JBC ARTICLE

⌘ *Author's Choice*

The major subunit of widespread competence pili exhibits a novel and conserved type IV pilin fold

Received for publication, March 3, 2020, and in revised form, April 7, 2020 Published, Papers in Press, April 9, 2020, DOI 10.1074/jbc.RA120.013316

Devon Sheppard[‡], **Jamie-Lee Berry**[‡], **Rémi Denise**[§¶], **Eduardo P. C. Rocha**[§], **Steve Matthews**[‖], **and Vladimir Pelicic**[‡1]

From the [‡]*Medical Research Council Centre for Molecular Bacteriology and Infection, Imperial College London, London SW7 2AZ, United Kingdom,* [§]*Microbial Evolutionary Genomics, Institut Pasteur, CNRS UMR3525, Paris 75015, France, the* [¶]*Sorbonne Université, Collège doctoral, Paris 75005, France, and the* [‖]*Centre for Structural Biology, Imperial College London, London SW7 2AZ, United Kingdom*

Edited by Chris Whitfield

Type IV filaments (T4F), which are helical assemblies of type IV pilins, constitute a superfamily of filamentous nanomachines virtually ubiquitous in prokaryotes that mediate a wide variety of functions. The competence (Com) pilus is a widespread T4F, mediating DNA uptake (the first step in natural transformation) in bacteria with one membrane (monoderms), an important mechanism of horizontal gene transfer. Here, we report the results of genomic, phylogenetic, and structural analyses of ComGC, the major pilin subunit of Com pili. By performing a global comparative analysis, we show that Com pili genes are virtually ubiquitous in Bacilli, a major monoderm class of Firmicutes. This also revealed that ComGC displays extensive sequence conservation, defining a monophyletic group among type IV pilins. We further report ComGC solution structures from two naturally competent human pathogens, *Streptococcus sanguinis* (ComGC$_{SS}$) and *Streptococcus pneumoniae* (ComGC$_{SP}$), revealing that this pilin displays extensive structural conservation. Strikingly, ComGC$_{SS}$ and ComGC$_{SP}$ exhibit a novel type IV pilin fold that is purely helical. Results from homology modeling analyses suggest that the unusual structure of ComGC is compatible with helical filament assembly. Because ComGC displays such a widespread distribution, these results have implications for hundreds of monoderm species.

Filamentous nanomachines composed of type IV pilins are virtually ubiquitous in Bacteria and Archaea (1), to which they confer a variety of unrelated functions including adhesion, motility, protein secretion, and DNA uptake. These type IV filaments (T4F)[2] are assembled by conserved multiprotein machineries, which further underlines their phylogenetic relationship (2).

Much of our current understanding of this superfamily of nanomachines comes from the study of type IV pili (T4P), the best characterized T4F (1). In brief, T4P are micrometer-long and thin surface-exposed filaments, which are polymers of type IV pilins. Type IV pilins (simply named pilins hereafter) are defined by an N-terminal sequence motif known as class III signal peptide (3). This motif, IPR012902 entry in the InterPro database (4), consists of a hydrophilic leader peptide ending with a tiny residue (Gly or Ala), followed by a tract of 21 mostly hydrophobic residues, except for a negatively charged Glu^5. This hydrophobic tract represents the N-terminal portion (α1N) of an extended α-helix of ~50 residues (α1), which is the universally conserved structural feature of type IV pilins (3). Although some small pilins consist solely of this extended α-helix (5), most pilins have a globular head consisting of the C-terminal half of α1 (α1C) packed against a β-sheet composed of several antiparallel β-strands, which gives them their typical "lollipop" 3D architecture (3). Upon translocation of prepilins across the cytoplasmic membrane (CM), where they remain embedded via their protruding hydrophobic α1N, the leader peptide is processed by an integral membrane aspartic acid protease named prepilin peptidase (IPR000045) (6). Processing primes pilins for polymerization into filaments. Filament assembly, which remains incompletely understood, is mediated by a multiprotein machinery in the CM, centered on an integral membrane platform protein (IPR003004) and a cytoplasmic extension ATPase (IPR007831) (1). As revealed by recent cryo-EM structures of several T4P (7, 8), filaments are right-handed helical polymers where pilins are held together by extensive interactions between their α1 helices, which are partially melted and run approximately parallel to each other within the filament core.

One of the key functional roles of T4F is their involvement in natural transformation in prokaryotes, the ability of species defined as "competent" to take up exogenous DNA across their membrane(s) and incorporate it stably into their genomes (9). This widespread property in bacteria (10) is key for horizontal gene transfer, an important factor in bacterial evolution and the

This work was supported by Medical Research Council Grant MR/P022197/1 (to V. P.), Doctoral School Complexité du vivant-ED515 contract number 2449/2016 (to R. D.), and INCEPTION project PIA/ANR-16-CONV-0005 (to E. P. C. R.). The authors declare that they have no conflicts of interest with the contents of this article.

This article contains Figs. S1–S8, Tables S1–S3, and Spreadsheets S1 and S2.

[1] To whom correspondence should be addressed. Tel.: 44-20-7594-2080; E-mail: v.pelicic@imperial.ac.uk.

[2] The abbreviations used are: T4F, type IV filaments; T4P, type IV pili; Com, competence; UFBoot, ultrafast bootstrap; RDC, residual dipolar couplings; RMSD, root mean square deviation; PDB, Protein Data Bank; LB, Lysogenic Broth; IPTG, isopropyl β-D-1-thiogalactopyranoside; CM, cytoplasmic membrane; MSH, mannose-sensitive hemagglutinin pili; T2SS, type II secretion system; RMSD, root mean square deviation; Ni-NTA, nickel-nitrilotriacetic acid; HSQC, heteronuclear single quantum coherence.

spread of antibiotic resistance. T4F are involved in the very first step of natural transformation, *i.e.* binding of free extracellular DNA and its translocation close to the CM (9). DNA is subsequently bound by the DNA receptor ComEA and further translocated across the CM through the ComEC channel (9). In diderm-competent species, the T4F involved in DNA uptake is a subtype of T4P, known as T4aP (11), which rapid depolymerization is powered by the retraction ATPase PilT (IPR006321), generating exceptionally large tensile forces (12). In brief, T4aP bind DNA directly, via one of their major or minor (low abundance) pilin subunits (13), and then are retracted by PilT, bringing DNA to the ComEA receptor (14). In monoderm-competent species, DNA uptake is mediated by a distinct T4F named competence (Com) pilus (9), much less well-characterized than T4P. Com pili are composed mainly of the major pilin (ComGC) (15, 16), and are assembled by a simple machinery composed of four minor pilins (ComGD, ComGE, ComGF, and ComGG), a prepilin peptidase (ComC), an extension ATPase (ComGA) and a platform protein (ComGB) (17, 18). Filaments morphologically similar to T4aP, several micrometer in length and 60 Å in width, have been observed in *Streptococcus pneumoniae* (15, 19).

How Com pili are assembled, bind DNA, and presumably retract in the absence of a PilT retraction motor is not understood. One important limitation is the absence of high-resolution structural information. Therefore, in the present study, we have focused on ComGC, the major subunit of the Com pilus. We report (i) a global comparative and phylogenetic analysis of ComGC, and (ii) 3D structures for two orthologs, ComGC_{SP} from the model competent species *S. pneumoniae* and ComGC_{SS} from *Streptococcus sanguinis*, a common cause of infective endocarditis in humans that has recently emerged as a monoderm model for the study of T4F. Finally, we discuss the general implications of these findings.

Results

Com pili genes are almost ubiquitous in monoderm Bacilli, including the T4F model S. sanguinis

So far, Com pili have been mainly studied in two model-competent species: *Bacillus subtilis* and *S. pneumoniae*. *S. sanguinis* is a naturally competent species that has recently emerged as a monoderm T4F model because it expresses retractable T4aP (20). Functional analysis of *S. sanguinis* T4aP showed that they are dispensable for DNA uptake, which is instead mediated by Com pili because competence was abolished in a Δ*comGB* mutant (21). A closer inspection of *S. sanguinis* genome revealed that all the genes encoding the Com pilus are present. These genes are organized in two loci (Fig. 1*A*), *comC* and the *comGABCDEDFG* operon, showing perfect synteny with the corresponding loci in model competent species (22, 23). Multiple sequence alignments of the corresponding proteins with orthologs in *B. subtilis* and *S. pneumoniae* showed extensive conservation (Table S1). Detailed sequence analysis of the N termini of the five ComG pilins identified clear class III signal peptides (Fig. 1*B*), *i.e.* short (8–15 residues) and hydrophilic leader peptides ending with an Ala, followed by a tract of 21 mostly hydrophobic residues. ComGG is the only pilin that does not have a negatively charged Glu^5 and displays a noncanonical class III signal peptide (Fig. 1*B*), which is not identified by InterPro or PilFind that is dedicated to the prediction of type IV pilins (24). This is a conserved property for ComGG orthologs.

