Same, Same, but Different*:*

Molecular Analyses of *Streptococcus pneumoniae* Immune Evasion Proteins Identifies new Domains and Reveals Structural Differences between PspC and Hic Variants

Shanshan Du[1], Claudia Vilhena[1], Samantha King[2,3], Alfredo Sahagun[1,4], Sven Hammerschmidt[5] Christine Skerka[1] and Peter F. Zipfel[16]

[1] Department of Infection Biology, Leibniz Institute for Natural Product Research and Infection Biology, Jena, Germany.

[2] Center for Microbial Pathogenesis, Abigail Wexner Research Institute at Nationwide Children's Hospital, Columbus, Ohio, United States of America.

[3]Department of Pediatrics, The Ohio State University, Columbus, Ohio, United States of America.

[4] Molecular Immunology Laboratory, Department of Microbiology and Immunology, Faculty of Veterinary Medicine and Animal Husbandry, National Autonomous University of Mexico, Mexico City, Mexico.

[5] Department of Molecular Genetics and Infection Biology, Interfaculty Institute for Genetics and Functional Genomics, Center for Functional Genomics of Microbes, University of Greifswald, Greifswald, Germany.

[6] Institute of Microbiology, Friedrich-Schiller-University, Jena, Germany

running title: Variations of S. pneumoniae PspC and Hic proteins

Keywords: Immune evasion, complement escape, Factor H binding, pIg receptor, plasminogen.

Abstract

PspC and Hic proteins of *Streptococcus pneumoniae* are some of the most variable microbial immune evasion proteins identified to date. Due to structural similarities and conserved binding profiles it was assumed over a long time that these pneumococcal surface proteins represent a protein family, comprising eleven subgroups. Recently, however, by evaluating more proteins larger diversity of individual proteins became apparent. In contrast to previous assumptions a pattern evaluation of six PspC and five Hic variants, each representing one of the previously defined subgroups, revealed distinct structural and likely functionally regions of the proteins, and identified nine new domains and new domain alternates. Several domains are unique to PspC and Hic variants, while other domains are shared with other *S. pneumoniae* and bacterial virulent determinants. This understanding improved pattern evaluation on the level of full-length proteins, allowed a sequence comparison on the domain level and furthermore identified domains with a modular composition. This novel concept allows a better characterization of variability, and modular domain composition of individual proteins, enables a structural and functional characterization at the domain level and furthermore shows substantial structural differences between PspC and Hic proteins. Such knowledge will also be useful for molecular strain typing, characterizing PspC and Hic proteins from new clinical *S. pneumoniae* strains, including those derived from patients who present with pneumococcal hemolytic uremic syndrome. Furthermore this analysis explains the role of multifaceted intact PspC and Hic proteins in pathogen host interactions. and can provide a basis for rational vaccine design.

Author Summary

The human pathobiont *Streptococcus pneumoniae* expresses highly polymorphic PspC or Hic proteins, which bind a repertoire of host immune regulators and combine antigenic variation with conserved immune evasion features. Understanding domain composition of each protein encoded by more than 60 000 *pspC* or *hic* genes deposited in the data banks defines their diversity, a role in immune escape and can furthermore delineate structure function approach for single protein domains. PspC and Hic proteins show variable domain composition and sequence diversity, which explain differences in binding of human regulators and likely in immune escape. The results of our analyses provide insights in the domain composition of these diverse immune evasion proteins, identifies new domains, defines domains which are unique to PspC or Hic variants, and identifies domains which are shared with other bacterial immune evasion proteins. These data have implication on cell wall attachment, surface distribution and in immune escape.

Introduction

The pathobiont *Streptococcus pneumoniae.* *Streptococcus pneumoniae* (the pneumococcus) is the major cause of community-acquired pneumonia. In addition, this human pathogenic Gram-positive bacterium can cause otitis media and may also cause acute life-threatening invasive infections such as meningitis and even sepsis (1-4). Malnutrition and *S. pneumoniae* infections are the major cause of childhood mortality worldwide. Pneumonia account for approximately 16 percent of the 5.6 million of deaths among children under five years old, killing around 808,000 children in 2016 according to the United Nations Children's Fund (UNICEF) and the World Health Organization (WHO)(5-7). At any point in time pneumococci can reside asymptomatically in the upper respiratory tract of about 50% of children, from where they can be transmitted to other persons. Based on differences of the polysaccharide capsule so far 97 serotypes are identified (8).

Pneumococcal diseases are widespread both in developing and developed countries and antibiotic resistant strains are arising. The increase in microbial resistance to antibiotics makes it important to identify new virulence determinants, to understand the diversity of these determinants and also to define the immune escape strategies of this relevant pathogenic bacterium (9-11). In addition, vaccines with higher serotype coverage or serotype-independent vaccines are needed in order to combat the pathogen.

Immune and in particular complement evasion is critical for *S. pneumoniae* and for all the ability of all human pathogenic microbes to cause infections. Common patterns regarding complement evasion and binding of human complement and immune regulators are emerging (12-16). Thus, it is important to understand the

exact role of individual pneumococcal virulence determinants, in particular their role of complement evasion, the topology of the capsule and surface location of virulence determinants (17-19).

