms from ideality Bond length (Å) Angles (°)	 0.017 1.96	 0.017 1.48	 0.023 2.15
No. of atoms Protein Ligand/ion Water	 661 - 104	 1442 15 247	 1404 33 200
Average B (Å^2) Protein Ligand/ion Water	 14.9 - 28.3	 18.1 19.7 27.8	 18 22.4 34.0
Ramachandran angles (%) Favoured Allowed Outliers	 98.0 2.0 0.0	 98.15 1.85 0.0	 97.0 3.0 0.0

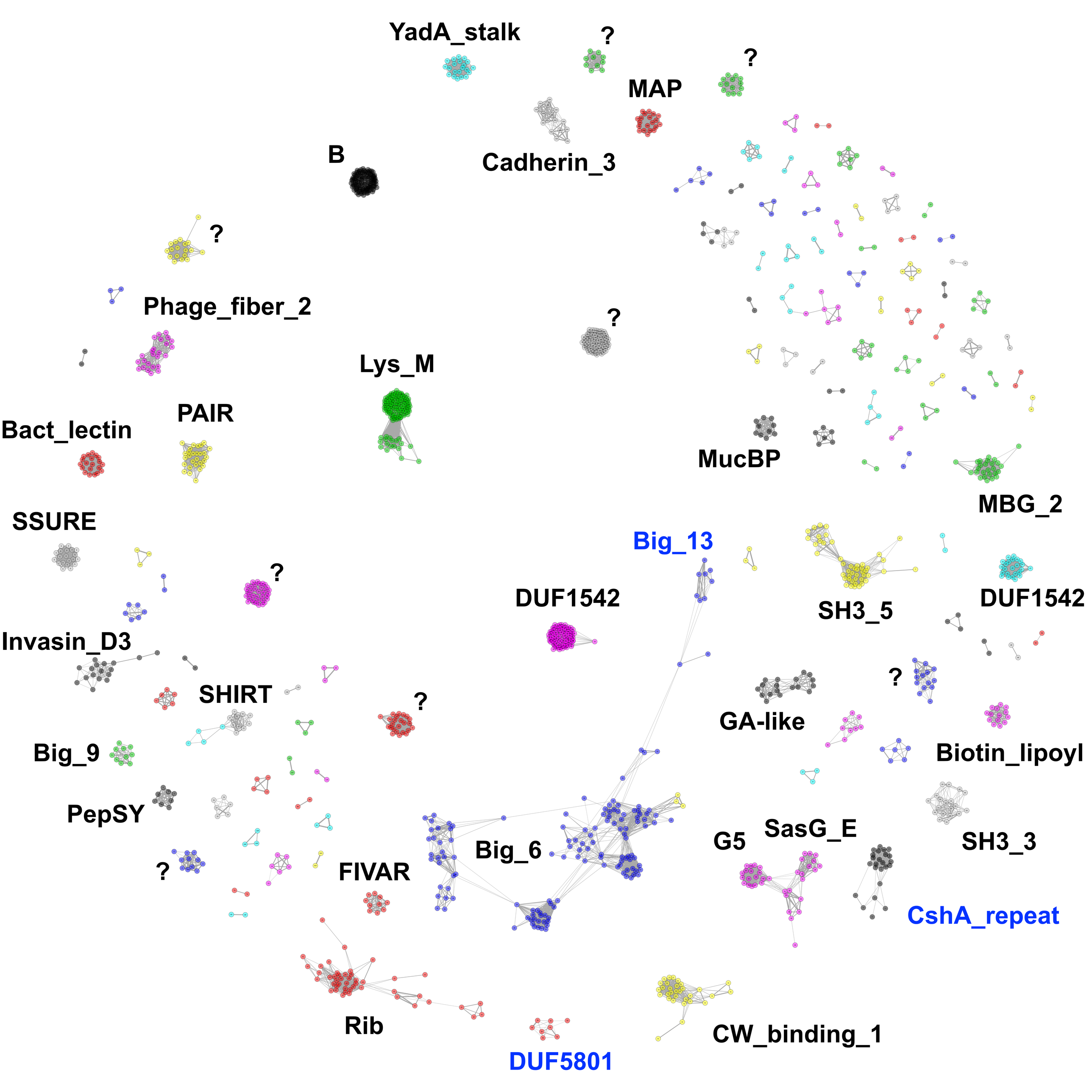
[1] Values in parentheses are for the highest resolution shell.