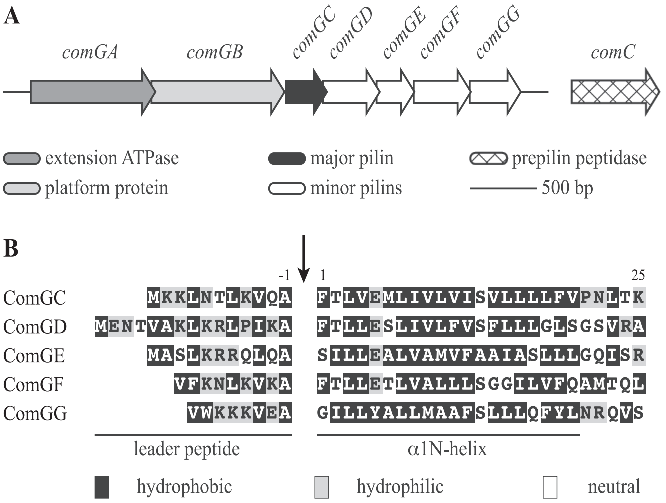
A
comGA
comGB
comGC
comGD
comGE
comGF
comGG
comC
extension ATPase
platform protein
major pilin
minor pilins
prepilin peptidase
500 bp
B
-1
1
25
ComGC MKKLNTLKVQA FTLVEMLIVLVISVLLLLFVPNLTK
ComGD MENTVAKLKRLPIKA FTLLESLIVLFVSFLLLGLSGSVRA
ComGE MASLKRRQLQA SILLEALVAMVFAAIASLLLGQISR
ComGF VFKNLKVKA FTLLETLVALLLSGGILVFQAMTQL
ComGG VWKKKVEA GILLYALLMAAFSLLLQFYLNRQVS
leader peptide
α1N-helix
hydrophobic
hydrophilic
neutral

Figure 1. Com pilus machinery in *S. sanguinis*. *A,* genomic organization of the genes involved in the biogenesis of the Com pilus in *S. sanguinis* 2908. All the genes are drawn to scale, with the *scale bar* representing 500 bp. The functions of the corresponding proteins are listed at the *bottom*. *B,* sequence alignment of the putative N-terminal class III signal peptides of the five ComG pilins in *S. sanguinis* 2908. The 8–15–amino acid long leader peptides, which contain a majority of hydrophilic (shaded in *gray*) or neutral (*no shading*) residues, end with a conserved Ala^{-1}. Leader peptides are processed (indicated by the *vertical arrow*) by the prepilin peptidase ComC. The mature proteins start with a tract of 21 predominantly hydrophobic residues (shaded in *black*), which invariably form the protruding N-terminal portion of an extended α-helix that is the main assembly interface within filaments.

We next determined the global distribution of the Com system in publicly available complete bacterial genomes using MacSyFinder (25). Specifically, we used the MacSyFinder model built for the identification of Com systems (2), which takes into account the genetic composition and organization of its components. This showed that the Com system is restricted to Firmicutes, a phylum comprising a vast majority of monoderms, where it is exceptionally widespread because it was detected in 2,333 genomes (Spreadsheet S1). An overwhelming majority of the corresponding species (99.7%) belong to the taxonomic class Bacilli (equally distributed among the Bacillales and Lactobacillales orders). As many as 88.7% of the sequenced Bacilli have a Com system. We also detected Com systems in one Clostridia (of 336) and six Erysipelotrichia (of 14). In total, 349 different species have the potential to express a Com pilus (Spreadsheet S2). Taken together, these findings suggest that the Com pilus is almost ubiquitous in Bacilli and can be advantageously studied in *S. sanguinis*.

ComGC, the major subunit of Com pili, is highly conserved and defines a monophyletic group among type IV pilins

We next focused specifically on the major subunit of Com pili, the pilin ComGC (15, 16). Compared with major pilins from T4aP, ComGC is ∼40% shorter, with 94 or 93 amino acids for the processed ComGC_{SS} and ComGC_{SP}, respectively (10.2

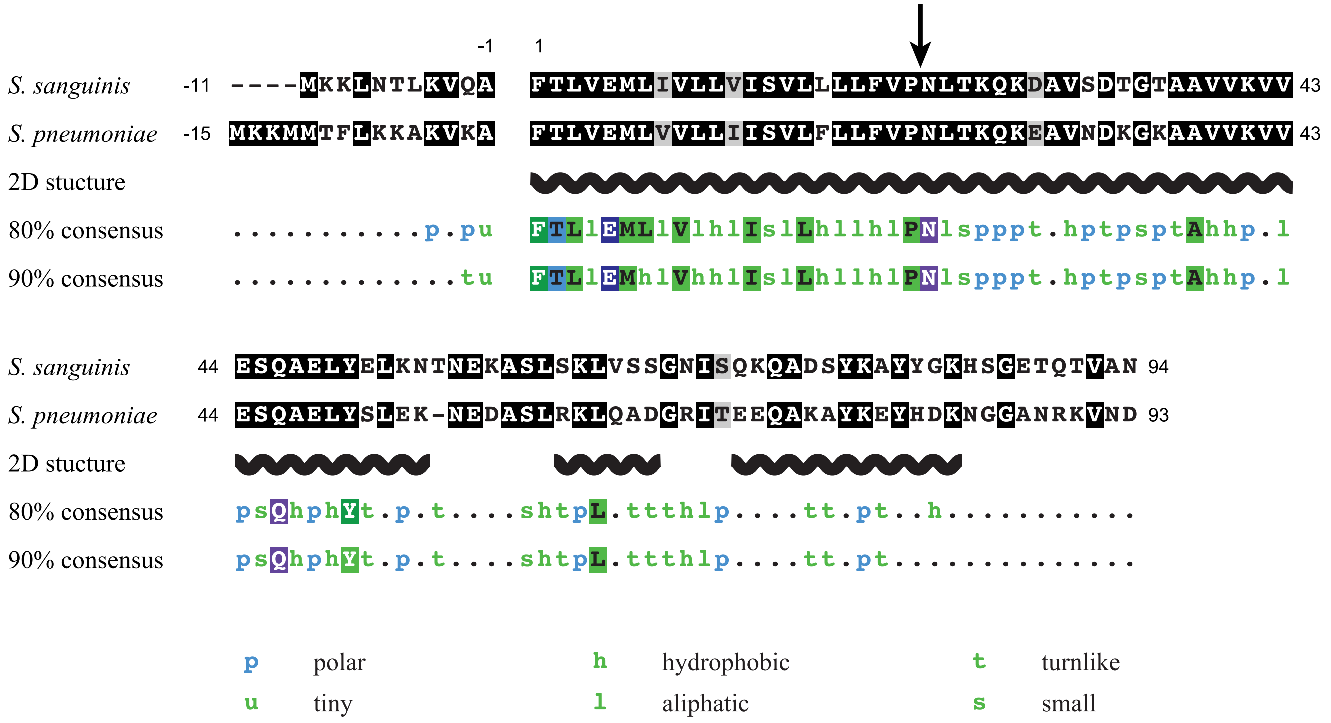
-1
1
S. sanguinis -11 ----MKKLNTLKVQA FTLVEMLIVLLVISVLLLLFVPNLTKQKDAVSDTGTAAVVKVV 43
S. pneumoniae -15 MKKMMTFLKKAKVKA FTLVEMLVVLLIISVLFLLFVPNLTKQKEAVNDKGKAAVVKVV 43
2D stucture
80% consensus ....p.pu FTLlEMLlVlhlIslLhllhlPNlspppt.hptpsptAhhp.l
90% consensus ....tu FTLlEMhlVhhlIslLhllhlPNlspppt.hptpsptAhhp.l
S. sanguinis 44 ESQAELYELKNTNEKASLSKLVSSGNISQKQADSYKAYYGKHSGETQTVAN 94
S. pneumoniae 44 ESQAELYSLEK-NEDASLRKLQADGRITEEQAKAYKEYHDKNGGANRKVND 93
2D stucture
80% consensus psQhphYt.p.t....shtpL.ttthlp....tt.pt..h...........
90% consensus psQhphYt.p.t....shtpL.ttthlp....tt.pt..............
p polar
h hydrophobic
t turnlike
u tiny
l aliphatic
s small

Figure 2. Global sequence analysis of ComGC pilins. Sequence alignments of ComGC in *S. sanguinis* and *S. pneumoniae* is represented in the *top two rows*. Residues were shaded in *black* (identical), *gray* (conserved), or *unshaded* (different). The leader peptide is *highlighted*. In the recombinant proteins that were produced for structure determination, the N-terminal 22 residues invariably forming a protruding hydrophobic α-helix were truncated (depicted by an *arrow*) to promote solubility. The 2D structural motifs predicted using JPred are depicted in the *third row*. *Fourth* and *fifth rows* represent the 80 and 90% ComGC consensus sequences, computed from 2,809 ComGC entries in InterPro, and aligned to ComGC$_{SS}$ and ComGC$_{SP}$. Multiple alignments were generated using Clustal Omega and formatted with MView. Polar: C, D, E, H, K, N, Q, R, S, or T. Tiny: A or G. Hydrophobic: A, C, F, G, H, I, K, L, M, R, T, V, W, or Y. Aliphatic: I, L or V. Turn-like: A, C, D, E, G, H, K, N, Q, R, S, or T. Small: A, C, D, G, N, P, S, T, or V. Single letter abbreviations are used.

and 10.4 kDa). Moreover, unlike most other pilins, in which the only detectable sequence homology is usually in the α1N portion of the class III signal peptide (3), ComGC orthologs show extensive sequence identity. For example, processed ComGC$_{SS}$ and ComGC$_{SP}$ display 65.6% overall sequence identity (Fig. 2). Similarly, processed ComGC$_{SS}$ and ComGC$_{BS}$ (from *B. subtilis*) show 33.3% sequence identity overall (Fig. S1). This is consistent with the existence of a ComGC signature in the InterPro database (IPR016940) (4), which lists 2,809 ComGC entries. Global multiple alignment of these ComGC proteins shows that most of the sequence is conserved in ~90% of the entries (Fig. 2). In Fig. 2, the consensus sequences have been aligned to ComGC$_{SS}$ and ComGC$_{SP}$. Strikingly, some residues show sequence identity in virtually all the entries, including residues outside of the α1N portion (such as Ala^{38}, Gln^{46}, Tyr^{50}, and Leu^{64} in ComGC$_{SS}$).

