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

Devon Sheppard[1], Jamie-Lee Berry[1], Rémi Denise[2,3], Eduardo P. C. Rocha[2], Stephen J. Matthews[4] and Vladimir Pelicic[1,*]

[1]MRC Centre for Molecular Bacteriology and Infection, Imperial College London, London SW7 2AZ, United Kingdom

[2]Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525, Paris 75015, France

[3]Sorbonne Université, Collège doctoral, Paris 75005, France

[4]Centre for Structural Biology, Imperial College London, London SW7 2AZ, United Kingdom

*Correspondence: v.pelicic@imperial.ac.uk

ABSTRACT

Type IV filaments (Tff) - a superfamily of filamentous nanomachines virtually ubiquitous in prokaryotes - are helical assemblies of type IV pilins that mediate a wide variety of functions. The competence (Com) (pseudo)pilus is a widespread Tff mediating DNA uptake in bacteria with one membrane (monoderms), which is the first step in natural transformation, an important mechanism of horizontal gene transfer. Here, we report the genomic, phylogenetic and structural analysis of ComGC, the major pilin subunit of Com (pseudo)pili. By performing a global comparative analysis, we show that Com (pseudo)pili are virtually ubiquitous in Bacilli, a major monoderm class of Firmicutes, and ComGC displays extensive sequence conservation, defining a monophyletic group among type IV pilins. We also report two 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, which is purely helical. Modelling of ComGC packing into Tff confirms that its unusual structure is compatible with helical filament assembly. Owing to the widespread distribution of ComGC, these results have implications for hundreds of monoderm species.

Keywords: natural transformation, type IV pilin, type IV filaments, DNA uptake

INTRODUCTION

Filamentous nanomachines composed of type IV pilins are virtually ubiquitous in Bacteria and Archaea (Berry and Pelicic, 2015), to which they confer a variety of unrelated functions including adhesion, motility, protein secretion, DNA uptake. These type IV filaments (Tff) are assembled by conserved multi-protein machineries, which further underlines their phylogenetic relationship (Denise et al., 2019).

Much of our current understanding of this superfamily of nanomachines comes from the study of type IV pili (Tfp), the best characterised Tff (Berry and Pelicic, 2015). In brief, Tfp are µm-long and thin surface-exposed filaments, which are polymers of usually one major type IV pilin. Type IV pilins (simply named pilins hereafter) are defined by an N-terminal sequence motif known as class III signal peptide (Giltner et al., 2012). This motif - IPR012902 entry in the InterPro database (Jones et al., 2014) - 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 (Giltner et al., 2012). Although some small pilins consist solely of this extended α-helix (Reardon and Mueller, 2013), 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 (Giltner et al., 2012). 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) (LaPointe and Taylor, 2000). Processing primes pilins for polymerisation into filaments. Filament assembly, which remains poorly understood, is mediated by a multi-protein machinery in the CM, centred on an integral membrane platform protein (IPR003004) and a cytoplasmic extension ATPase (IPR007831) (Berry and Pelicic, 2015). As revealed by recent

67 cryo-EM structures of several Tfp (Kolappan et al., 2016; Wang et al., 2017),
68 filaments are right-handed helical polymers where pilins are held together by
69 extensive interactions between their α1 helices, which are partially melted and run
70 approximately parallel to each other within the filament core.

71 One of the key functional roles of Tff is their involvement in natural
72 transformation in prokaryotes, the ability of species defined as "competent" to take up
73 exogenous DNA across their membrane(s) and incorporate it stably into their
74 genomes (Dubnau and Blokesch, 2019). This widespread property in bacteria
75 (Johnston et al., 2014) is key for horizontal gene transfer, an important factor in
76 bacterial evolution and the spread of antibiotic resistance. Tff are involved in the very
77 first step of natural transformation, *i.e.* binding of free extracellular DNA and its
78 translocation close to the cytoplasmic membrane (Dubnau and Blokesch, 2019).
79 DNA is subsequently bound by the DNA receptor ComEA and further translocated
80 across the CM through the ComEC channel (Dubnau and Blokesch, 2019). In diderm
81 competent species, the Tff involved in DNA uptake is a sub-type of Tfp, known as
82 Tfpa (Berry et al., 2019), which rapid depolymerisation is powered by the retraction
83 ATPase PilT (IPR006321), generating exceptionally large tensile forces (Merz et al.,
84 2000). In brief, Tfpa bind DNA directly, via one of their major or minor (low
85 abundance) pilin subunits (Cehovin et al., 2013), and then are retracted by PilT,
86 bringing DNA to the ComEA receptor (Ellison et al., 2018). In monoderm competent
87 species, DNA uptake is mediated by a distinct Tff named competence (Com)
88 (pseudo)pilus (Dubnau and Blokesch, 2019), much less well characterised than Tfp.
89 Com (pseudo)pili are composed mainly of the major pilin (ComGC) (Chen et al.,
90 2006; Laurenceau et al., 2013), and are assembled by a simple machinery
91 composed of four minor pilins (ComGD, ComGE, ComGF, ComGG), a prepilin
92 peptidase (ComC), an extension ATPase (ComGA) and a platform protein (ComGB)
93 (Chung and Dubnau, 1995, 1998). Com filaments are elusive Tff, since µm-long
94 filaments have been visualised only in *S. pneumoniae* so far (Laurenceau et al.,

95 2013; Muschiol et al., 2017), while filaments are thought to be much shorter in the
96 other model competent species *Bacillus subtilis* (Chen et al., 2006). For this reason,
97 until this points is addressed in other competent species, we refer to Com as
98 (pseudo)pili in this study.

99 How Com (pseudo)pili are assembled, bind DNA and presumably retract in the
100 absence of a PilT retraction motor is not understood. One important limitation is the
101 absence of high-resolution structural information. Therefore, in the present study, we
102 have focused on ComGC, the major subunit of the Com (pseudo)pilus. We report (i)
103 a global comparative and phylogenetic analysis of ComGC, and (ii) 3D structures for
104 two orthologs, $ComGC_{SP}$ from the model competent species *S. pneumoniae* and
105 $ComGC_{SS}$ from *S. sanguinis*, a common cause of infective endocarditis in humans
106 that has recently emerged as a monoderm model for the study of Tff. Finally, we
107 discuss the general implications of these findings.

RESULTS

Com (pseudo)pili are almost ubiquitous in monoderm Bacilli, including the Tff model S. sanguinis

So far, Com (pseudo)pili have been mainly studied in two model competent species: *B. subtilis* and *S. pneumoniae*. *S. sanguinis* is a naturally competent species that has recently emerged as a monoderm Tff model since it expresses retractable Tfpa and Com (pseudo)pili (Pelicic, 2019). Functional analysis of *S. sanguinis* Tfpa showed that they are dispensable for DNA uptake, which is instead mediated by the Com (pseudo)pilus since competence was abolished in a *ΔcomGB* mutant (Gurung et al., 2016). A closer inspection of *S. sanguinis* genome revealed that all the genes encoding the Com (pseudo)pilus are present. These genes are organised in two loci (Fig. 1A), *comC* and the *comGABCDEDFG* operon, showing shared synteny with the corresponding loci in model competent species (Albano et al., 1989; Mohan et al., 1989). 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. 1B), *i.e.* with 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 non-canonical class III signal peptide (Fig. 1B), which is not identified by InterPro or PilFind that is dedicated to the prediction of type IV pilins (Imam et al., 2011). This is a conserved property for ComGG orthologs.

We next determined the global distribution of the Com system in publicly available complete bacterial genomes using MacSyFinder (Abby et al., 2014). Specifically, we used the MacSyFinder model built for the identification of Com systems (Denise et al., 2019), which takes into account the genetic composition and organisation of its components. This showed that the Com system is restricted to

136 Firmicutes, a phylum comprising a vast majority of monoderms, where it is
137 exceptionally widespread since it was detected in 2,333 genomes (Supplemental
138 Spreadsheet 1). An overwhelming majority of the corresponding species (99.7%)
139 belong to the taxonomic class Bacilli (equally distributed among the Bacillales and
140 Lactobacillales orders). As many as 88.7% of the sequenced Bacilli have a Com
141 system. We also detected Com systems in one Clostridia (out of 336) and six
142 Erysipelotrichia (out of 14). In total, 349 different species have the potential to
143 express a Com (pseudo)pilus (Supplemental Spreadsheet 2).

144 Taken together, these findings suggest that the Com (pseudo)pilus is almost
145 ubiquitous in Bacilli, and can be advantageously studied in the Tff model monoderm
146 species *S. sanguinis*.