PspC and Hic represent related *S. pneumoniae* surface proteins. PspC and Hic proteins are important pneumococcal immune evasion proteins and adhesins and represent promising vaccine candidates (20). The majority of virulent *S. pneumoniae* strains express at least one Psp or Hic variant, and strains that have the *pspC/hic* genes deleted show significant amelioration of lung infection in mice, nasopharyngeal colonization, and bacteremia (21).

PspC and Hic proteins as central pneumococcal immune evasion proteins. Initially, PspC was identified as an adhesin, which targets the secretory component of the secreted Immunoglobulin A (sIgA) and polymeric IgA receptor (pIgR)(22). Because *pspC* and *hic* genes were identified independently by several groups, different names were originally given, including CbpA (choline-binding protein A), SpsA (secretory IgA binding protein), PbcA (C3-binding protein A), or Hic (Factor H binding inhibitor of complement)(23-33).

PspC and Hic proteins are attached to the bacterial cell wall, and are surface-exposed adhesins and immune evasion proteins. PspC proteins with their C-terminal choline-binding anchors attach non-covalently to the phosphorylcoline (PCho) moiety of the teichoic acids (TAs) and Hic proteins, displaying a C-terminal LPsTG motif, are covalently linked to the peptidoglycan via the sortase A. This suggests that proteins attach with different strengths and likely also to a distinct depth in the cell wall, which in turn may influence surface distribution. The fact that both proteins anchor via the

C-terminal region suggests that the preceding part of the protein spans the capsular polysaccharides and that the N-terminal part is extending beyond the capsule.

As central immune evasion proteins, PspC and Hic proteins bind several human plasma proteins including Factor H, C3, C4BP, Plasminogen, thrombospondin-1, and vitronectin (22-36). These multifunctional proteins represent one of the most diverse group of immune evasion and adhesive proteins recognized to date (35,36). PspC and Hic proteins have *a* mosaic structure, comprising distinct regions, consisting of multiple domains. Furthermore a substantial overlap of domains exist between PspC and Hic variants. Standard domain or sequence-based comparison among members of this protein family is complex due to structural differences and variable domain composition. Currently, the protein NCBI databank lists 54909 entries for PspC or Hic and 11817 entries for CbpA, encoding both full-length proteins and partial protein sequences (march 06, 2020; NCBI www.ncbi.nlm.nih.gov/protein). The individual entries show homology, but also exhibit considerable variations in structure and sequence. Single PspC and Hic proteins show variable domain patterns, different variants of these proteins combine domains in different ways, and apparently not all domains are identified so far.

***Mosaic-structured PspC and Hic proteins*.** Our understanding of these important pneumococcal immune evasion proteins is currently still fragmented. Thus, defining the exact domain composition of individual PspC and Hic variants, or to correlate phenotypes with disease forms, is important for better understanding the role of each protein, for structural predictions, for localizing binding sites of host ligands, for understanding precise domain function(s), and for characterizing strain specific differences.

Based on overall sequence similarities PspC and Hic variants were initially considered to belong to one group of pneumococcal immune evasion proteins. Initial analyses by Brooks Walter in 1999 and Iannelli et al. in 2001 revealed both sequence similarity and diversity among PspC and Hic proteins (37,38). Ianelli et al. identified several domains for the evaluated 43 PspC and Hic proteins including the leader peptide, α-helical regions with a seven-amino acid periodicity, repeat domains, a proline-rich stretch followed by either a choline-binding or sortase-dependent anchor (38). At that time, the cell wall anchors were used as criterion to differentiate between PspC and Hic family proteins and based on sequence differences six PspC-type and five Hic-type clusters were defined. However, still today no precise criteria exist regarding cluster specific domain composition or domain characteristics. Because the domain patterns as well as borders of single domains are not well-defined, a straightforward variant designation e.g. of existing but also of newly identified *pspC and hic* genes, or genes from novel clinical pneumococcal isolates, is difficult or even impossible (39).

***Aim of the study*:** Thus far, the internal or external position of each domain remains unclear, as do the precise borders of the regions and of the domains. We do not exactly know which domain(s) are indeed integrated into the bacterial cell wall, which domain(s) span the capsule, or which domains are externally positioned. Given these limitations, and the heterogeneity among the proteins, we aimed to evaluate the structure and domain composition of six PspC and five Hic variants, each representing one of the clusters defined by Ianelli *et al*. We further aimed to obtain evidence on domain composition and position. By evaluating the domain pattern of each variant, we identified nine new domains, illustrate structural as well as

compositional differences between the full length proteins, between N and C-terminal regions, and between PspC and Hic proteins. Furthermore this comparison also identified subvariants on the domain level.