The above observations suggest that Com pili form a highly homogeneous T4F subfamily. This was tested by performing a phylogenetic analysis based on the protein sequences of major pilins from different T4F found in a wide variety of bacteria, including T4aP, T4bP, T4cP (also known as Tad pili), mannose-sensitive hemagglutinin pili (MSH), type II secretion systems (T2SS), and Com pili. The phylogeny tree that was generated (Fig. 3), using IQ-TREE (26), reveals that several T4F are in clear monophyletic groups with good branch support, >96% ultrafast bootstrap (UFBoot) (27). Of particular interest, Com pili define a highly supported clade (99% UFBoot), clearly distinct from all other T4F systems. Taken together, these findings show that ComGC is a small pilin with a highly conserved sequence, which defines a monophyletic group.

Solution structure of two ComGC orthologs reveal a conserved and new type IV pilin fold

Because high-resolution structural information is needed to improve our understanding of Com pili, we decided to solve the 3D structure of ComGC$_{SS}$. To facilitate protein purification, we used a synthetic *comGC$_{SS}$* gene codon-optimized for expression in *Escherichia coli*, and produced a soluble protein in which the first 22 residues of ComGC$_{SS}$ that form a hydrophobic α-helix (α1N) were replaced by a noncleavable N-terminal hexahistidine tag (His$_6$). This commonly used truncation approach is predicted to have minimal structural impact on the rest of the protein, as previously shown for the *Pseudomonas aeruginosa* PAK pilin (28). The resulting 8.8-kDa His$_6$-ComGC$_{SS}$ protein could be readily purified using a combination of affinity and gel-filtration chromatography. After purification of isotopically labeled protein with ^{13}C and ^{15}N for backbone and side chain NMR resonance assignments, we could assign 99.5% of the backbone and 92% of assignable protons overall. Structural ensembles were determined with 962 NOE-based restraints, 50 hydrogen bonds, 110 dihedral angles restraints, and 39 residual dipolar couplings (RDC) (Table 1). As can be seen in Fig. 4, ComGC$_{SS}$ 3D structure is unlike that of any type IV pilin present in the PDB, as it is purely helical, with three distinct helices connected by loops. The helices present

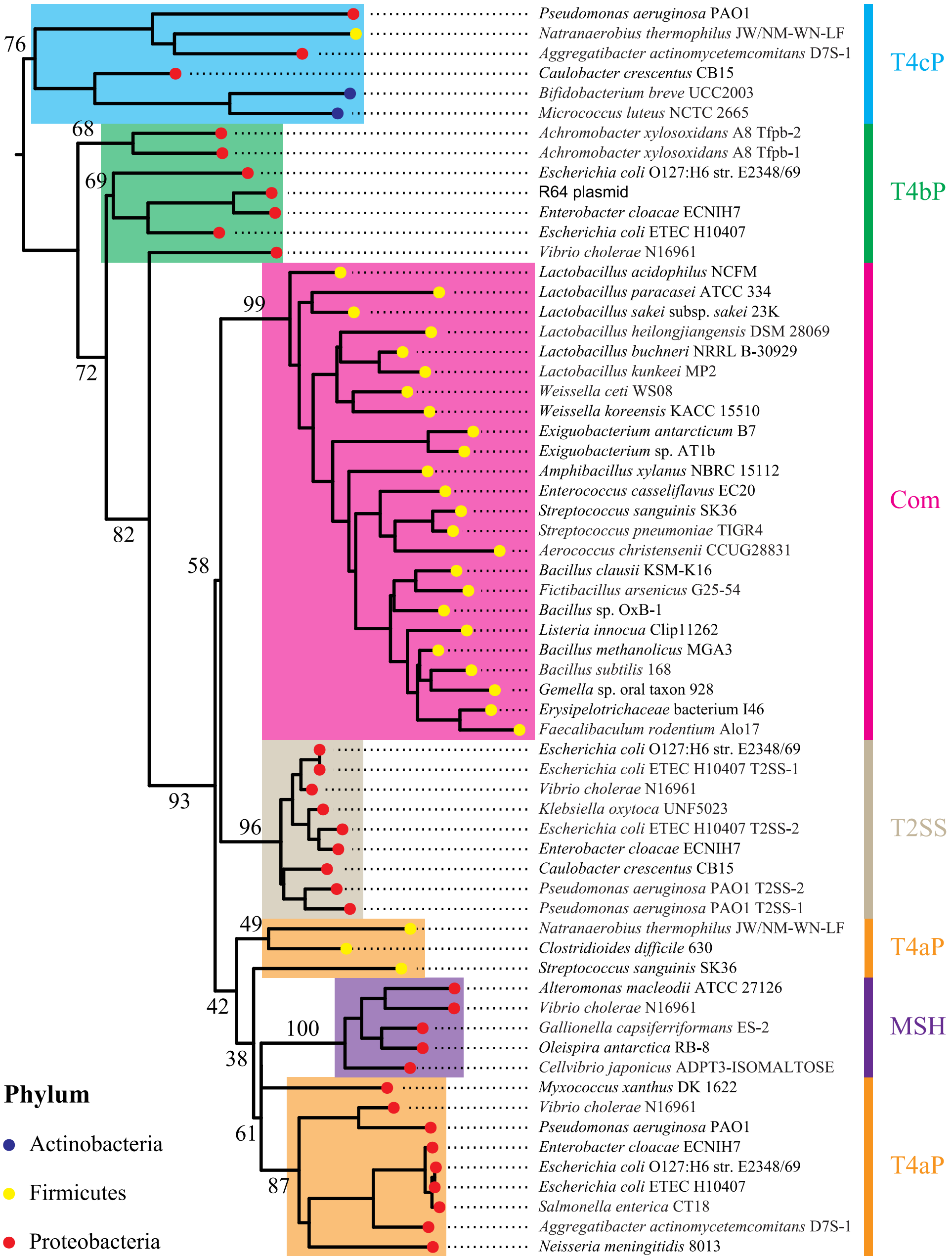
76
68
69
72
82
58
99
93
96
49
42
100
38
61
87
Pseudomonas aeruginosa PAO1
Natranaerobius thermophilus JW/NM-WN-LF
Aggregatibacter actinomycetemcomitans D7S-1
Caulobacter crescentus CB15
Bifidobacterium breve UCC2003
Micrococcus luteus NCTC 2665
Achromobacter xylosoxidans A8 Tfpb-2
Achromobacter xylosoxidans A8 Tfpb-1
Escherichia coli O127:H6 str. E2348/69
R64 plasmid
Enterobacter cloacae ECNIH7
Escherichia coli ETEC H10407
Vibrio cholerae N16961
Lactobacillus acidophilus NCFM
Lactobacillus paracasei ATCC 334
Lactobacillus sakei subsp. sakei 23K
Lactobacillus heilongjiangensis DSM 28069
Lactobacillus buchneri NRRL B-30929
Lactobacillus kunkeei MP2
Weissella ceti WS08
Weissella koreensis KACC 15510
Exiguobacterium antarcticum B7
Exiguobacterium sp. AT1b
Amphibacillus xylanus NBRC 15112
Enterococcus casseliflavus EC20
Streptococcus sanguinis SK36
Streptococcus pneumoniae TIGR4
Aerococcus christensenii CCUG28831
Bacillus clausii KSM-K16
Fictibacillus arsenicus G25-54
Bacillus sp. OxB-1
Listeria innocua Clip11262
Bacillus methanolicus MGA3
Bacillus subtilis 168
Gemella sp. oral taxon 928
Erysipelotrichaceae bacterium I46
Faecalibaculum rodentium Alo17
Escherichia coli O127:H6 str. E2348/69
Escherichia coli ETEC H10407 T2SS-1
Vibrio cholerae N16961
Klebsiella oxytoca UNF5023
Escherichia coli ETEC H10407 T2SS-2
Enterobacter cloacae ECNIH7
Caulobacter crescentus CB15
Pseudomonas aeruginosa PAO1 T2SS-2
Pseudomonas aeruginosa PAO1 T2SS-1
Natranaerobius thermophilus JW/NM-WN-LF
Clostridioides difficile 630
Streptococcus sanguinis SK36
Alteromonas macleodii ATCC 27126
Vibrio cholerae N16961
Gallionella capsiferriformans ES-2
Oleispira antarctica RB-8
Cellvibrio japonicus ADPT3-ISOMALTOSE
Myxococcus xanthus DK 1622
Vibrio cholerae N16961
Pseudomonas aeruginosa PAO1
Enterobacter cloacae ECNIH7
Escherichia coli O127:H6 str. E2348/69
Escherichia coli ETEC H10407
Salmonella enterica CT18
Aggregatibacter actinomycetemcomitans D7S-1
Neisseria meningitidis 8013
T4cP
T4bP
Com
T2SS
T4aP
MSH
T4aP
Phylum
● Actinobacteria
● Firmicutes
● Proteobacteria

Figure 3. Rooted phylogeny of the major pilins from various bacterial T4F. The tree was build using IQ-Tree, with 1,000 replicates of UFBoot and LG+F+R4 model. Numeric values (in %) indicate UFBoot of the corresponding branches. The *color of the bullet points* indicates the taxonomic group of the corresponding species. The *color of the strips* and *highlights* indicate the classification of the different T4F systems. *T4aP*, type IVa pilus. *T4bP*, type IVb pilus. *T4cP*, type IVc pilus (also known as Tad).

are consistent with JPred secondary structure prediction (Fig. 2) (29). The N-terminal α1-helix, which involves residues 37–53 of the processed protein, corresponds to α1C because the hydrophobic α1N has been truncated in His$_6$-ComGC$_{SS}$. Tightly packed against this α1-helix, in a parallel plane, are α2-helix (residues 61–67) and α3-helix (residues 72–85), which stack against each other in antiparallel fashion (Fig. 4*A*) and orthogonally to α1. Except for the N-terminal unstructured residues, the ComGC$_{SS}$ structures within the NMR ensemble superpose well onto each other (Fig. 4*B*), with a root mean square deviation (RMSD) of 1.2 Å for Cα atoms, which suggests that there is no significant flexibility in this portion of the structure (30). The unstructured N terminus, which lacks long and medium NOEs present in the ordered regions of the protein, was predicted to be highly dynamic based on TALOS+ (31), with an average S^2 order parameter of 0.49 ± 0.10.