147

148 ***ComGC, the major subunit of Com (pseudo)pili, is highly conserved and***
149 ***defines a monophyletic group among type IV pilins***

150 We next focused specifically on the major subunit of Com (pseudo)pili, the pilin
151 ComGC (Chen et al., 2006; Laurenceau et al., 2013). Compared to major pilins from
152 Tfpa, ComGC is ~40% shorter, with 94 or 93 aa for the processed $ComGC_{SS}$ and
153 $ComGC_{SP}$, respectively (10.18 and 10.42 kDa). Moreover, unlike most other pilins, in
154 which the only detectable sequence homology is usually in the α1N portion of the
155 class III signal peptide (Giltner et al., 2012), ComGC orthologs show extensive
156 sequence identity. For example, processed $ComGC_{SS}$ and $ComGC_{SP}$ display 65.6%
157 overall sequence identity (Fig. 2). Similarly, processed $ComGC_{SS}$ and $ComGC_{BS}$
158 (from *B. subtilis*) show 33.3% sequence identity overall (Fig. S1). This is consistent
159 with the existence of a ComGC signature in the InterPro database (IPR016940)
160 (Jones et al., 2014), which lists 2,809 ComGC entries. Global multiple alignment of
161 these ComGC proteins shows that most of the sequence is conserved in ~90% of the
162 entries (Fig. 2). In Fig. 2, the consensus sequences have been aligned to $ComGC_{SS}$
163 and $ComGC_{SP}$. Strikingly, some residues show sequence identity in virtually all the

164 entries, including residues outside of the α1N portion (such as Ala_{38}, Gln_{46}, Tyr_{50} and
165 Leu_{64} in $ComGC_{SS}$), which is highly unusual in type IV pilins.

166 The above observations suggest that Com (pseudo)pili form a highly
167 homogeneous Tff sub-family. This was tested by performing a phylogenetic analysis
168 based on the protein sequences of major pilins from different Tff found in a wide
169 variety of bacteria, including Tfpa, Tfpb, Tfpc (also known as Tad pili), mannose-
170 sensitive hemagglutinin pili (MSH), type II secretion systems (T2SS) and Com
171 (pseudo)pili. The phylogeny tree that was generated (Fig. 3), using IQ-TREE
172 (Nguyen et al., 2015), reveals that several Tff are in clear monophyletic groups with
173 good branch support, >96% ultrafast bootstrap (UFBoot) (Hoang et al., 2018). Of
174 particular interest, Com (pseudo)pili define a highly supported clade (99% UFBoot),
175 clearly distinct form all other Tff systems.

176 Taken together, these findings show that ComGC is a small pilin with a highly
177 conserved sequence, which defines a monophyletic group.

178

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

181 Since high-resolution structural information is needed to improve our understanding
182 of Com (pseudo)pili, we decided to solve the 3D structure of $ComGC_{SS}$. To facilitate
183 protein purification, we used a synthetic *$comGC_{SS}$* gene codon-optimised for
184 expression in *Escherichia coli*, and fused the 72 aa-long soluble portion of $ComGC_{SS}$
185 to a non-cleavable N-terminal hexahistidine tag (6His). This truncation, which
186 removes the first 22 residues that invariably form a protruding hydrophobic α-helix (α-
187 1N), is known not to affect the structural fold of the rest of the protein (Giltner et al.,
188 2012). The resulting 8.79 kDa 6His-$ComGC_{SS}$ protein could be purified using a
189 combination of affinity and gel-filtration chromatography, as a well-behaved and
190 soluble protein. After purification of isotopically labelled protein with ^{13}C and ^{15}N for
191 backbone and side-chain NMR resonance assignments, we could assign 99.5% of

192 the backbone and 92% of assignable protons overall. Structural ensembles were
193 determined with 962 NOE based restraints, 50 hydrogen bonds, 110 dihedral angles
194 restraints and 39 residual dipolar couplings (RDC) (Table 1). As can be seen in Fig.
195 4, $ComGC_{SS}$ 3D structure is unlike that of any type IV pilin present in PDB, as it is
196 purely helical, with three distinct helices connected by loops. The helices present are
197 consistent with JPred secondary structure prediction (Fig. 2) (Drozdetskiy et al.,
198 2015). The N-terminal α1-helix, which involves residues 37-53 of the processed
199 protein, corresponds to α1C since the hydrophobic α1N has been truncated in 6His-
200 $ComGC_{SS}$. Tightly packed against this α1-helix, in a parallel plane, are α2-helix
201 (residues 61-67) and α3-helix (residues 72-85), which stack against each other in
202 antiparallel fashion (Fig. 4A) and orthogonally to α1. Except for the N-terminal
203 unstructured residues, the $ComGC_{SS}$ structures within the NMR ensemble superpose
204 well onto each other (Fig. 4B), with a root mean square deviation (RMSD) of 1.22 Å
205 for Cα atoms, which suggests that there is no significant flexibility in this region of the
206 structure (Krissinel and Henrick, 2004). The unstructured N-terminus, which lacks
207 long and medium NOEs present in the ordered regions of the proteins, was predicted
208 to be highly dynamic based on TALOS+ (Shen et al., 2009), with an average S^2 order
209 parameter of 0.49 ± 0.10.

210 Our $ComGC_{SS}$ structure differs markedly from the recently reported solution
211 structure of $ComGC_{SP}$ (PDB 5NCA) (Muschiol et al., 2017), which is surprising
212 considering the high sequence identity between these two proteins (Fig. 2).
213 Therefore, in order to define the structural relationship between ComGC orthologs,
214 we decided to solve the structure of $ComGC_{SP}$. As above, we used a synthetic
215 *$comGC_{SP}$* gene codon-optimised for expression in *E. coli*, we fused the 71 aa-long
216 soluble portion of $ComGC_{SP}$ to a non-cleavable N-terminal 6His tag and purified
217 doubly labelled 6His-$ComGC_{SP}$ (8.99 kDa). Again, assignment was excellent since
218 98.1% of the backbone and 90% of assignable protons overall could be assigned.
219 Structural ensembles were determined with 880 NOE based restraints, 54 hydrogen

220 bonds, 102 dihedral angles restraints and 38 RDC (Table 1). As can be seen in Fig.
221 5, our $ComGC_{SP}$ 3D structure is highly similar to the structure of $ComGC_{SS}$, but very
222 different from the solution structure that was recently determined from a low number
223 of restraints (Fig. S2) (Muschiol et al., 2017). In brief, $ComGC_{SP}$ display three distinct
224 helices, with α2-helix (residues 60-66) and α3-helix (residues 71-82) stacking against
225 each other and packing orthogonal to the N-terminal α1-helix (Fig. 5A). As for
226 $ComGC_{SS}$, except for the unstructured N-terminus, there is no significant flexibility in
227 $ComGC_{SP}$ since the structures within the NMR ensemble superpose well onto each
228 other, with a RMSD of 1.60 Å for Cα atoms (Fig. 5B). Our $ComGC_{SS}$ and $ComGC_{SP}$
229 averaged structures are highly similar (Fig. 5C), with 1.79 Å RMSD between their
230 ordered regions and 1.54 Å RMSD for the helical regions, which is consistent with the
231 high sequence identity between these two proteins.

232 As determined by GETAREA (Fraczkiewicz and Braun, 1998) with a probe
233 radius of 1.4 Å, the average ratio of solvent exposure for the ordered portion of
234 $ComGC_{SS}$ is 48.3%, relative to 6.7% for those residues determined to be on the
235 interior. In our $ComGC_{SS}$ structure, conserved residues Val_{43}, Gln_{46}, Tyr_{50}, Leu_{64} and
236 Ile_{70} are deeply buried, with an average of only 6.0% solvent exposure, forming a
237 critical portion of an hydrophobic core contributing to the globular fold of ComGC.
238 (Fig. 6). In contrast, conserved Gly_{68} is solvent exposed, which is important for the
239 formation of the α2-helix-turn-α3-helix motif where a tiny residue at the beginning of
240 the turn is necessary to provide the flexibility and lack of steric restrictions required
241 for turning. These observations also apply to our $ComGC_{SP}$ structure, and are
242 surprisingly reflected in the conservation of multiple chemical shifts between the
243 conserved residues in our two structures (Fig. S3). In addition, modelling of the
244 globular head of $ComGC_{BS}$ (Fig. S4), which predicts a globular fold similar to
245 $ComGC_{SS}$ and $ComGC_{SP}$, shows that Cys_{36} and Cys_{76} are in close enough proximity
246 to form a disulfide bond. Such disulfide bond, which is absent in $ComGC_{SS}$ and
247 $ComGC_{SP}$ that do not have Cys residues, is expected to stabilise the globular fold

248 and was reported to stabilise ComGC in *B. subtilis* (Chen et al., 2006; Meima et al.,
249 2002).