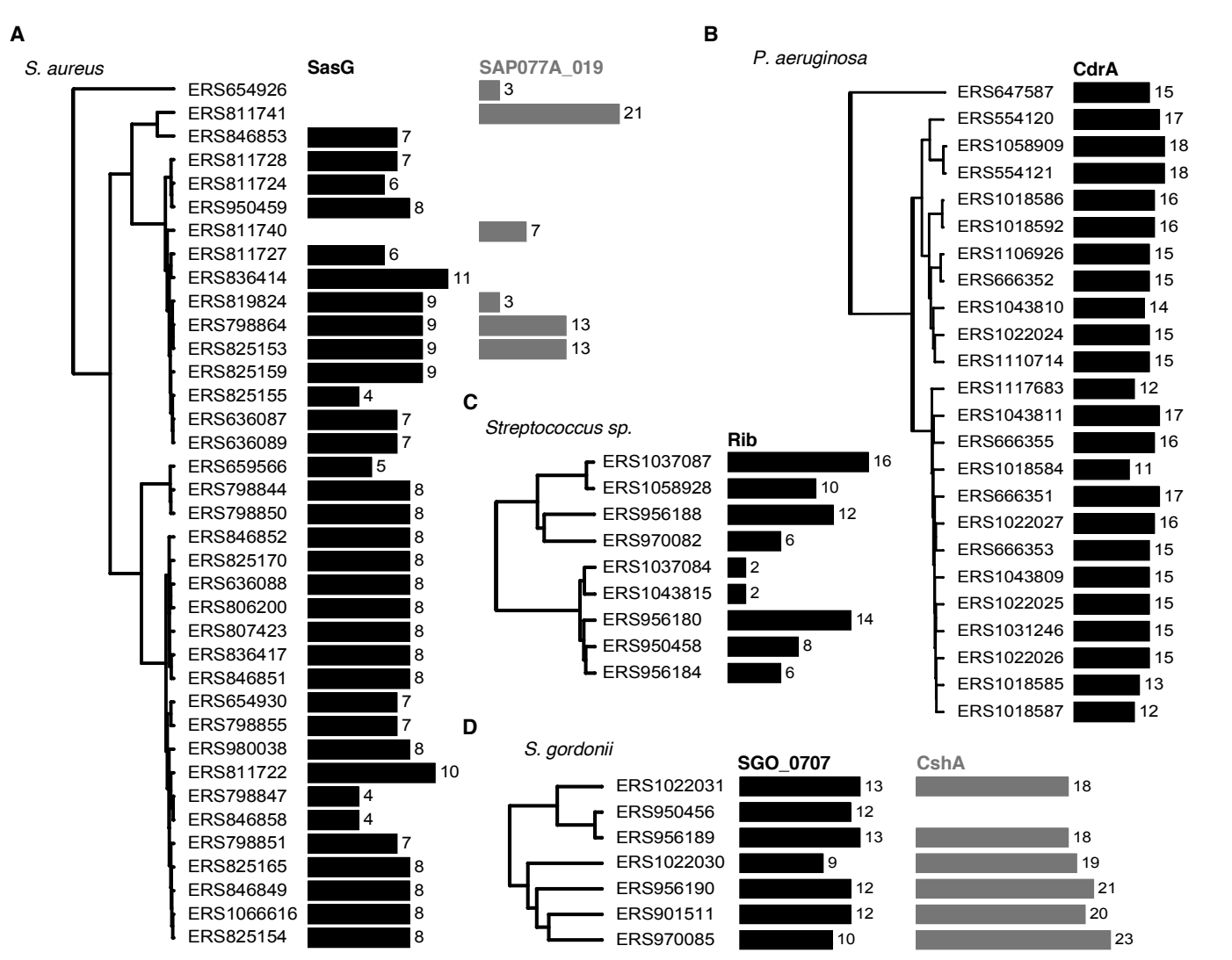
[2] $R_{merge} = \Sigma|I - \langle I\rangle| / \Sigma I$