Our ComGC$_{SS}$ structure differs markedly from the recently reported solution structure of ComGC$_{SP}$ (PDB 5NCA) (19), which is surprising considering the high sequence identity between these two proteins (Fig. 2). Therefore, to define the structural relationship between ComGC orthologs, we decided to solve the structure of ComGC$_{SP}$. As above, we used a synthetic *comGC$_{SP}$* gene codon-optimized for expression in *E. coli*, we fused the 71-amino acid long soluble portion of ComGC$_{SP}$ to a noncleavable N-terminal His$_6$ tag and purified doubly labeled His$_6$-ComGC$_{SP}$ (9 kDa). Again, assignment was excel-

Table 1
NMR structural statistics

	$ComGC_{SS}$	$ComGC_{SP}$
NOE-derived distance constraints		
Long [(*i-j*) > 5]	128	95
Medium [5 ≥ (*i-j*) > 1]	414	404
Intraresidue (*i* = *j*)	420	381
Total	962	880
Hydrogen bonds	50	54
Dihedral constraints (Φ and Ψ)	110	102
Residual dipolar couplings (RDC)	39	38
Ramachandran statistics (from PROCHECK)		
Most favored (%)	93.4	83.0
Additionally allowed (%)	6.4	15.9
Generously allowed (%)	0.2	1.1
Disallowed (%)	0.0	0.0
Structure statistics		
RMSD backbone (all residues)	3.3	4.0
RMSD backbone (ordered residues)[a]	0.6	0.8
RMS bond angles (°)	1.8	1.9
RMS bond lengths (Å)	0.012	0.017
Restraint statistics (RMSD of violations)		
NOE restraints	0.060 ± 0.003	0.179 ± 0.008
Hydrogen bonds	0.075 ± 0.015	0.100 ± 0.017
Dihedral restraints	1.805 ± 0.075	1.827 ± 0.318
RDC	0.748 ± 0.138	0.716 ± 0.256
Q value	0.146 ± 0.028	0.150 ± 0.054

[a] PROCHECK ordered residues are 37–53, 61–66, and 72–85 for $ComGC_{SS}$, and 36–54, 60–65, and 71–82 for $ComGC_{SP}$.

lent because 98.1% of the backbone and 90% of assignable protons overall could be assigned. Structural ensembles were determined with 880 NOE-based restraints, 54 hydrogen bonds, 102 dihedral angles restraints, and 38 RDC (Table 1). As can be seen in Fig. 5, our $ComGC_{SP}$ 3D structure is highly similar to the structure of $ComGC_{SS}$. It is, however, very different (RMSD of 3.6 Å) from the $ComGC_{SP}$ solution structure that was recently determined from a low number of restraints (Fig. S2) (19). In brief, our structure shows that $ComGC_{SP}$ displays three distinct helices, with α2-helix (residues 60–66) and α3-helix (residues 71–82) stacking against each other and packing orthogonal to the N-terminal α1-helix (Fig. 5*A*). As for $ComGC_{SS}$, except for the unstructured N terminus, there is no significant flexibility in $ComGC_{SP}$ because the structures within the NMR ensemble superpose well onto each other, with a RMSD of 1.6 Å for Cα atoms (Fig. 5*B*). Our $ComGC_{SS}$ and $ComGC_{SP}$ averaged structures are highly similar (Fig. 5*C*), with 1.8 Å RMSD between their ordered regions and 1.5 Å RMSD for the helical regions, which is consistent with the high sequence identity between these two proteins.

As determined by GETAREA (32) with a probe radius of 1.4 Å, the average ratio of solvent exposure for the ordered portion of $ComGC_{SS}$ is 48.3%, relative to 6.7% for those residues determined to be on the interior. In our $ComGC_{SS}$ structure, conserved residues Val^{43}, Gln^{46}, Tyr^{50}, Leu^{64}, and Ile^{70} are deeply buried, with an average of only 6% solvent exposure, forming a critical portion of a hydrophobic core contributing to the globular fold of ComGC. (Fig. 6). In contrast, the conserved Gly^{68} is solvent exposed, which is important for the formation of the α2-helix-turn-α3-helix motif where a tiny residue at the beginning of the turn is necessary to provide the flexibility and lack of steric restrictions required for turning. These observations also apply to our $ComGC_{SP}$ structure and are surprisingly reflected in the conservation of multiple chemical shifts between the conserved residues in our two structures (Fig. S3). In addition, modeling of the globular head of $ComGC_{BS}$ (Fig. S4), which predicts a globular fold similar to $ComGC_{SS}$ and $ComGC_{SP}$, shows that Cys^{36} and Cys^{76} are in close enough proximity to form a disulfide bond. Such disulfide bond, which is absent in $ComGC_{SS}$ and $ComGC_{SP}$ that do not have Cys residues, is expected to stabilize the globular fold and was reported to stabilize ComGC in *B. subtilis* (16, 33).

Because the hydrophobic α1N that has been truncated in His_6-$ComGC_{SS}$ is highly similar to the corresponding portion of several other bacterial T4F major pilins, including the PilE major pilin from *Neisseria gonorrheae* (Fig. S5) for which a full-length crystal structure is available (34), we could model the structure of the portion of α1 truncated in our construct to produce a full-length model of $ComGC_{SS}$ (Fig. 7). Comparison with the two different pilin folds identified so far, pilins from *N. gonorrheae* and *Geobacter sulfurreducens* have been chosen as representative models, clearly shows that ComGC adopts a radically different type IV pilin fold (Fig. 7). All three pilins have in common an extended N-terminal α1-helix, the universal defining structural feature of type IV pilins (3). In addition, whereas the very short *G. sulfurreducens* pilin almost exclusively consists of α1, both ComGC and PilE display a typical lollipop shape with a globular head mounted onto a "stick" (the α1-helix). However, unlike in canonical pilins where the globular head consists of a 4–7–stranded antiparallel β-sheet in a parallel plane to α1, oriented 45° or more relative to the long axis of α1 (3), in ComGC the structural backbone of the globular head is an helix-turn-helix roughly orthogonal to α1 (Fig. 7). This fold, which falls within the class of mainly α and the architecture of orthogonal bundles, represents a novel pilin fold. Taken together, these structural findings show that ComGC orthologs display conserved 3D structures, with a previously unreported type IV pilin fold.

ComGC novel pilin fold is compatible with helical T4F assembly

Because ComGC represents a novel type IV pilin structural fold, it was important to determine whether it could be modeled into recent cryo-EM structures obtained for a variety of bacterial T4F (7, 8, 35). These similar structures, *i.e.* filaments are right-handed helical polymers where pilins are held together by interactions between their α1 helices within the filament core, have revealed that a segment of α1N is melted during filament assembly, centered on helix-breaking residue Pro^{22}. That portion of α1 is highly conserved in ComGC, including the helix-breaking Pro^{22} (Fig. S5). Using SWISS-MODEL (36) and the cryo-EM structure of *N. gonorrheae* T4P (8) as a template, we produced a full-length 3D structural model of $ComGC_{SS}$ with a melted α1N segment (Fig. 8*A*). Considering that ComGC defines a monophyletic group and is highly conserved, it is very likely that all ComGC orthologs will display a similar 3D structure. This notion was strengthened by producing structural models for a range of different species expressing more or less distant ComGC (21.3–65.6% sequence identity), which were used to generate the phylogeny tree in Fig. 3. As seen in Fig. S6, all the models display the same lollipop shape

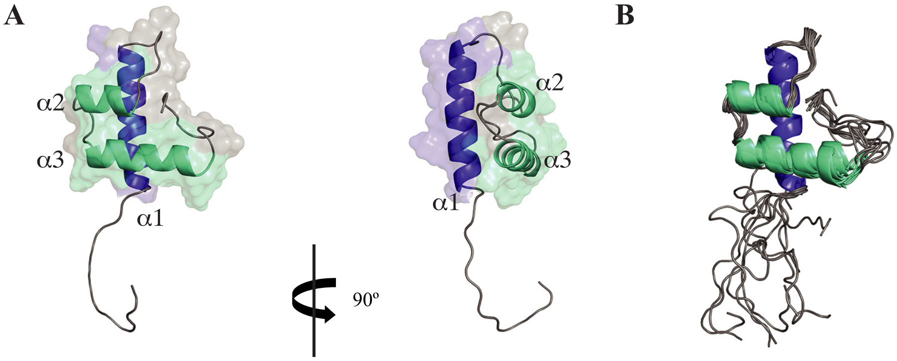

Figure 4. 3D solution structure of ComGC$_{SS}$. *A*, cartoon representation of the ComGC$_{SS}$ structure: face and side views are shown. A dimmed surface representation of the protein is superimposed. The three consecutive α-helices have been named α1, α2, and α3, and highlighted in *blue* (α1) or *cyan* (α2 and α3). *B*, cartoon representation of the superposition of the ensemble of 10 ComGC$_{SS}$ structures determined by NMR, which highlights that there is no significant flexibility in the structure except for the unstructured N terminus.