250 Since the hydrophobic α1N that has been truncated in 6His-ComGC$_{SS}$ is highly
251 similar to the corresponding portion of the PilE major pilin from *Neisseria*
252 *gonorrhoeae* (Fig. S5), we could model the structure of the portion of α1 truncated in
253 our construct, which produced a reliable model of full-length ComGC$_{SS}$ (Fig. 7).
254 Comparison with the two different pilin folds identified so far - pilins from *N.*
255 *gonorrhoeae* and *Geobacter sulfurreducens* have been chosen as representative
256 models - clearly shows that ComGC adopts a radically different type IV pilin fold (Fig.
257 7). All three pilins have in common an extended N-terminal α1-helix, the universal
258 defining structural feature of type IV pilins (Giltner et al., 2012). In addition, while the
259 very short *G. sulfurreducens* pilin almost exclusively consists of α1, both ComGC and
260 PilE display a typical lollipop shape with a globular head mounted onto a "stick" (the
261 α1-helix). However, unlike in canonical pilins where the globular head consists of a 4-
262 to 7-stranded antiparallel β-sheet in a parallel plane to α1, oriented 45° or more
263 relative to the long axis of α1 (Giltner et al., 2012), in ComGC the structural
264 backbone of the globular head is an helix-turn-helix roughly orthogonal to α1 (Fig. 7).
265 While this fold falls within the class of mainly α and the architecture of orthogonal
266 bundles, it represents a novel fold.

267 Taken together, these structural findings show that ComGC orthologs display
268 conserved 3D structures, with a previously unreported type IV pilin fold.

269

270 ***ComGC novel pilin fold is compatible with helical Tff assembly***

271 Since ComGC represents a novel type IV pilin structural fold, it was important to
272 determine whether it could be modelled into recent cryo-EM structures obtained for a
273 variety of bacterial Tff (Kolappan et al., 2016; Lopez-Castilla et al., 2017; Wang et al.,
274 2017). These structures have revealed that a segment of α1 is melted during filament

assembly, centred on helix-breaking residues Pro_{22} and confirmed that the N-terminal Phe_1 is methylated. Since that portion α1 is highly conserved in ComGC, including the helix-breaking Pro_{22} (Fig. S5), and ComGC has been shown to be N-terminally methylated in *S. pneumoniae* (Laurenceau et al., 2013), we used the PulG pilin from the cryo-EM structure of *Klebsiella oxytoca* T2SS pseudopili (Lopez-Castilla et al., 2017) to produce a reliable full-length 3D structural model of $ComGC_{SS}$ with a melted segment and an N-terminal methyl-Phe_1 (Fig. 8), using SWISS-MODEL (Waterhouse et al., 2018). Apart from these two modifications, this full-length $ComGC_{SS}$ model is very similar to the one in Fig. 7 for which a different template has been used. 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 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 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 Tff 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 *K. oxytoca* T2SS pseudopili, which have a morphology similar to Com (pseudo)pili observed in *S. pneumoniae* (Laurenceau et al., 2013; Muschiol et al., 2017). This produced a homology model with good Ramachandran plot statistics, i.e. allowed (95.3%), generously allowed (3.5%) and disallowed (1.1%) based on PROCHECK (Laskowski et al., 1993). As can be seen in Fig. 8, 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 and establish extensive hydrophobic interactions. In addition, the Glu_5 side chain of subunit S establishes a salt bridge and

303 a hydrogen bond with Phe_1 and Thr_2, respectively, of S+1. Importantly, the globular
304 heads are stacked on top of each other along the long axis of the filaments and their
305 helix-turn-helix structural backbone forms the outer surface of the filaments (Fig. 8).

306 **DISCUSSION**

307 Their virtual ubiquity in prokaryotes and role in a variety of key biological processes
308 make Tff an important research topic (Berry and Pelicic, 2015; Denise et al., 2019).
309 Com (pseudo)pili are involved in DNA uptake in naturally competent monoderm
310 bacteria (Dubnau and Blokesch, 2019). Imported DNA, which usually leads to
311 genome diversification via transformation, can also be used as a source of food or as
312 a template for repair of damaged genomic DNA (Johnston et al., 2014). Compared to
313 Tff in diderms, most notably Tfp and T2SS that have been extensively studied, Com
314 (pseudo)pili have been understudied, including from a structural point of view. In this
315 report, we focused on the major subunit of Com (pseudo)pili, the ComGC pilin, which
316 we analysed genomically, phylogenetically and structurally. This led to the notable
317 findings discussed below.

318 Although Com (pseudo)pili have been primarily studied in two model competent
319 species (*B. subtilis* and *S. pneumoniae*), the present study makes it clear that they
320 are widespread since complete sets of Com-encoding genes are readily detected in
321 more than 2,300 genomes corresponding to almost 350 different species. However,
322 unlike promiscuous Tff such as Tfpa and Tfpc that are found in virtually all phyla of
323 Bacteria (Denise et al., 2019), Com (pseudo)pili are restricted to a single phylum
324 (Firmicutes) and almost exclusively to a single underlying class of monoderms
325 (Bacilli), where they are almost ubiquitous. Indeed, an overwhelming majority of
326 Bacilli genomes (88%) have Com-encoding genes. Interestingly, the major subunit of
327 Com (pseudo)pili (ComGC) shows extensive sequence conservation in the
328 corresponding genomes and define a clear monophyletic group within type IV pilins.
329 Taken together, these observations suggest that the Com (pseudo)pilus is a Tff that
330 has emerged only once, very early during the diversification of Firmicutes, where it
331 has remained largely confined ever since. Since the Com-encoding genes have not
332 become pseudogenes, it is likely that most Bacilli have the ability to assemble a Com
333 (pseudo)pilus and take up DNA. However, since only a handful of these species have

been experimentally shown to be competent (Johnston et al., 2014), 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 (pseudo)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 (pseudo)pilus, displays an entirely novel major pilin fold where the extended N-terminal α1-helix, the universal defining structural feature of type IV pilins (Giltner et al., 2012), is topped by a purely helical globular head. ComGC thus appears to be a "middle ground" between longer canonical pilins (e.g. *N. gonorrhoeae*), 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, the major subunit of *S. sanguinis* Tfp, which has two short α-helices in the loop connecting α1 and the antiparallel β-sheet (Berry et al., 2019). Importantly, this putative "truncation" does not interfere with the ability of ComGC to be assembled into helical filaments, since ComGC could be readily modelled into recent Tff structures (Kolappan et al., 2016; Lopez-Castilla et al., 2017; Wang et al., 2017). Com (pseudo)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 which make contacts between chains in our (pseudo)pilus model beyond the classically conserved α1N, and the fact that the two structures that we solved are virtually

362 identical, strongly suggest that these structural features apply to the whole ComGC
363 clade, including species such as *B. subtilis* where extended filaments have not been
364 observed (Chen et al., 2006). It is therefore surprising that a recently published NMR
365 structure of $ComGC_{SP}$ (PDB 5NCA) (Muschiol et al., 2017) differs dramatically from
366 ours. While the previous structure is purely helical as well, the orientation of the α2
367 and α3 helices is entirely different, resulting in an absence of packing of the
368 conserved hydrophobic core. Therefore, PDB 5NCA which resembles a one-sided
369 "pick-axe" with no globular head cannot be readily modelled into a helical Tff (Fig.
370 S7). Indeed, the pick-axe points towards the filament core, which is sterically
371 disallowed and incompatible with filament assembly. Interestingly, our assignments
372 vary only slightly from those previously produced for PDB 5NCA (Fig. S8). However,
373 while we have managed to successfully assign 90% assignable protons overall, the
374 previous assignment was merely 65% (Muschiol et al., 2017), which probably
375 accounts for the apparently "unfolded" state of PDB 5NCA. Indeed, without a high
376 degree of proton identification, the assignment of NOESY peaks and production of
377 distance restraints fails. Local hydrogen bonds and dihedral restraints often cannot
378 compensate for lack of long-range NOEs within the protein interior or between
379 elements of secondary structure.

380 Together with these conserved structural features, the conservation *en bloc* of
381 the genes encoding the Com (pseudo)pilus strongly suggests that the molecular
382 mechanisms of filament assembly and DNA uptake are widely conserved in
383 Firmicutes. These mechanisms, which remain poorly understood, can be
384 advantageously studied in *S. sanguinis*, which has recently emerged as a monoderm
385 Tfp model (Pelicic, 2019). Actually, *S. sanguinis* is the only monoderm shown to
386 express two distinct Tff, retractable Tfpa and Com (pseudo)pili, which further
387 cements it as a prime Tff model species. Comparison with other Tff systems shows
388 that the machinery involved in biogenesis of Com (pseudo)pili is one of the simplest,
389 by far. Since ComGD, ComGE, ComGF and ComGG pilins are likely to be minor

390 (pseudo)pilus components important for filament stability and function (a conserved
391 role for minor pilins in various Tff) (Berry and Pelicic, 2015), and ComC is the prepilin
392 peptidase processing pilins (Chung et al., 1998), it appears that assembly of ComGC
393 into filaments is mediated by two proteins only. Namely, an extension ATPase
394 (ComGA) and a platform protein (ComGB), which together will assemble processed
395 ComGC in a right-handed helical filament. Upon DNA binding, which has been
396 visualised for *S. pneumoniae* Com (pseudo)pili, but the receptor is yet to be identified
397 (Laurenceau et al., 2013), uptake will be initiated by filament retraction (Ellison et al.,
398 2018). Since there is no dedicated retraction ATPase, ComGA might therefore be a
399 bifunctional motor powering both extension and retraction like recently suggested for
400 the Tfpc motor (Ellison et al., 2019). It would be interesting to image Com filaments
401 dynamics and DNA-binding ability in live cells, using a labelling strategy that has
402 recently enabled the visualisation of these steps for Tfpa involved in competence in
403 naturally competent diderm species (Ellison et al., 2018).