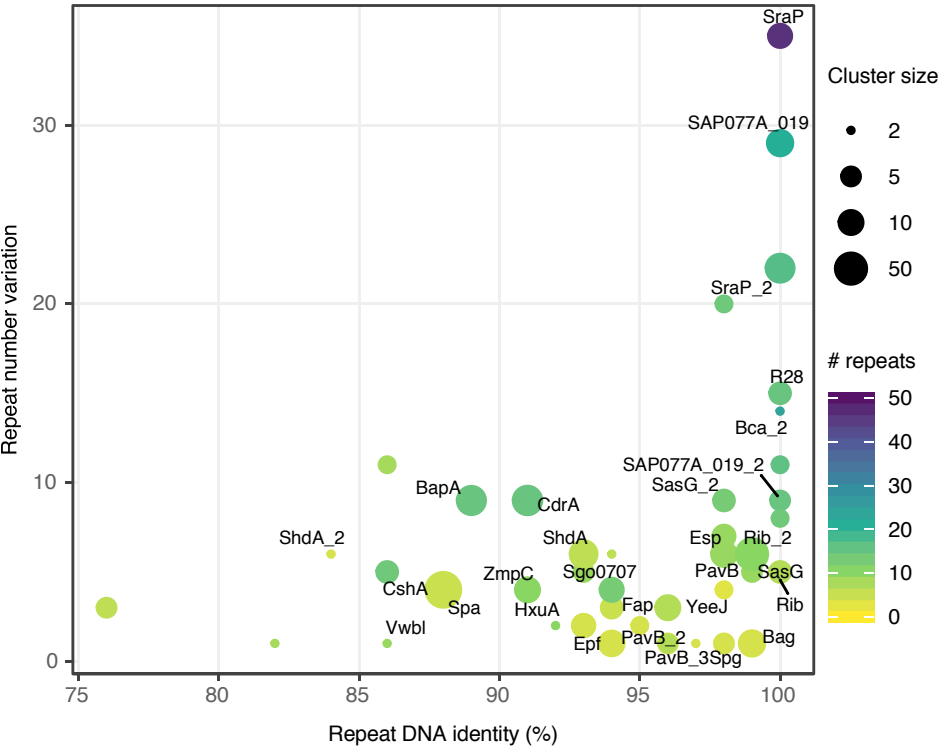

A
tandem repeat region
1
N1
N2
L
1643
B
β3
β6
β7
C
β5
N
β2
β1
β4
α1
β3
β4
β6
β5
C
β7
N
β1
β2
α1
C
SGO_0707
fibA
SGO_0707 S. gordonii str. Challis substr. CH1
HMPREF0847_01079 S. sp. 2_1_36FAA
TZ86_00090 S. gordonii G9B
RN86_10690 S. gordonii KCOM 1506
WH25_04385 S. gordonii IE35
UA00_01863 S. gordonii I141
AXF18_01495 S. sp. oral taxon 064
SORDD15_01302 S. oralis DD15
HMPREF1047_1371 S. oralis SK1074
TZ95_01757 S. oralis subsp. tigurinus
V470_10525 S. sp. VT162
SOR_0366 S. oralis Uo5
HMPREF8578_0049 S. oralis ATCC 49296
SSV_0571 S. sanguinis 2908
HMPREF9382_1128 S. sanguinis SK115
HMPREF9381_0540 S. sanguinis SK72
HMPREF9383_1788 S. sanguinis SK150
HMPREF9395_1672 S. sanguinis SK1058
HMPREF9398_1728 S. sanguinis VMC66
SSA_0684 S. sanguinis SK36
HMPREF9394_1795 S. sanguinis SK1057
HMPREF9384_1170 S. sanguinis SK160
HMPREF9397_1121 S. sanguinis SK1087
HMPREF9386_1124 S. sanguinis SK330
HMPREF9388_1149 S. sanguinis SK353
0
7
14