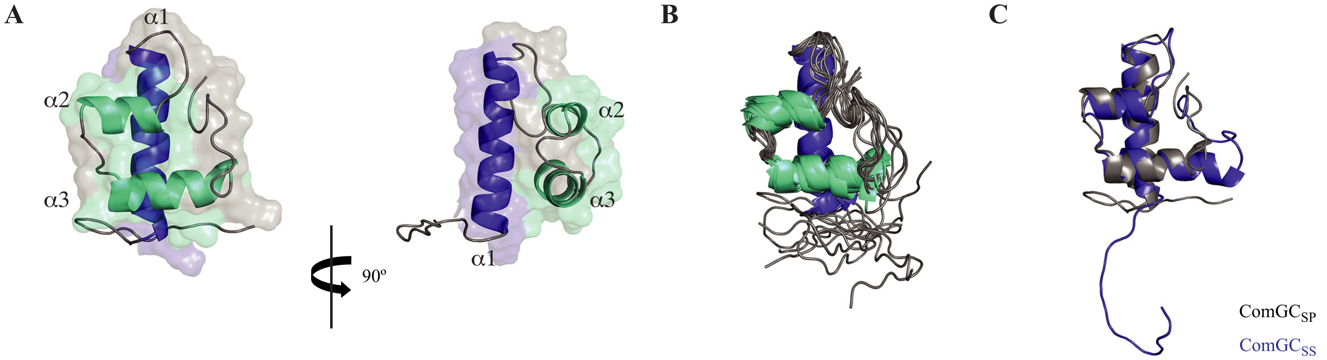

Figure 5. 3D solution structure of ComGC$_{SP}$. *A*, cartoon representation of the ComGC$_{SP}$ structure: face and side views are shown. A dimmed surface representation of the protein is superimposed. Nomenclature and color scheme are the same as described in the legend to Fig. 4. *B*, cartoon representation of the superposition of the ensemble of 10 ComGC$_{SP}$ structures determined by NMR. *C*, cartoon representation of the overlay of ComGC$_{SP}$ and ComGC$_{SS}$ representative structures. This highlights the high structural similarity between the two proteins, with 1.5 Å RMSD for the helical regions.

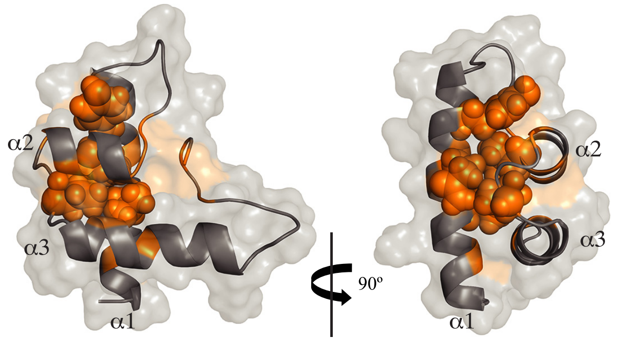

Figure 6. Conserved residues contributing to the globular fold of ComGC. Cartoon representation of the ordered portion of ComGC$_{SS}$, where residues determined to be on the interior using GETAREA, with surface accessibility ratios of less than 20%, are highlighted in *orange*. The consensus residues Val^{43}, Gln^{46}, Tyr^{50}, Leu^{64}, and Ile^{70} are shown with space filling representation.

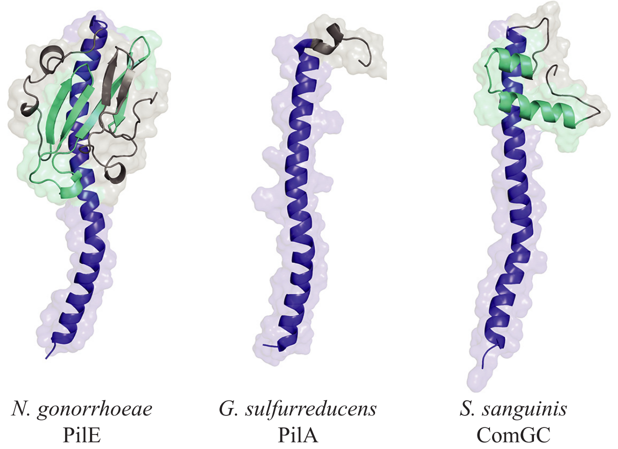

Figure 7. ComGC display a novel type IV pilin fold. 3D structure of the three different structural types of type IV pilins identified so far. The canonical type IV pilin fold is represented by the major pilin of T4aP in *N. gonorrheae* (PDB 2PIL). *G. sulfurreducens* T4P pilin (PDB 2M7G) is the chosen representative of the very short pilins almost exclusively consisting of α1. The full-length 3D structure of ComGC$_{SS}$ has been modeled. The conserved α1 is highlighted in *blue*. Distinctive structural features in the globular heads of PilE (antiparallel β-sheet) and ComGC (antiparallel α2-α3 orthogonal to α1) have been highlighted in *cyan*.

with a globular head mounted onto a α1 stick. As for ComGC$_{SS}$ and ComGC$_{SP}$, the structural backbone of the globular head is always a helix-turn-helix roughly orthogonal to α1.

We next assessed whether full-length ComGC would be compatible with helical T4F assembly and found that to be the case. Despite its novel pilin fold, we were able to model packing of ComGC within the cryo-EM structure of *N. gonorrheae* T4P (8). This produced a homology model with good Ramachandran plot statistics based on PROCHECK (37), *i.e.* allowed (89.5%), additional allowed (8.2%), generously allowed (2.3%), and disallowed (0%). As can be seen in Fig. 8*B*, the model revealed a right-handed helical packing of the conserved N-terminal α1-helices of ComGC$_{SS}$ in the filament core, which run approximately parallel to each other

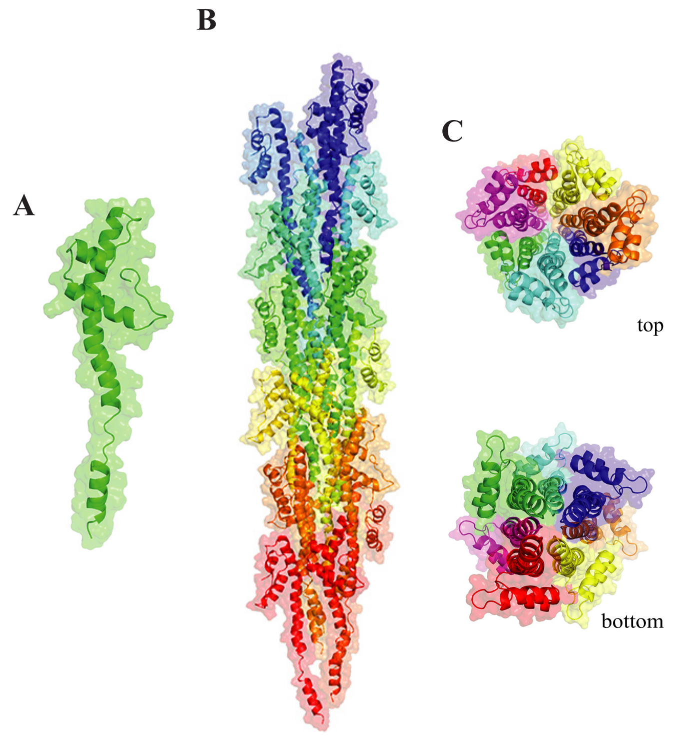

Figure 8. 3D model of ComGC filaments. The cryo-EM structure of the *N. gonorrheae* T4P (PDB 5VXX) has been used as a template to generate a model of ComGC$_{SS}$ pili. *A*, full-length ComGC$_{SS}$ in filaments with a melted segment in α1N. *B*, ComGC$_{SS}$ pili with a right-handed helical packing of the conserved α1-helices, which run approximately parallel to each other in the filament core. *C*, top and bottom views of ComGC$_{SS}$ pili highlighting the globular heads forming the outer surface of the filaments and the extensive interactions between α1-helices in the filament core.

and establish extensive hydrophobic interactions (Fig. 8*C*). In addition, the Glu^5 side chain of subunit S establishes a salt bridge and a hydrogen bond Phe^1 and Thr^2, respectively, of S+1. Importantly, the globular heads are stacked on top of each other along the long axis of the filaments and their helix-turn-helix structural backbone forms the outer surface of the filaments (Fig. 8*C*). A very similar model was obtained using *P. aeruginosa* T4P as a template (8).

Discussion

Their virtual ubiquity in prokaryotes and role in a variety of key biological processes make T4F an important research topic (1, 2). Com pili are involved in DNA uptake in naturally competent monoderm bacteria (9). Imported DNA, which usually leads to genome diversification via transformation, can also be used as a source of food or as a template for repair of damaged genomic DNA (10). Compared with T4F in diderms, most notably T4P and T2SS that have been extensively studied, Com pili have been understudied, including from a structural point of view. In this report, we focused on the major subunit of Com pili, the ComGC pilin, which we analyzed genomically, phylogenetically, and structurally. This led to the notable findings discussed below.