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

410 **EXPERIMENTAL PROCEDURES**

411

412 **Bioinformatic analyses**

413 Protein sequences were routinely analysed using the DNA Strider program. Protein
414 sequence alignments were done using the Clustal Omega server at EMBL-EBI.
415 Pretty-printing of alignment files was done using BoxShade server at ExPASy.
416 Reformatting of large multiple alignment files was done using the MView server at
417 EMBL-EBI. Prediction of functional domains was done using the InterProScan server
418 at EMBL-EBI, which was also used to download all the ComGC protein entries with
419 an IPR0160940 domain. Protein secondary structure prediction was done using
420 JPred server at University of Dundee. Protein 3D structures were downloaded from
421 the RCSB PDB server. Molecular visualisation of protein 3D structures was done
422 using PyMOL (Schrödinger). The GETAREA server, at UTMB, was used for
423 calculating the solvent accessible surface area of ComGC proteins.

424 Detection of the Com systems in genomes available in NCBI RefSeq database
425 (last accessed in April 2019, 13,512 genomes of Bacteria and Archaea) was done as
426 described previously (Denise et al., 2019), using MacSyFinder (Abby et al., 2014)
427 and the relevant HMM Com model (Denise et al., 2019). Phylogenetic analysis based
428 on protein sequences of major pilins of different Tff involved an initial alignment of the
429 sequences using MAFFT v7.273 (Katoh and Standley, 2013), specifically the linsi
430 algorithm. Multiple alignments were analysed using Noisy v1.5.12 (Dress et al., 2008)
431 with default parameters, in order to select the informative sites. Next, we inferred
432 maximum likelihood trees from the curated alignments using IQ-TREE v 1.6.7.2
433 (Nguyen et al., 2015), with option -allnni. We evaluated the node supports using the
434 options -bb 1,000 for ultra-fast bootstraps, and -alrt 1,000 for SH-aLRT (Hoang et al.,
435 2018). The best evolutionary model was selected with ModelFinder
436 (Kalyaanamoorthy et al., 2017), option -MF and BIC criterion. We used the option -
437 wbtl to conserve all optimal trees and their branches length.

438

439 **Protein expression and purification**

440 A synthetic gene, codon-optimised for *E. coli* expression, encoding $ComGC_{SS}$ from *S.*
441 *sanguinisis* 2908 (Gurung et al., 2016) was synthesised and cloned by GeneArt,
442 yielding pMA-T-$comGC_{SS}$ (Table S2). The portion of the gene encoding residues 23-
443 94 from the mature protein was PCR-amplified using $comGC_{SS}$-F and $comGC_{SS}$-R
444 primers (Table S3), cut with NcoI and BamHI and cloned into the pET28b vector
445 (Novagen) cut with the same enzymes. The forward primer was designed to fuse a
446 non-cleavable N-terminal 6His tag to $ComGC_{SS}$. The resulting plasmid was verified
447 by sequencing and transformed into chemically competent *E. coli* BL21(DE3) cells. A
448 single colony was transferred to 10 ml of LB supplemented with 50 $\mu g.ml^{-1}$
449 kanamycin and grown at 37°C overnight (O/N). This pre-culture was back-diluted
450 100-fold into 1 l M9 minimal medium, supplemented with antibiotic, a mixture of
451 vitamins and trace elements, and ^{13}C D-glucose and ^{15}N NH_4Cl for isotopic labelling.
452 Cells were grown in an orbital shaker at 37°C until the OD_{600} reached 0.7, before
453 adding 0.4 mM IPTG (Merck Chemicals) to induce protein expression during 16 h at
454 18°C. Cells were then harvested by centrifugation at 8,000 *g* for 20 min and
455 subjected to one freeze/thaw cycle in lysis buffer (PBS pH 7.4, with EDTA-free
456 protease inhibitors). This lysate was further disrupted by repeated cycles of
457 sonication, pulses of 5 sec on and 5 sec off during 5 min, until the cell suspension
458 was visibly less viscous. The cell lysate was then centrifuged for 20 min at 18,000 *g*
459 to remove cell debris. The clarified lysate was then passed using an ÄKTA Purifier
460 FPLC through a 1 ml HisTrap HP column (GE Healthcare), pre-equilibrated in lysis
461 buffer. The column was then washed extensively with lysis buffer to remove unbound
462 material before 6His-$ComGC_{SS}$ was eluted using elution buffer (PBS pH 7.4, 200 mM
463 NaCl, 300 mM imidazole). Affinity-purified $ComGC_{SS}$ was further purified by gel-
464 filtration chromatography on an HiLoad 16/600 Superdex 75 column (GE

465 Healthcare), using (25 mM Na_2HPO_4/NaH_2PO_4 pH 6, 200 mM NaCl) buffer for
466 elution. For RDC measurements we produced ^{15}N labelled protein as follows.
467 Bacteria grown O/N in 5 ml LB with antibiotic were sub-cultured at 37ºC in 0.8 l LB to
468 0.6 OD_{600}, and then transferred to 0.4 l M9 with ^{15}N NH_4Cl, unlabelled D-glucose, and
469 10 µg.l^{-1} thiamine. Cultures were induced with 0.3 mM IPTG at 16ºC for 18 h. After
470 the production of a clarified lysate, protein was purified as above, except for the use
471 of hand-made Ni-NTA agarose (Qiagen) in (50 mM Tris pH 8, 300 mM NaCl) and
472 eluted using (50 mM Tris pH 8, 200 mM NaCl, 300 mM imidazole), and Superdex 75
473 10/300 GL (GE Healthcare) columns in (25 mM Tris pH 8, 200 mM NaCl) and
474 dialysed into (25 mM Na_2HPO_4/NaH_2PO_4 pH 6, 50 mM NaCl).

475 For $ComGC_{SP}$, a codon-optimised synthetic gene based on the gene from *S.*
476 *pneumoniae* R6 was synthesised and cloned by GeneArt, yielding pMA-T-$comGC_{SP}$
477 (Table S2). The portion of the gene encoding residues 23-93 from the mature protein
478 was PCR-amplified using $comGC_{SP}$-F and $comGC_{SP}$-R primers (Table S3), cut with
479 NcoI and BamHI and cloned into the pET28b vector (Novagen) cut with the same
480 enzymes. The forward primer was designed to fuse a non-cleavable N-terminal 6His
481 tag to $ComGC_{SS}$. The resulting plasmid was verified by sequencing and transformed
482 into chemically competent *E. coli* BL21(DE3) cells. A single colony was transferred to
483 5 ml of LB supplemented with 50 µg.ml^{-1} kanamycin and grown O/N at 37°C. Bacteria
484 were sub-cultured at 37ºC in 0.8 l LB with antibiotic to OD_{600} 0.7, and then transferred
485 into 0.4 l M9 with10 µg.l^{-1} thiamine, and either ^{15}N NH_4Cl and unlabelled D-glucose,
486 or ^{15}N NH_4Cl and ^{13}C D-glucose. Cultures were induced with 0.3 mM IPTG at 16ºC
487 for 18 h. After the production of a clarified lysate, $ComGC_{SS}$ was purified as above
488 using hand-made Ni-NTA agarose (Qiagen) and Superdex 75 10/300 GL (GE
489 Healthcare) columns.

490

491 **NMR spectroscopy and structure determination**

492 All data was collected on Bruker Avance III HD 800 MHz and 600 MHz triple
493 resonance spectrometers with cryoprobes operated at 25ºC. For $ComGC_{SS}$, a
494 sample containing ^{13}C, ^{15}N labelled protein at 1 mM in NMR buffer (25 mM
495 Na_2HPO_4/NaH_2PO_4 pH 6, 50 mM NaCl, 5% D_2O) was used for assignment
496 experiments and structure determination. For $ComGC_{SP}$, a sample containing ^{13}C,
497 ^{15}N labelled protein at 1.8 mM in NMR buffer was used for assignment experiments
498 and structure determination. Resonance assignments for $ComGC_{SS}$ were performed
499 using ^{15}N HSQC, ^{13}C aliphatic HSQC, HNCACB, CBCACONH, HBHA, HNCO,
500 HNCACO, HCCCONH, CCCONH and CCH. For $ComGC_{SP}$, assignments were
501 performed using ^{15}N HSQC, ^{13}C aliphatic HSQC, HNCA, CBCANH, CBCACONH,
502 HBHA, HNCO, HNCACO, HCCCONH, CCCONH and CCH. All data was processed
503 using MddNMR (Orekhov and Jaravine, 2011) for reconstruction after Non-Uniform
504 Sampling and NMRPipe (Delaglio et al., 1995). Peak picking and assignments were
505 performed in SPARKY (Lee et al., 2015).