A
623
Sgo_0707 R3
Sgo_0707 R4
789
P 704
A 705
P 706
B
Frequency
Length (Å)
Sgo_0707 R3
Sgo_0707 R4
Sgo_0707 R3-4
C
Frequency
Angle (°)
α
β
γ
D
E
R_g (Å)
Number of SHIRT domains
R_g ~N$^{0.8}$
R_g ~N$^{0.6}$
R_g ~N$^{0.4}$


A
Sgo_R3-4
Sgo_R2-8
Porod exponent = 1.2
Porod exponent = 1.2
Log I(q)
Log q (Å-1)
B
Rgc = 6.4 ± 0.0 Å
Ln[I(q)*q]
q2 x 102 (Å-2)
C
Rgc = 7.0 ± 0.0 Å
Ln[I(q)*q]
q2 x 102 (Å-2)
D
Frequency (%)
Distance (nm)
Repeats 2–8
N1
N2
L
~27.5 nm


YadA_stalk
?
?
MAP
Cadherin_3
B
?
Phage_fiber_2
?
Lys_M
PAIR
Bact_lectin
MucBP
MBG_2
SSURE
Big_13
?
DUF1542
SH3_5
DUF1542
Invasin_D3
?
SHIRT
?
GA-like
Big_9
Biotin_lipoyl
PepSY
SH3_3
G5
SasG_E
Big_6
?
FIVAR
CshA_repeat
Rib
CW_binding_1
DUF5801


A
S. aureus
SasG
SAP077A_019
ERS654926
ERS811741
ERS846853 7
ERS811728 7
ERS811724 6
ERS950459 8
ERS811740
ERS811727 6
ERS836414 11
ERS819824 9
ERS798864 9
ERS825153 9
ERS825159 9
ERS825155 4
ERS636087 7
ERS636089 7
ERS659566 5
ERS798844 8
ERS798850 8
ERS846852 8
ERS825170 8
ERS636088 8
ERS806200 8
ERS807423 8
ERS836417 8
ERS846851 8
ERS654930 7
ERS798855 7
ERS980038 8
ERS811722 10
ERS798847 4
ERS846858 4
ERS798851 7
ERS825165 8
ERS846849 8
ERS1066616 8
ERS825154 8
3
21
7
3
13
13
B
P. aeruginosa
CdrA
ERS647587 15
ERS554120 17
ERS1058909 18
ERS554121 18
ERS1018586 16
ERS1018592 16
ERS1106926 15
ERS666352 15
ERS1043810 14
ERS1022024 15
ERS1110714 15
ERS1117683 12
ERS1043811 17
ERS666355 16
ERS1018584 11
ERS666351 17
ERS1022027 16
ERS666353 15
ERS1043809 15
ERS1022025 15
ERS1031246 15
ERS1022026 15
ERS1018585 13
ERS1018587 12
C
Streptococcus sp.
Rib
ERS1037087 16
ERS1058928 10
ERS956188 12
ERS970082 6
ERS1037084 2
ERS1043815 2
ERS956180 14
ERS950458 8
ERS956184 6
D
S. gordonii
SGO_0707
CshA
ERS1022031 13 18
ERS950456 12
ERS956189 13 18
ERS1022030 9 19
ERS956190 12 21
ERS901511 12 20
ERS970085 10 23


SraP
SAP077A_019
SraP_2
R28
Bca_2
SAP077A_019_2
SasG_2
BapA
CdrA
ShdA_2
ShdA
Esp
Rib_2
PavB
SasG
Rib
ZmpC
Sgo0707
CshA
Spa
HxuA
Fap
YeeJ
Vwbl
Epf
PavB_2
PavB_3
Spg
Bag
Repeat number variation
Repeat DNA identity (%)
Cluster size
2
5
10
50
repeats
50
40
30
20
10
0
0
10
20
30
75
80
85
90
95
100