Although Com pili have been primarily studied in two model competent species (*B. subtilis* and *S. pneumoniae*), the present study suggests that they are widespread because complete sets of Com pilus-encoding genes are readily detected in more than 2,300 genomes corresponding to almost 350 different species. However, unlike promiscuous T4F such as T4aP and T4cP that are found in virtually all phyla of Bacteria (2), Com pili are restricted to a single phylum (Firmicutes) and almost exclusively to a single underlying class of monoderms (Bacilli), where they are virtually ubiquitous. Indeed, an overwhelming majority of Bacilli genomes (88%) have Com-encoding genes. Interestingly, the major subunit of Com pili (ComGC) shows extensive sequence conservation in the corresponding genomes and define a clear monophyletic group within type IV pilins. Taken together, these observations suggest that the Com pilus is a T4F that has emerged only once, very early during the diversification of Firmicutes, where it has remained largely confined ever since. Because the Com-encoding genes have not become pseudogenes, it is likely that most Bacilli have the ability to assemble a Com pilus and take up DNA. However, because only a handful of these species have been experimentally shown to be competent (10), this implies that either the imported DNA is primarily used as food or for genome repair instead of genome diversification, or that the inducing cues leading to transformation are yet to be established for most species of Firmicutes. Alternatively, Com pili might have evolved in some of these species to take up other macromolecules, which is, however, at odds with the conservation of the five pilins.

Perhaps the most important finding in this study is that ComGC, the major subunit of the Com pilus, displays an entirely novel major pilin fold where the extended N-terminal α1-helix, the universal defining structural feature of type IV pilins (3), is topped by a purely helical globular head. ComGC thus appears to be a "middle ground" between longer canonical pilins (*e.g. N. gonorrheae*), in which the globular head consists of an antiparallel β-sheet, and the very short pilins where a globular head is missing altogether (*e.g. G. sulfurreducens*). These structures point to a hypothetical evolutionary scenario during which truncation of the antiparallel β-sheet in a canonical type IV pilin might have led to a purely helical ComGC proto-structure. Intriguingly, this scenario "works" particularly well with PilE1, a major subunit of *S. sanguinis* T4P, which has two short α-helices in the loop connecting α1 and the antiparallel β-sheet (11). Importantly, this putative "truncation" would not interfere with the expected ability of ComGC to be assembled into helical filaments, because this pilin could be readily modeled into recent T4F structures (7, 8, 35). Com pili are thus likely to result from the right-handed helical packing of ComGC α1-helices within the filament core, running parallel to each other and establishing extensive hydrophobic interactions, with a melted central portion. Such packing will stack the globular heads on top of each other, forming the surface of the filaments. Extensive sequence conservation, including for residues beyond the classically conserved α1N, and the fact that the two structures that we have solved are virtually identical, strongly suggest that these structural features apply to the whole ComGC clade, including species such as *B. subtilis* where extended filaments have not been observed (16). It is therefore surprising that a recently published NMR structure of ComGC$_{SP}$ (PDB 5NCA) (19) differs dramatically from ours. Although the previous structure is purely helical as well, the orientation of the α2 and α3 helices is entirely different, resulting in an absence of packing of the conserved hydrophobic core. Therefore, PDB 5NCA, which resembles a one-sided "pick-axe" with no globular head, cannot be readily modeled into recent T4F structures (Fig. S7). Interestingly, our assignments vary only slightly from those previously produced for PDB 5NCA (Fig. S8). However, whereas we have managed to successfully assign 90% assignable protons overall, the previous assignment was merely 65% (19), which probably accounts for the apparently "unfolded" state of PDB 5NCA. Indeed, without a high degree of proton identification, the assignment of NOESY peaks and production of distance restraints fails. Local hydrogen bonds and dihedral restraints often cannot compensate for lack of long-range NOEs within the protein interior or between elements of secondary structure.

Together with these conserved structural features, the conservation *en bloc* of the genes encoding the Com pilus strongly suggests that the molecular mechanisms of filament assembly and DNA uptake are widely conserved in Firmicutes. These mechanisms, which remain poorly understood, can be advantageously studied in *S. sanguinis*, which has recently emerged as a monoderm T4P model (20). Actually, *S. sanguinis* is so far the only monoderm expressing two distinct T4F, Com pili and retractable T4aP, which further cements it as a prime model species. Comparison with other T4F systems shows that the machinery involved in biogenesis of Com pili is one of the simplest, by far. Because ComGD, ComGE, ComGF, and ComGG pilins are likely to be minor pilus components important for filament stability and function (a conserved role for minor pilins in various T4F) (1), and ComC is the prepilin peptidase processing pilins (38), it appears that assembly of ComGC into filaments is mediated by two proteins only. Namely, an extension ATPase (ComGA) and a platform protein (ComGB), which together will assemble processed ComGC into a right-handed helical filament. Upon DNA binding, which has been visualized for *S. pneumoniae* Com pili, but the receptor is yet to be identified (15), uptake will be initiated by filament retraction (14). Because there is no dedicated retraction ATPase, one possibility is that ComGA might be a bifunctional motor powering both extension and retraction like recently suggested for the T4cP motor (39). It would be interesting to image Com filaments dynamics and DNA-binding ability in live cells, using a labeling strategy that has recently enabled the visualization of these steps for T4aP involved in competence in naturally competent diderm species (14).

In conclusion, by providing high-resolution structural information for the ComGC pilins, this study has shed light on an understudied T4F involved in DNA uptake found in hundreds of monoderm bacterial species and has led to the surprising discovery of a novel type IV pilin fold. This paves the way for further investigations of this minimalist T4F, which are expected to improve our understanding of a fascinating superfamily of filamentous nanomachines ubiquitous in prokaryotes.

Experimental procedures

Bioinformatic analyses

Protein sequences were routinely analyzed using the DNA Strider program. Protein sequence alignments were done using the Clustal Omega server at EMBL-EBI. Pretty-printing of alignment files was done using the BoxShade server at ExPASy. Reformatting of large multiple alignment files was done using the MView server at EMBL-EBI. Prediction of functional domains was done using the InterProScan server at EMBL-EBI, which was also used to download all the ComGC protein entries with an IPR016940 domain. Protein secondary structure prediction was done using the JPred server at the University of Dundee. Protein 3D structures were downloaded from the RCSB PDB server. Molecular visualization of protein 3D structures was done using PyMOL (Schrödinger). The GETAREA server, at UTMB, was used for calculating the solvent accessible surface area of ComGC proteins.

Detection of the Com systems in genomes available in the NCBI RefSeq database (last accessed in April 2019, 13,512 genomes of Bacteria and Archaea) was done as described previously (2), using MacSyFinder (25) and the relevant HMM Com model (2). Phylogenetic analysis based on protein sequences of major pilins of different T4F involved an initial alignment of the sequences using MAFFT version 7.273 (40), specifically the linsi algorithm. Multiple alignments were analyzed using Noisy version 1.5.12 (41) with default parameters, to select the informative sites. Next, we inferred maximum likelihood trees from the curated alignments using IQ-TREE version 1.6.7.2 (26), with option -allnni. We evaluated the node supports using the options -bb 1,000 for ultra-fast bootstraps, and -alrt 1,000 for SH-aLRT (27). The best evolutionary model was selected with ModelFinder (42), option -MF and BIC criterion. We used the option -wbtl to conserve all optimal trees and their branches length.

Protein expression and purification

A synthetic gene, codon-optimized for *E. coli* expression, encoding ComGC$_{SS}$ from *S. sanguinis* 2908 (21) was synthesized and cloned by GeneArt, yielding pMA-T-*comGC*$_{SS}$ (Table S2). The portion of the gene encoding residues 23–94 from the mature protein was PCR-amplified using *comGC*$_{SS}$-F and *comGC*$_{SS}$-R primers (Table S3), cut with NcoI and BamHI, and cloned into the pET28b vector (Novagen) cut with the same enzymes. The forward primer was designed to fuse a noncleavable N-terminal His$_6$ tag to ComGC$_{SS}$. The resulting plasmid was verified by sequencing and transformed into chemically competent *E. coli* BL21(DE3) cells. A single colony was transferred to 10 ml of LB supplemented with 50 μg ml^{-1} of kanamycin and grown at 37 °C overnight. This pre-culture was back-diluted 100-fold into 1 liter of M9 minimal medium, supplemented with antibiotic, a mixture of vitamins, and trace elements, and D-[^{13}C]glucose and [^{15}N]NH_4Cl for isotopic labeling. Cells were grown in an orbital shaker at 37 °C until the OD$_{600}$ reached 0.7, before adding 0.4 mM IPTG (Merck Chemicals) to induce protein expression during 16 h at 18 °C. Cells were then harvested by centrifugation at 8,000 × *g* for 20 min and subjected to one freeze/thaw cycle in lysis buffer (PBS, pH 7.4, with EDTA-free protease inhibitors). This lysate was further disrupted by repeated cycles of sonication, pulses of 5 s on and 5 s off during 5 min, until the cell suspension was visibly less viscous. The cell lysate was then centrifuged for 20 min at 18,000 × *g* to remove cell debris. The clarified lysate was then passed using an ÄKTA Purifier FPLC through a 1-ml HisTrap HP column (GE Healthcare), pre-equilibrated in lysis buffer. The column was then washed extensively with lysis buffer to remove unbound material before His_6-$ComGC_{SS}$ was eluted using elution buffer (PBS, pH 7.4, 200 mM NaCl, 300 mM imidazole). Affinity-purified $ComGC_{SS}$ was further purified by gel-filtration chromatography on an HiLoad 16/600 Superdex 75 column (GE Healthcare), using (25 mM Na_2HPO_4/NaH_2PO_4, pH 6, 200 mM NaCl) buffer for elution. For RDC measurements we produced ^{15}N-labeled protein as follows. Bacteria grown overnight in 5 ml of LB with antibiotic were subcultured at 37 °C in 0.8 liters of LB to 0.6 OD_{600}, and then transferred to 0.4 liters of M9 with $[^{15}N]NH_4Cl$, unlabeled D-glucose, and 10 μg liter^{-1} of thiamine. Cultures were induced with 0.3 mM IPTG at 16 °C for 18 h. After the production of a clarified lysate, protein was purified as above, except for the use of hand-made Ni-NTA-agarose (Qiagen) in 50 mM Tris, pH 8, 300 mM NaCl and eluted using 50 mM Tris, pH 8, 200 mM NaCl, 300 mM imidazole, and Superdex 75 10/300 GL (GE Healthcare) columns in 25 mM Tris, pH 8, 200 mM NaCl, and dialyzed into 25 mM Na_2HPO_4/NaH_2PO_4, pH 6, 50 mM NaCl.