506 NOE peak lists were used, with mixing time of 140 msec, from 3D ^{13}C HSQC-
507 NOESY, 3D ^{15}N HSQC-NOESY for $ComGC_{SP}$, and simultaneous $^{13}C/^{15}N$ chemical
508 shift evolution NOESY for $ComGC_{SS}$. For both proteins, RDC lists were derived from
509 ^{15}N HSQC-IPAP experiments on ^{15}N labelled isotropic and aligned sample in 3%
510 PEG/hexanol liquid crystal, with D_2O splitting of ~7 Hz. RDCs were included in the
511 structure calculations if there was baseline resolution and for residues where
512 TALOS+ predicted order parameter of >0.8. Angular constraints from TALOS+ were
513 used in the structure calculations. Both $ComGC_{SS}$ and $ComGC_{SP}$ structures were
514 determined using Ponderosa-C/S (Lee et al., 2015), refined using Xplor-NIH 2.52
515 (Schwieters et al., 2006), aligned using Theseus (Theobald and Wuttke, 2008), and
516 secondary structure checked using Stride (Frishman and Argos, 1995). Structure
517 validation was performed using PSVS (Bhattacharya et al., 2007), PROCHECK
518 (Laskowski et al., 1993) and in-house scripts.

519

520 **Modelling**

521 SWISS-MODEL server at ExPASy was used for modelling protein 3D structures. In
522 brief, the full-length $ComGC_{SS}$ was modelled with using *N. gonorrhoeae* major pilin
523 (PDB 2PIL) as a template (Forest et al., 1999). We first modelled the missing α1
524 residues in our structure, which was aligned to our Xplor-NIH-produced average
525 NMR structure (without the first unstructured α1 residues) using PyMol and finally
526 merged using Coot (Emsley et al., 2010).

527 Similarly, the full-length $ComGC_{SS}$ structure within filaments was modelled by
528 using one *K. oxytoca* PulG subunit from the T2SS pseudopilus (PDB 5WDA) as a
529 template for the missing α1 residues in our structure (Lopez-Castilla et al., 2017).
530 This full-length $ComGC_{SS}$ was then used to produce models of a variety of more or
531 less distant ComGC orthologs. The Com (pseudo)pilus model was produced after
532 alignment of the averaged NMR structure ComGC α1-helices to the α1-helices of
533 SWISS-MODEL PulG-based homology model subunits in the T2SS pseudopilus.
534 This was also done for the recently published $ComGC_{SP}$ structure (PDB 5NCA). The
535 structures were fused and we added the N-terminal methyl-Phe_1 using Coot (Emsley
536 et al., 2010). In addition, we modelled packing of full-length $ComGC_{SS}$ in the PAK
537 pilus from *P. aeruginosa* (PDB 5VXY) (Wang et al., 2017).

538 **ACCESSION NUMBERS**

539 The NMR solution structures of $ComGC_{SS}$ and $ComGC_{SP}$ have been deposited in the
540 Protein Data Bank under entries 6TXT and 6Y1H, respectively. Chemical shift
541 assignments and NOE-based restraints used in structure calculations are available
542 from the BMRB under accession numbers 50194 and 7441, respectively.

AUTHOR CONTRIBUTIONS

V.P. designed the research. E.P.C.R., S.J.M. and V.P. directed the research. Experiments were done by D.S., J.L.B and R.D. All authors contributed to writing the manuscript.

547 **ACKNOWLEDGMENTS**

548 This work was funded by a Medical Research Council grant (MR/P022197/1) to V.P.
549 R.D. was funded by the doctoral school Complexité du vivant-ED515 (contract
550 number 2449/2016). E.P.C.R. was funded by the INCEPTION project (PIA/ANR-16-
551 CONV-0005). This work relied heavily on the use of the Cross-Faculty NMR Centre
552 at Imperial College London. We are grateful to Nicolas Biais (City University of New
553 York) and Romé Voulhoux (CNRS Marseille) for critical reading of the manuscript.

554 **REFERENCES**

555 Abby, S.S., Neron, B., Menager, H., Touchon, M., and Rocha, E.P. (2014).
556 MacSyFinder: a program to mine genomes for molecular systems with an application
557 to CRISPR-Cas systems. PLoS One *9*, e110726.

558 Albano, M., Breitling, R., and Dubnau, D.A. (1989). Nucleotide sequence and genetic
559 organization of the *Bacillus subtilis comG* operon. J Bacteriol *171*, 5386-5404.

560 Berry, J.L., Gurung, I., Anonsen, J.H., Spielman, I., Harper, E., Hall, A.M.J.,
561 Goosens, V.J., Raynaud, C., Koomey, M., Biais, N.*, et al.* (2019). Global biochemical
562 and structural analysis of the type IV pilus from the Gram-positive bacterium
563 *Streptococcus sanguinis*. J Biol Chem *294*, 6796-6808.

564 Berry, J.L., and Pelicic, V. (2015). Exceptionally widespread nano-machines
565 composed of type IV pilins: the prokaryotic Swiss Army knives. FEMS Microbiol Rev
566 *39*, 134-154.

567 Bhattacharya, A., Tejero, R., and Montelione, G.T. (2007). Evaluating protein
568 structures determined by structural genomics consortia. Proteins *66*, 778-795.

569 Cehovin, A., Simpson, P.J., McDowell, M.A., Brown, D.R., Noschese, R., Pallett, M.,
570 Brady, J., Baldwin, G.S., Lea, S.M., Matthews, S.J.*, et al.* (2013). Specific DNA
571 recognition mediated by a type IV pilin. Proc Natl Acad Sci USA *110*, 3065-3070.

572 Chen, I., Provvedi, R., and Dubnau, D. (2006). A macromolecular complex formed by
573 a pilin-like protein in competent *Bacillus subtilis*. J Biol Chem *281*, 21720-21727.

574 Chung, Y.S., Breidt, F., and Dubnau, D. (1998). Cell surface localization and
575 processing of the ComG proteins, required for DNA binding during transformation of
576 *Bacillus subtilis*. Mol Microbiol *29*, 905-913.

577 Chung, Y.S., and Dubnau, D. (1995). ComC is required for the processing and
578 translocation of ComGC, a pilin-like competence protein of *Bacillus subtilis*. Mol
579 Microbiol *15*, 543-551.

580 Chung, Y.S., and Dubnau, D. (1998). All seven *comG* open reading frames are
581 required for DNA binding during transformation of competent *Bacillus subtilis*. J
582 Bacteriol *180*, 41-45.
583 Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. (1995).
584 NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J
585 Biomol NMR *6*, 277-293.
586 Denise, R., Abby, S.S., and Rocha, E.P.C. (2019). Diversification of the type IV
587 filament superfamily into machines for adhesion, protein secretion, DNA uptake, and
588 motility. PLoS Biol *17*, e3000390.
589 Dress, A.W., Flamm, C., Fritzsch, G., Grunewald, S., Kruspe, M., Prohaska, S.J., and
590 Stadler, P.F. (2008). Noisy: identification of problematic columns in multiple
591 sequence alignments. Algorithm Mol Biol *3*, 7.
592 Drozdetskiy, A., Cole, C., Procter, J., and Barton, G.J. (2015). JPred4: a protein
593 secondary structure prediction server. Nucleic Acids Res *43*, W389-W394.
594 Dubnau, D., and Blokesch, M. (2019). Mechanisms of DNA uptake by naturally
595 competent bacteria. Annu Rev Genet *53*, 217-237.
596 Ellison, C.K., Dalia, T.N., Vidal Ceballos, A., Wang, J.C., Biais, N., Brun, Y.V., and
597 Dalia, A.B. (2018). Retraction of DNA-bound type IV competence pili initiates DNA
598 uptake during natural transformation in *Vibrio cholerae*. Nat Microbiol *3*, 773-780.
599 Ellison, C.K., Kan, J., Chlebek, J.L., Hummels, K.R., Panis, G., Viollier, P.H., Biais,
600 N., Dalia, A.B., and Brun, Y.V. (2019). A bifunctional ATPase drives tad pilus
601 extension and retraction. Sci Adv *5*, eaay2591.
602 Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and
603 development of Coot. Acta Crystallogr Sect D: Biol Crystallogr *66*, 486-501.
604 Forest, K.T., Dunham, S.A., Koomey, M., and Tainer, J.A. (1999). Crystallographic
605 structure reveals phosphorylated pilin from *Neisseria*: phosphoserine sites modify
606 type IV pilus surface chemistry and fibre morphology. Mol Microbiol *31*, 743-752.

607 Fraczkiewicz, R., and Braun, W. (1998). Exact and efficient analytical calculation of
608 the accessible surface areas and their gradients for macromolecules. J Comput
609 Chem *19*, 319-333.
610 Frishman, D., and Argos, P. (1995). Knowledge-based protein secondary structure
611 assignment. Proteins *23*, 566-579.
612 Giltner, C.L., Nguyen, Y., and Burrows, L.L. (2012). Type IV pilin proteins: versatile
613 molecular modules. Microbiol Mol Biol Rev *76*, 740-772.
614 Gurung, I., Spielman, I., Davies, M.R., Lala, R., Gaustad, P., Biais, N., and Pelicic, V.
615 (2016). Functional analysis of an unusual type IV pilus in the Gram-positive
616 *Streptococcus sanguinis*. Mol Microbiol *99*, 380-392.
617 Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018).
618 UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol *35*, 518-522.
619 Imam, S., Chen, Z., Roos, D.S., and Pohlschröder, M. (2011). Identification of
620 surprisingly diverse type IV pili, across a broad range of Gram-positive bacteria.
621 PLoS One *6*, e28919.
622 Johnston, C., Martin, B., Fichant, G., Polard, P., and Claverys, J.P. (2014). Bacterial
623 transformation: distribution, shared mechanisms and divergent control. Nat Rev
624 Microbiol *12*, 181-196.
625 Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H.,
626 Maslen, J., Mitchell, A., Nuka, G.*, et al.* (2014). InterProScan 5: genome-scale
627 protein function classification. Bioinformatics *30*, 1236-1240.
628 Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S.
629 (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat
630 Methods *14*, 587-589.
631 Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software
632 version 7: improvements in performance and usability. Mol Biol Evol *30*, 772-780.
633 Kolappan, S., Coureuil, M., Yu, X., Nassif, X., Egelman, E.H., and Craig, L. (2016).
634 Structure of the *Neisseria meningitidis* type IV pilus. Nat Commun *7*, 13015.

635 Krissinel, E., and Henrick, K. (2004). Secondary-structure matching (SSM), a new
636 tool for fast protein structure alignment in three dimensions. Acta Crystallogr Sect D:
637 Biol Crystallogr *60*, 2256-2268.
638 LaPointe, C.F., and Taylor, R.K. (2000). The type 4 prepilin peptidases comprise a
639 novel family of aspartic acid proteases. J Biol Chem *275*, 1502-1510.
640 Laskowski, R.A., W., M.M., Moss, D.S., and Thornton, J.M. (1993). PROCHECK - a
641 program to check the stereochemical quality of protein structures. J Appl Crystallogr
642 *26*, 283-291.
643 Laurenceau, R., Pehau-Arnaudet, G., Baconnais, S., Gault, J., Malosse, C.,
644 Dujeancourt, A., Campo, N., Chamot-Rooke, J., Le Cam, E., Claverys, J.P.*, et al.*
645 (2013). A type IV pilus mediates DNA binding during natural transformation in
646 *Streptococcus pneumoniae*. PLoS Pathog *9*, e1003473.
647 Lee, W., Tonelli, M., and Markley, J.L. (2015). NMRFAM-SPARKY: enhanced
648 software for biomolecular NMR spectroscopy. Bioinformatics *31*, 1325-1327.
649 Lopez-Castilla, A., Thomassin, J.L., Bardiaux, B., Zheng, W., Nivaskumar, M., Yu, X.,
650 Nilges, M., Egelman, E.H., Izadi-Pruneyre, N., and Francetic, O. (2017). Structure of
651 the calcium-dependent type 2 secretion pseudopilus. Nat Microbiol *2*, 1686-1695.
652 Meima, R., Eschevins, C., Fillinger, S., Bolhuis, A., Hamoen, L.W., Dorenbos, R.,
653 Quax, W.J., van Dijl, J.M., Provvedi, R., Chen, I.*, et al.* (2002). The *bdbDC* operon of
654 *Bacillus subtilis* encodes thiol-disulfide oxidoreductases required for competence
655 development. J Biol Chem *277*, 6994-7001.
656 Merz, A.J., So, M., and Sheetz, M.P. (2000). Pilus retraction powers bacterial
657 twitching motility. Nature *407*, 98-102.
658 Mohan, S., Aghion, J., Guillen, N., and Dubnau, D. (1989). Molecular cloning and
659 characterization of *comC*, a late competence gene of *Bacillus subtilis*. J Bacteriol
660 *171*, 6043-6051.
661 Muschiol, S., Erlendsson, S., Aschtgen, M.S., Oliveira, V., Schmieder, P., de
662 Lichtenberg, C., Teilum, K., Boesen, T., Akbey, U., and Henriques-Normark, B.

663 (2017). Structure of the competence pilus major pilin ComGC in *Streptococcus*
664 *pneumoniae*. J Biol Chem *292*, 14134-14146.
665 Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a
666 fast and effective stochastic algorithm for estimating maximum-likelihood
667 phylogenies. Mol Biol Evol *32*, 268-274.
668 Orekhov, V.Y., and Jaravine, V.A. (2011). Analysis of non-uniformly sampled spectra
669 with multi-dimensional decomposition. Prog Nucl Mag Res Spect *59*, 271-292.
670 Pelicic, V. (2019). Monoderm bacteria: the new frontier for type IV pilus biology. Mol
671 Microbiol *112*, 1674-1683.
672 Reardon, P.N., and Mueller, K.T. (2013). Structure of the type IVa major pilin from
673 the electrically conductive bacterial nanowires of *Geobacter sulfurreducens*. J Biol
674 Chem *288*, 29260-29266.
675 Schwieters, C.D., Kuszewski, J.J., and Clore, G.M. (2006). Using Xplor-NIH for NMR
676 molecular structure determination. Prog Nucl Mag Res Spect *48*, 47-62.
677 Shen, Y., Delaglio, F., Cornilescu, G., and Bax, A. (2009). TALOS+: a hybrid method
678 for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol
679 NMR *44*, 213-223.
680 Theobald, D.L., and Wuttke, D.S. (2008). Accurate structural correlations from
681 maximum likelihood superpositions. PLoS Comput Biol *4*, e43.
682 Wang, F., Coureuil, M., Osinski, T., Orlova, A., Altindal, T., Gesbert, G., Nassif, X.,
683 Egelman, E.H., and Craig, L. (2017). Cryoelectron microscopy reconstructions of the
684 *Pseudomonas aeruginosa* and *Neisseria gonorrhoeae* type IV pili at sub-nanometer
685 resolution. Structure *25*, 1423-1435.
686 Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R.,
687 Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L.*, et al.* (2018). SWISS-MODEL:
688 homology modelling of protein structures and complexes. Nucleic Acids Res *46*,
689 W296-W303.

690 **LEGENDS TO FIGURES**

691

692 **Fig. 1. Com (pseudo)pilus machinery in *S. sanguinis*.** (**A**) Genomic organisation
693 of the genes involved in the biogenesis of the Com (pseudo)pilus in *S. sanguinis*
694 2908. All the genes are drawn to scale, with the scale bar representing 500 bp. The
695 names of the corresponding proteins are listed at the bottom. (**B**) Sequence
696 alignment of the putative N-terminal class III signal peptides of the five ComG pilins
697 in *S. sanguinis* 2908. The 8-15 aa-long leader peptides, which contain a majority of
698 hydrophilic (shaded in grey) or neutral (no shading) residues, end with a conserved
699 Ala_{-1}. Leader peptides are processed (indicated by the vertical arrow) by the prepilin
700 peptidase ComC. The mature proteins start with a tract of 21 predominantly
701 hydrophobic residues (shaded in black), which invariably form the protruding N-
702 terminal portion of an extended α-helix that is the main assembly interface within
703 filaments.

704

705 **Fig. 2. Global sequence analysis of ComGC pilins.** Sequence alignments of
706 ComGC in *S. sanguinis* and *S. pneumoniae* is represented in the top two rows.
707 Residues were shaded in black (identical), grey (conserved) or unshaded (different).
708 The leader peptide is highlighted. In the recombinant proteins that were produced for
709 structure determination, the N-terminal 22 residues invariably forming a protruding
710 hydrophobic α-helix were truncated (depicted by an arrow) to promote solubility. The
711 2D structural motifs predicted using JPred are depicted in the third row. Fourth and
712 fifth rows represent the 80 and 90% ComGC consensus sequences, computed from
713 2,809 ComGC entries in InterPro, and aligned to $ComGC_{SS}$ and $ComGC_{SP}$. Multiple
714 alignments were generated using Clustal Omega and formatted with MView. Polar:
715 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,
716 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:
717 A, C, D, G, N, P, S, T or V.