For $ComGC_{SP}$, a codon-optimized synthetic gene based on the gene from *S. pneumoniae* R6 was synthesized and cloned by GeneArt, yielding pMA-T-*$comGC_{SP}$* (Table S2). The portion of the gene encoding residues 23–93 from the mature protein was PCR-amplified using *$comGC_{SP}$*-F and *$comGC_{SP}$*-R primers (Table S3), cut with NcoI and BamHI, and cloned into the pET28b vector (Novagen) cut with the same enzymes. The forward primer was designed to fuse a noncleavable N-terminal His_6 tag to $ComGC_{SS}$. The resulting plasmid was verified by sequencing and transformed into chemically competent *E. coli* BL21(DE3) cells. A single colony was transferred to 5 ml of LB supplemented with 50 μg ml^{-1} of kanamycin and grown overnight at 37 °C. Bacteria were subcultured at 37 °C in 0.8 liters of LB with antibiotic to OD_{600} 0.7, and then transferred into 0.4 liters of M9 with 10 μg liter^{-1} of thiamine, and either $[^{15}N]NH_4Cl$ and unlabeled D-glucose, or $[^{15}N]NH_4Cl$ and D-$[^{13}C]$glucose. Cultures were induced with 0.3 mM IPTG at 16 °C for 18 h. After the production of a clarified lysate, $ComGC_{SS}$ was purified as above using hand-made Ni-NTA-agarose (Qiagen) and Superdex 75 10/300 GL (GE Healthcare) columns.

NMR spectroscopy and structure determination

All data were collected on Bruker Avance III HD 800 MHz and 600 MHz triple resonance spectrometers with cryoprobes operated at 25 °C. For $ComGC_{SS}$, a sample containing $^{13}C/^{15}N$-labeled protein at 1 mM in NMR buffer (25 mM Na_2HPO_4/NaH_2PO_4, pH 6, 50 mM NaCl, 5% D_2O) was used for assignment experiments and structure determination. For $ComGC_{SP}$, a sample containing $^{13}C/^{15}N$-labeled protein at 1.8 mM in NMR buffer was used for assignment experiments and structure determination. Resonance assignments for $ComGC_{SS}$ were performed using ^{15}N HSQC, ^{13}C aliphatic HSQC, HNCACB, CBCACONH, HBHA, HNCO, HNCACO, HCCCONH, CCCONH, and CCH. For $ComGC_{SP}$, assignments were performed using ^{15}N HSQC, ^{13}C aliphatic HSQC, HNCA, CBCANH, CBCACONH, HBHA, HNCO, HNCACO, HCCCONH, CCCONH, and CCH. All data were processed using MddNMR (43) for reconstruction after Non-Uniform Sampling and NMRPipe (44). Peak picking and assignments were performed in SPARKY (45).

NOE peak lists were used, with mixing time of 140 ms, from 3D ^{13}C-HSQC-NOESY, 3D ^{15}N-HSQC-NOESY for $ComGC_{SP}$, and simultaneous $^{13}C/^{15}N$ chemical shift evolution NOESY for $ComGC_{SS}$. For both proteins, RDC lists were derived from ^{15}N-HSQC-IPAP experiments on ^{15}N-labeled isotropic and aligned sample in 3% PEG/hexanol liquid crystal, with D_2O splitting of ~7 Hz. RDCs were included in the structure calculations if there was baseline resolution and for residues where TALOS+ predicted order parameter of >0.8. Angular constraints from TALOS+ were used in the structure calculations. Both $ComGC_{SS}$ and $ComGC_{SP}$ structures were determined using Ponderosa-C/S (45), refined using Xplor-NIH 2.52 (46), aligned using Theseus (47), and the secondary structure checked using Stride (48). Structure validation was performed using PSVS (49), PROCHECK (37), and in-house scripts.

Modeling

SWISS-MODEL server at ExPASy was used for modeling protein 3D structures. In brief, the full-length $ComGC_{SS}$ was modeled using *N. gonorrheae* major pilin PilE (PDB 2PIL) as a template (50). We first modeled the missing α1 residues in our structure, which was aligned to our Xplor-NIH–produced average NMR structure (without the first unstructured α1 residues) using PyMOL and finally merged using Coot (51).

Similarly, the full-length $ComGC_{SS}$ structure within filaments was modeled by using one of the PilE subunits from the cryo-EM model of *N. gonorrheae* T4P (PDB 5VXX) (8) as a template for the missing α1 residues in our structure. The Com pilus model was produced after alignment of the averaged NMR structure $ComGC_{SS}$ α1-helices to the α1-helices of SWISS-MODEL PilE-based homology model subunits in the *N. gonorrheae* T4P. This was also done for the recently published $ComGC_{SP}$ structure (PDB 5NCA). The structural elements were fused using Coot (51). In addition, we modeled packing of full-length $ComGC_{SS}$ in the PAK pilus from *P. aeruginosa* (PDB 5VXY) (8).

Data availability

The NMR solution structures of $ComGC_{SS}$ and $ComGC_{SP}$ have been deposited in the Protein Data Bank under accession numbers 6TXT and 6Y1H, respectively. Chemical shift assignments and NOE-based restraints used in structure calculations are available from the Biological Magnetic Resonance Data Bank under accession numbers 34477 and 34490, respectively. All the other data described in the manuscript are either contained within the manuscript, or are to be shared upon request to corresponding author.

Author contributions—D. S. and V. P. data curation; D. S., E. P. C. R., and V. P. formal analysis; D. S., E. P. C. R., S. M., and V. P. validation; D. S., J.-L. B., and R. D. investigation; D. S., J.-L. B., R. D., S. M., and V. P. methodology; D. S. and V. P. writing-original draft; R. D., E. P. C. R., and S. M. writing-review and editing; E. P. C. R. and V. P. supervision; E. P. C. R. and V. P. funding acquisition; S. M. and V. P. resources; V. P. conceptualization; V. P. project administration.

Acknowledgments—This work relied heavily on the use of the Cross-Faculty NMR Centre at Imperial College London. We are grateful to Nicolas Biais (City University of New York) and Romé Voulhoux (CNRS Marseille) for critical reading of the manuscript.