718

719 **Fig. 3. Rooted phylogeny of the major pilins from various bacterial Tff.** The tree
720 was build using IQ-Tree, with 1,000 replicates of UFBoot and LG+F+R4 model.
721 Numeric values (in %) indicate UFBoot of the corresponding branches. The colour of
722 the bullet points indicates the taxonomic group of the corresponding species. The
723 colour of the strips and highlights indicate the classification of the different Tff
724 systems. Tfpa: type IVa pilus. Tfpb: type IVb pilus. Tfpc: type IVc pilus (also known
725 as Tad). MSH: mannose-sensitive hemagglutinin pilus. Com: competence
726 (pseudo)pilus. T2SS: type II secretion system pseudopilus.

727

728 **Fig. 4. 3D solution structure of $ComGC_{SS}$.** (**A**) Cartoon representation of the
729 $ComGC_{SS}$ structure: face and side views are shown. A dimmed surface
730 representation of the protein is superimposed. The three consecutive α-helices have
731 been named α1, α2 and α3, and highlighted in blue (α1) or cyan (α2 and α3). (**B**)
732 Cartoon representation of the superposition of the ensemble of 10 $ComGC_{SS}$
733 structures determined by NMR, which highlights that there is no significant flexibility
734 in the structure except for the unstructured N-terminus.

735

736 **Fig. 5. 3D solution structure of $ComGC_{SP}$.** (**A**) Cartoon representation of the
737 $ComGC_{SP}$ structure: face and side views are shown. A dimmed surface
738 representation of the protein is superimposed. Nomenclature and colour scheme are
739 the same than in Fig. 4. (**B**) Cartoon representation of the superposition of the
740 ensemble of 10 $ComGC_{SP}$ structures determined by NMR, (**C**) Cartoon
741 representation of the overlay of $ComGC_{SP}$ and $ComGC_{SS}$ representative structures.
742 This highlights the high structural similarity between the two proteins, with 1.54 Å
743 RMSD for the helical regions.

744

745 **Fig. 6. Conserved residues contributing to the globular fold of ComGC.** Cartoon
746 representation of the ordered portion of $ComGC_{SS}$, where residues determined to be
747 on the interior using GETAREA - with accessible surface accessibility ratios of less
748 than 20% - are highlighted in orange. The consensus residues Val_{43}, Gln_{46}, Tyr_{50},
749 Leu_{64} and Ile_{70} are shown with space filling representation.

750

751 **Fig. 7. ComGC display a novel type IV pilin fold.** 3D structure of the three different
752 structural types of type IV pilins identified so far. The canonical type IV pilin fold is
753 represented by the major pilin of Tfpa in *N. gonorrhoeae* (PDB 2PIL). *G.*
754 *sulfurreducens* Tfpa pilin (PDB 2M7G) is the representative member of the very short
755 pilins almost exclusively consisting of α1. The full-length 3D structure of $ComG_{SS}$ has
756 been modelled. The conserved α1 in all three sub-types is highlighted in blue.
757 Distinctive structural features in the globular heads of PilE (antiparallel β-sheet) and
758 ComGC (antiparallel α2-α3 orthogonal to α1) have been highlighted in cyan.

759

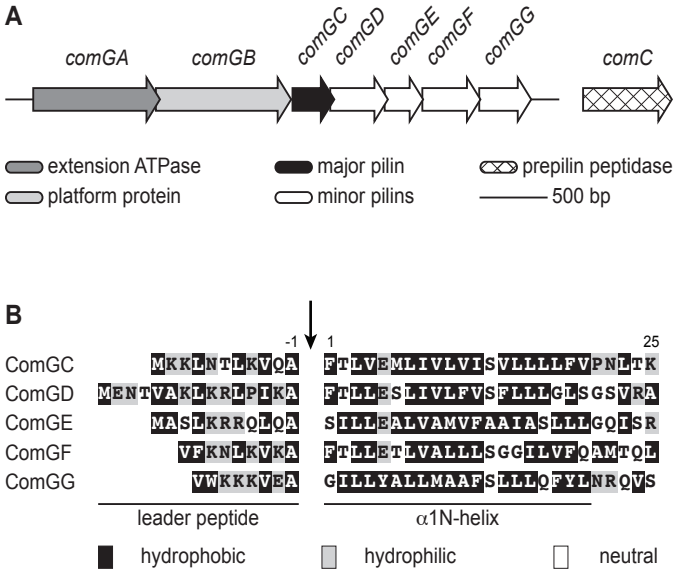
760 **Fig. 8 3D model of ComGC filaments.** The cryo-EM structure of the *K. oxytoca*
761 PulG T2SS pseudopilus (PDB 5WDA) has been used as a template to generate
762 models of $ComGC_{SS}$ (pseudo)pili. (**A**) Full-length $ComGC_{SS}$ in filaments with a melted
763 segment in α1N and an N-terminal methyl-Phe_1. (**B**) $ComGC_{SS}$ (pseudo)pili with a
764 right-handed helical packing of the conserved α1-helices which run approximately
765 parallel to each other in the filament core. (**C**) Top and bottom views of $ComGC_{SS}$
766 (pseudo)pili highlighting the extensive interactions between α1-helices and the
767 globular heads forming the outer surface of the filaments.

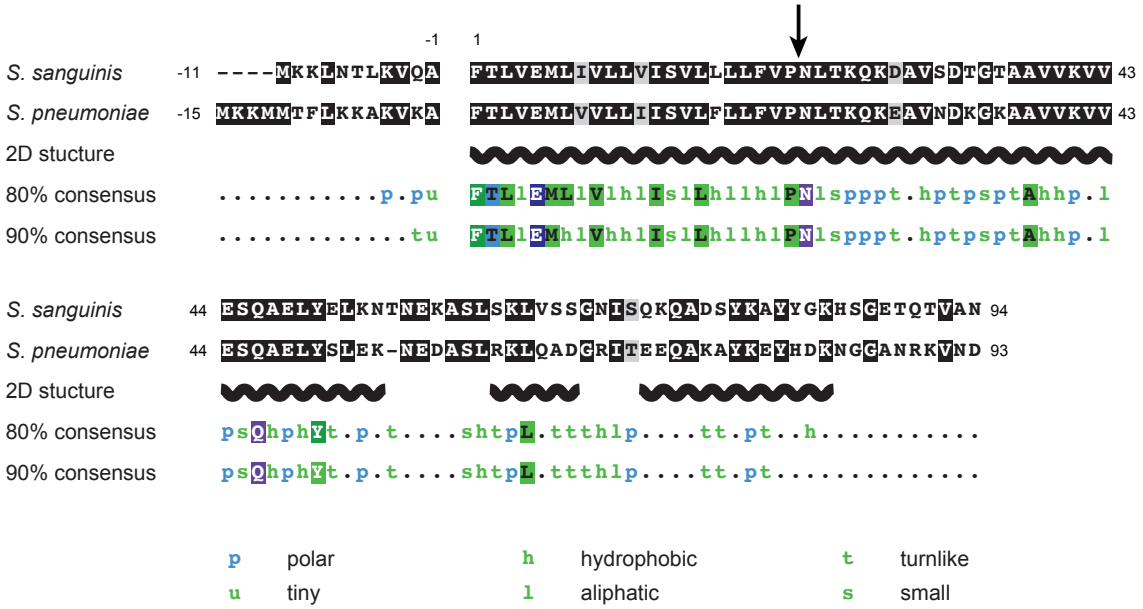
768 **Table 1. NMR structural statistics.**

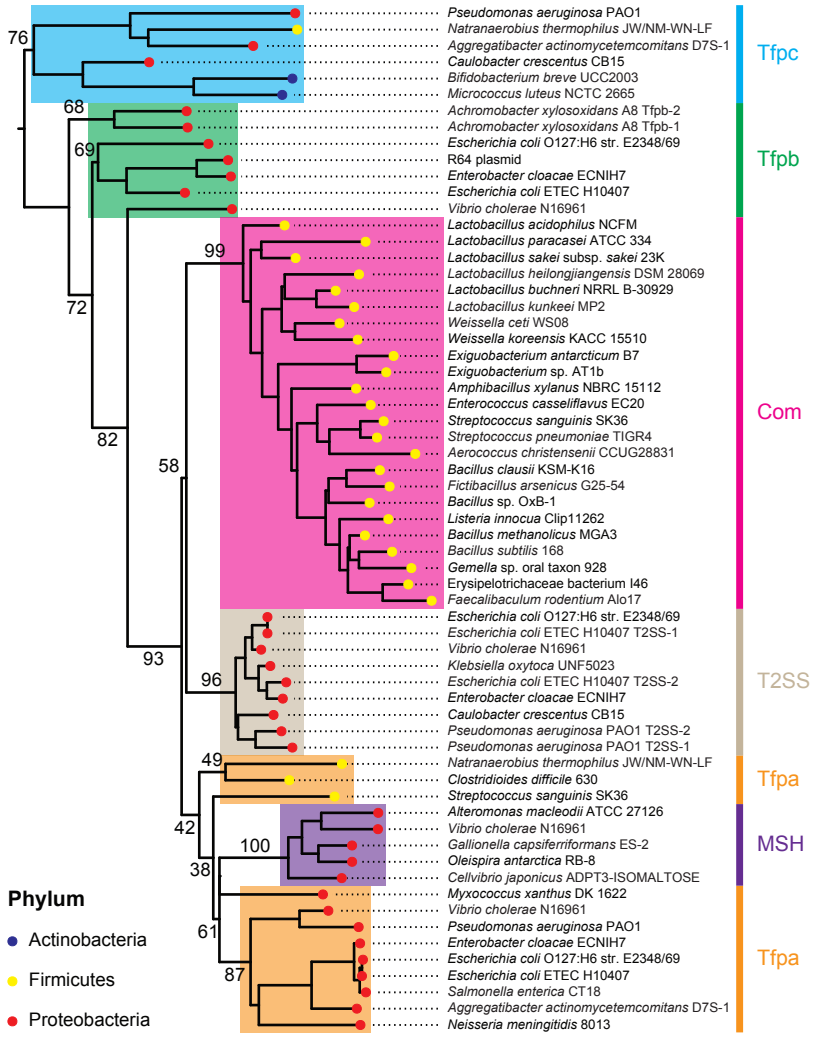
	$ComGC_{SS}$	$ComGC_{SP}$
NOE-derived distance constraints		
long [(i-j) > 5]	128	95
medium [5 ≥ (i-j) > 1]	414	404
intraresidue (i=j)	420	381
total	962	880
hydrogen bonds	50	54
dihedral constraints (Φ and Ψ)	110	102
residual dipolar couplings (RDC)	39	38
Ramachandran statistics (from PROCHECK)		
most favoured (%)	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*)	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

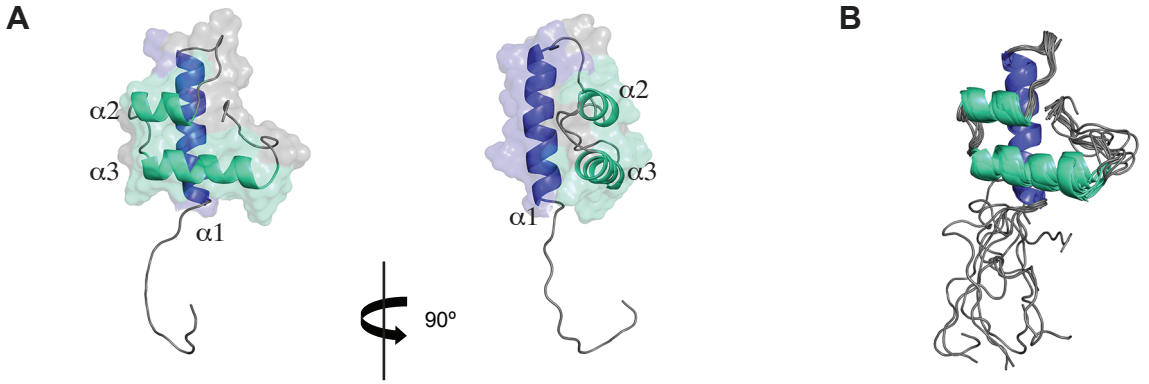
769 *PROCHECK ordered residues are 37-53, 61-66 and 72-85 for $ComGC_{SS}$, and 36-

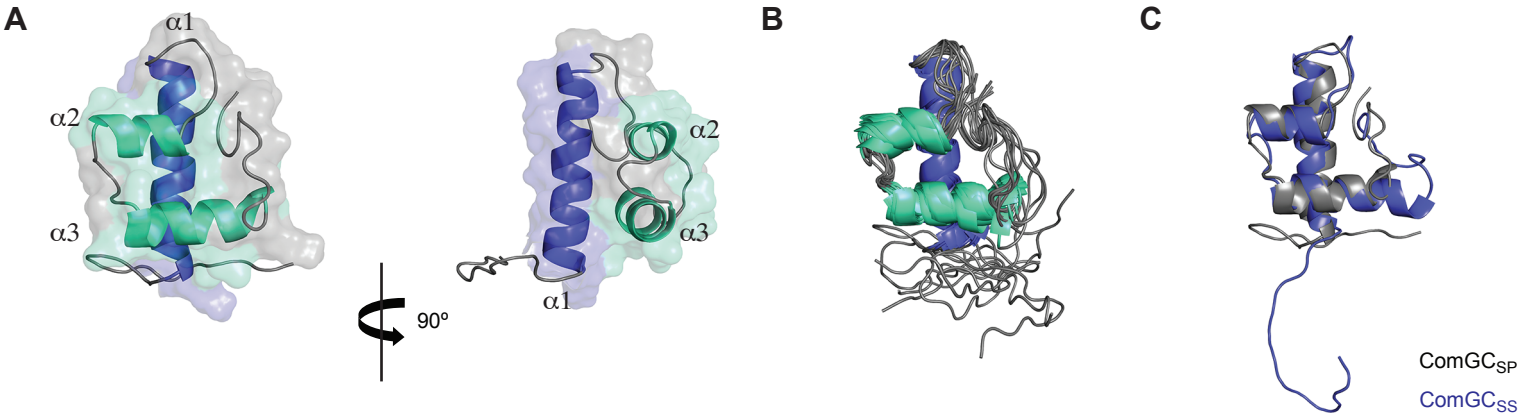
770 54, 60-65 and 71-82 for $ComGC_{SP}$.

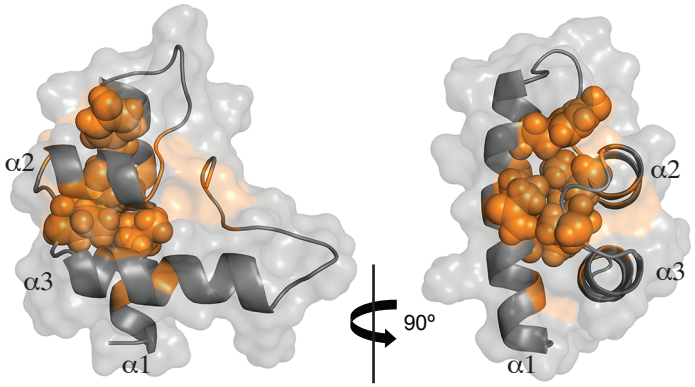

A
comGA
comGB
comGC
comGD
comGE
comGF
comGG
comC
extension ATPase
platform protein
major pilin
minor pilins
prepilin peptidase
500 bp
B
-1
1
25
ComGC MKKLNTLKVQA FTLVEMLIVLVISVLLLLFVPNLTK
ComGD MENTVAKLKRLPIKA FTLLESLIVLFVSFLLLGLSGSVRA
ComGE MASLKRRQLQA SILLEALVAMVFAAIASLLLGQISR
ComGF VFKNLKVKA FTLLETLVALLLSGGILVFQAMTQL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


-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


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
Tfpc
Tfpb
Com
T2SS
Tfpa
MSH
Tfpa
76
68
69
72
82
99
58
93
96
49
42
100
38
61
87
Phylum
Actinobacteria
Firmicutes
Proteobacteria


A
α2
α3
α1
90°
α2
α3
α1
B

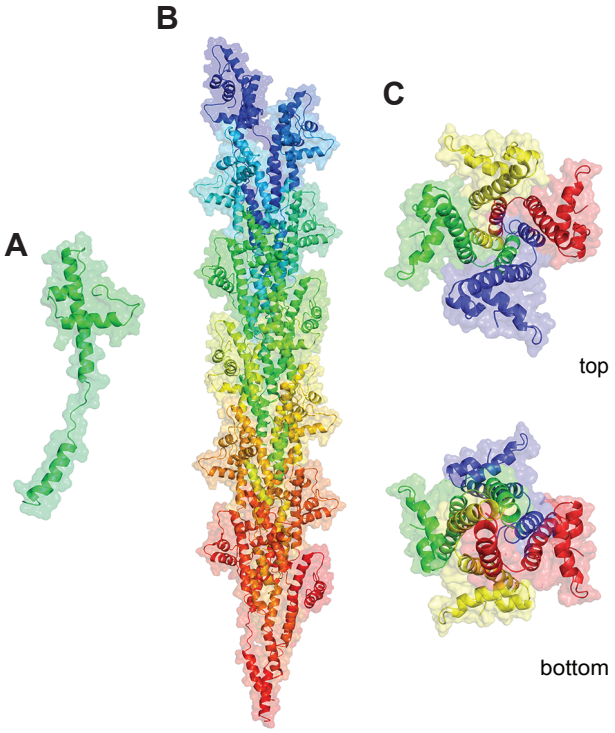

A
α1
α2
α3
90°
α2
α3
α1
B
C
ComGC$_{SP}$
ComGC$_{SS}$


α2
α3
α1
90º
α2
α3
α1

N. gonorrhoeae
PilE

G. sulfurreducens
PilE

S. sanguinis
ComGC


A
B
C
top
bottom