References

1. Berry, J. L., and Pelicic, V. (2015) Exceptionally widespread nano-machines composed of type IV pilins: the prokaryotic Swiss Army knives. *FEMS Microbiol. Rev.* **39,** 134–154 CrossRef Medline
2. Denise, R., Abby, S. S., and Rocha, E. P. C. (2019) Diversification of the type IV filament superfamily into machines for adhesion, protein secretion, DNA uptake, and motility. *PLos Biol.* **17,** e3000390 CrossRef Medline
3. Giltner, C. L., Nguyen, Y., and Burrows, L. L. (2012) Type IV pilin proteins: versatile molecular modules. *Microbiol. Mol. Biol. Rev.* **76,** 740–772 CrossRef Medline
4. Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., and Hunter, S. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30,** 1236–1240 CrossRef Medline
5. Reardon, P. N., and Mueller, K. T. (2013) Structure of the type IVa major pilin from the electrically conductive bacterial nanowires of *Geobacter sulfurreducens. J. Biol. Chem.* **288,** 29260–29266 CrossRef
6. LaPointe, C. F., and Taylor, R. K. (2000) The type 4 prepilin peptidases comprise a novel family of aspartic acid proteases. *J. Biol. Chem.* **275,** 1502–1510 CrossRef Medline
7. Kolappan, S., Coureuil, M., Yu, X., Nassif, X., Egelman, E. H., and Craig, L. (2016) Structure of the *Neisseria meningitidis* type IV pilus. *Nat. Commun.* **7,** 13015 CrossRef Medline
8. Wang, F., Coureuil, M., Osinski, T., Orlova, A., Altindal, T., Gesbert, G., Nassif, X., Egelman, E. H., and Craig, L. (2017) Cryoelectron microscopy reconstructions of the *Pseudomonas aeruginosa* and *Neisseria gonorrhoeae* type IV pili at sub-nanometer resolution. *Structure* **25,** 1423–1435.e4 CrossRef Medline
9. Dubnau, D., and Blokesch, M. (2019) Mechanisms of DNA uptake by naturally competent bacteria. *Annu. Rev. Genet.* **53,** 217–237 CrossRef Medline
10. Johnston, C., Martin, B., Fichant, G., Polard, P., and Claverys, J. P. (2014) Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat. Rev. Microbiol.* **12,** 181–196 CrossRef Medline
11. Berry, J. L., Gurung, I., Anonsen, J. H., Spielman, I., Harper, E., Hall, A. M. J., Goosens, V. J., Raynaud, C., Koomey, M., Biais, N., Matthews, S., and Pelicic, V. (2019) Global biochemical and structural analysis of the type IV pilus from the Gram-positive bacterium *Streptococcus sanguinis. J. Biol. Chem.* **294,** 6796–6808 CrossRef Medline
12. Merz, A. J., So, M., and Sheetz, M. P. (2000) Pilus retraction powers bacterial twitching motility. *Nature* **407,** 98–102 CrossRef Medline
13. Cehovin, A., Simpson, P. J., McDowell, M. A., Brown, D. R., Noschese, R., Pallett, M., Brady, J., Baldwin, G. S., Lea, S. M., Matthews, S. J., and Pelicic, V. (2013) Specific DNA recognition mediated by a type IV pilin. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 3065–3070 CrossRef Medline
14. Ellison, C. K., Dalia, T. N., Vidal Ceballos, A., Wang, J. C., Biais, N., Brun, Y. V., and Dalia, A. B. (2018) Retraction of DNA-bound type IV competence pili initiates DNA uptake during natural transformation in *Vibrio cholerae. Nat. Microbiol.* **3,** 773–780 CrossRef
15. Laurenceau, R., Pehau-Arnaudet, G., Baconnais, S., Gault, J., Malosse, C., Dujeancourt, A., Campo, N., Chamot-Rooke, J., Le Cam, E., Claverys, J. P., and Fronzes, R. (2013) A type IV pilus mediates DNA binding during natural transformation in *Streptococcus pneumoniae. PLoS Pathog.* **9,** e1003473 CrossRef
16. Chen, I., Provvedi, R., and Dubnau, D. (2006) A macromolecular complex formed by a pilin-like protein in competent *Bacillus subtilis. J. Biol. Chem.* **281,** 21720–21727 CrossRef Medline
17. Chung, Y. S., and Dubnau, D. (1995) ComC is required for the processing and translocation of ComGC, a pilin-like competence protein of *Bacillus subtilis. Mol. Microbiol.* **15,** 543–551 CrossRef
18. Chung, Y. S., and Dubnau, D. (1998) All seven *comG* open reading frames are required for DNA binding during transformation of competent *Bacillus subtilis. J. Bacteriol.* **180,** 41–45 CrossRef Medline
19. Muschiol, S., Erlendsson, S., Aschtgen, M. S., Oliveira, V., Schmieder, P., de Lichtenberg, C., Teilum, K., Boesen, T., Akbey, U., and Henriques-Normark, B. (2017) Structure of the competence pilus major pilin ComGC in *Streptococcus pneumoniae. J. Biol. Chem.* **292,** 14134–14146 Medline
20. Pelicic, V. (2019) Monoderm bacteria: the new frontier for type IV pilus biology. *Mol. Microbiol.* **112,** 1674–1683 CrossRef Medline
21. Gurung, I., Spielman, I., Davies, M. R., Lala, R., Gaustad, P., Biais, N., and Pelicic, V. (2016) Functional analysis of an unusual type IV pilus in the Gram-positive *Streptococcus sanguinis. Mol. Microbiol.* **99,** 380–392 CrossRef Medline
22. Albano, M., Breitling, R., and Dubnau, D. A. (1989) Nucleotide sequence and genetic organization of the *Bacillus subtilis comG* operon. *J. Bacteriol.* **171,** 5386–5404 CrossRef Medline
23. Mohan, S., Aghion, J., Guillen, N., and Dubnau, D. (1989) Molecular cloning and characterization of *comC*, a late competence gene of *Bacillus subtilis. J. Bacteriol.* **171,** 6043–6051 CrossRef Medline
24. Imam, S., Chen, Z., Roos, D. S., and Pohlschröder, M. (2011) Identification of surprisingly diverse type IV pili, across a broad range of Gram-positive bacteria. *PLoS ONE* **6,** e28919 CrossRef Medline
25. Abby, S. S., Néron, B., Ménager, H., Touchon, M., and Rocha, E. P. (2014) MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS ONE* **9,** e110726 CrossRef Medline
26. Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32,** 268–274 CrossRef Medline
27. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018) UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35,** 518–522 CrossRef Medline
28. Craig, L., Taylor, R. K., Pique, M. E., Adair, B. D., Arvai, A. S., Singh, M., Lloyd, S. J., Shin, D. S., Getzoff, E. D., Yeager, M., Forest, K. T., and Tainer, J. A. (2003) Type IV pilin structure and assembly. X-ray and EM analyses of *Vibrio cholerae* toxin-coregulated pilus and *Pseudomonas aeruginosa* PAK pilin. *Mol. Cell* **11,** 1139–1150 CrossRef Medline
29. Drozdetskiy, A., Cole, C., Procter, J., and Barton, G. J. (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43,** W389–394 CrossRef Medline
30. Krissinel, E., and Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60,** 2256–2268 CrossRef
31. Shen, Y., Delaglio, F., Cornilescu, G., and Bax, A. (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR* **44,** 213–223 CrossRef Medline
32. Fraczkiewicz, R., and Braun, W. (1998) Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comput. Chem.* **19,** 319–333 CrossRef
33. Meima, R., Eschevins, C., Fillinger, S., Bolhuis, A., Hamoen, L. W., Dorenbos, R., Quax, W. J., van Dijl, J. M., Provvedi, R., Chen, I., Dubnau, D., and Bron, S. (2002) The *bdbDC* operon of *Bacillus subtilis* encodes thiol-disulfide oxidoreductases required for competence development. *J. Biol. Chem.* **277,** 6994–7001 CrossRef Medline
34. Parge, H. E., Forest, K. T., Hickey, M. J., Christensen, D. A., Getzoff, E. D., and Tainer, J. A. (1995) Structure of the fibre-forming protein pilin at 2.6 Å resolution. *Nature* **378,** 32–38 CrossRef Medline
35. López-Castilla, A., Thomassin, J. L., Bardiaux, B., Zheng, W., Nivaskumar, M., Yu, X., Nilges, M., Egelman, E. H., Izadi-Pruneyre, N., and Francetic, O. (2017) Structure of the calcium-dependent type 2 secretion pseudopilus. *Nat. Microbiol.* **2,** 1686–1695 CrossRef Medline
36. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., and Schwede, T. (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46,** W296–W303 CrossRef Medline

37. Laskowski, R. A., MacArthur, M. W., Moss D. S, Thornton, J. M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26,** 283–291 CrossRef
38. Chung, Y. S., Breidt, F., and Dubnau, D. (1998) Cell surface localization and processing of the ComG proteins, required for DNA binding during transformation of *Bacillus subtilis. Mol. Microbiol.* **29,** 905–913 CrossRef
39. Ellison, C. K., Kan, J., Chlebek, J. L., Hummels, K. R., Panis, G., Viollier, P. H., Biais, N., Dalia, A. B., and Brun, Y. V. (2019) A bifunctional ATPase drives tad pilus extension and retraction. *Sci. Adv.* **5,** eaay2591 CrossRef Medline
40. Katoh, K., and Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 CrossRef Medline
41. Dress, A. W., Flamm, C., Fritzsch, G., Grunewald, S., Kruspe, M., Prohaska, S. J., and Stadler, P. F. (2008) Noisy: identification of problematic columns in multiple sequence alignments. *Algorithm. Mol. Biol.* **3,** 7 CrossRef
42. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14,** 587–589 CrossRef Medline
43. Orekhov, V. Y., and Jaravine, V. A. (2011) Analysis of non-uniformly sampled spectra with multi-dimensional decomposition. *Prog. Nucl. Mag. Res. Spect.* **59,** 271–292 CrossRef
44. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6,** 277–293 Medline
45. Lee, W., Tonelli, M., and Markley, J. L. (2015) NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **31,** 1325–1327 CrossRef Medline
46. Schwieters, C. D., Kuszewski, J. J., and Clore, G. M. (2006) Using Xplor-NIH for NMR molecular structure determination. *Prog. Nucl. Mag. Reson. Spect.* **48,** 47–62 CrossRef
47. Theobald, D. L., and Wuttke, D. S. (2008) Accurate structural correlations from maximum likelihood superpositions. *PLoS Comput. Biol.* **4,** e43 CrossRef Medline
48. Frishman, D., and Argos, P. (1995) Knowledge-based protein secondary structure assignment. *Proteins* **23,** 566–579 CrossRef Medline
49. Bhattacharya, A., Tejero, R., and Montelione, G. T. (2007) Evaluating protein structures determined by structural genomics consortia. *Proteins* **66,** 778–795 Medline
50. Forest, K. T., Dunham, S. A., Koomey, M., and Tainer, J. A. (1999) Crystallographic structure reveals phosphorylated pilin from *Neisseria*: phosphoserine sites modify type IV pilus surface chemistry and fibre morphology. *Mol. Microbiol.* **31,** 743–752 CrossRef Medline
51. Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010) Features and development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66,** 486–501 CrossRef