Results

Global Similarity of PspC and Hic Variant Proteins

Selection of PspC and Hic variants. One protein from each variant cluster as defined by Ianelli et al. was selected (38). The variants included the six PspC variants, i.e. PspC1.1, PspC2.2, PspC3.1, PspC4.2, PspC5.1, PspC6.1, and five Hic variants, Hic/PspC7.1, Hic/PspC8.1, Hic/PspC9.1, Hic/PspC10.1, Hic/PspC11.1. At the date of the cluster designation Ianelli *et al.* considered the PspC and Hic variants as one protein family and used a PspC nomenclature for both protein groups (38). To appreciate the Hic type character and at the same time follow the nomenclature suggested by Ianelli *et al* we combine the Hic and PspC designations **(Figure 1A)**. The selected proteins vary in size and mass, with PspC1.1 as the largest protein having a length of 929 aa and a molecular mass of 110 kDa, while Hic/PspC8.1 is the smallest protein with a length of 503 aa and a mass of 65 kDa (Supplementary Table I). When compared to the well-characterized PspC3.1 protein (strain D39), the overall sequence amino acid identity of the six PspC proteins ranged from 51 to 82%. In contrast, the five Hic variants showed a less pronounced identity which ranged from 15 to 26%. Thus suggesting functional differences between the PspC and Hic variants **(Figure 1B).**

PspC3.1 as a prototype PspC. PspC3.1 was selected as prototype and used for analyzing structure and domain composition. PspC3.1 has a signal peptide that directs the protein to export. The protein has an externally oriented N-terminal region and is integrated into the teichoic acids of the bacterial cell wall via the C-terminal Choline-Binding Domain. Because different regions of these membrane anchored proteins are facing different environments, we hypothesized, that hydrophilic and hydrophobic surroundings, could influence protein structure and composition.

Structure and residue composition of PspC3.1. PspC3.1, when evaluated *in silico,* showed three clearly different structural regions. The N-terminal 410 residues form mostly α–helices, followed by a 70 aa long predominately coiled-coil region and a 221 aa long region composed mainly of β-sheets (**Figure 2A).** Given these structural differences the 410 aa long mainly α–helical region was designated as N-terminal region, and the remaining part with the coiled-coil and β-sheet segments and almost lacking α–helices was termed C-terminal region. When the structural regions were aligned with the known domains of PspC3.1, the N-terminal α–helical region included the signal peptide, the Hypervariable Domain, the two Repeat Domains, and the Random Coil Domain. The Hypervariable Domain includes the binding sites for human Factor H and each Repeat Domain includes a binding site for sIgA/polymeric Ig receptor, which is in agreement with an external, orientation. The mostly coiled-coil structured region represents the Proline-Rich Domain (aa 411-482), which is considered a cell wall-spanning and flexible domain and the β-sheet region represented the Choline-Binding Domain (aa 483-701) used for cell wall attachment **(Figure 2B**). The C-terminal Proline-Rich Domain and the Choline-Binding Domain have an inside location (40, 41).

Amino acid composition. Next we evaluated if the proposed outside and inside environments influence the protein make up. The N-terminal region of PspC3.1 includes 45.3% charged, 18.0% polar and amphipathic residues and has a low fraction of Tyrosines (1,7**%**). The C-terminal region in contrast includes only 15.0% charged residues, has 3an increase percentage or (9.5%) polar and amphipathic amino acids (9.5%) and many Tyrosines (8.9%) (**Figure 2C).** Thus the N-terminal and C-terminal regions of PspC3.1 differ in structure, and amino acid composition.

The differences of the N and C-terminal regions are conserved in the PspC and Hic variants. Next we evaluated if the structural composition, as outlined for PspC3.1, is conserved in the other PspC and Hic variants. The N-terminal regions of all analyzed PspC and Hic variants have mostly α-helical structrues, and the C-terminal Proline-Rich Domains have predominantly coiled-coil structures. The PspC specific Choline-Binding Domains have mostly β–sheets, and the Hic specific LPsTG anchors have an α–helical segment following a coiled-coil stretch (**Supplementary Figures 1 and 2**).

In addition the amino acid composition was determined. The N-terminal regions of the six PspC variants contained 35-45% charged residues. In contrast their C-terminal regions contained 16% or less charged residues. The C-terminal domain also included more polar and amphipathic amino acids (32-36%) and were rich in Tyrosine (8.3-9.8%)(**Figure 3A**), The Hic variants contained 28-37% charged residues in their N-terminal regions, and their C-terminal regions had a high fraction of charged (28-41%) and less polar/amphipathic residues (15-21%) than the PspC variants (**Figure 3A)**. Thus, the N and C-terminal regions of the proteins differ in structure and amino acid composition and the C-terminal regions of the PspC and Hic proteins show differences in amino acid composition.

The N-terminal regions of the different variants ranged in length from **146** (Hic/PspC8.1) to **633** (PspC5.1) residues. A homology alignment of the N-terminal regions showed two distinct clusters. One N-terminal cluster included five PspC variants (PspC1.1, PspC6.1, PspC2.2, PspC5.1, PspC3.1) and the Hic/PspC11.1 variant while the second N-terminal panel included PspC4.2 and four Hic variants (Hic/PspC7.1, HicPspC9.1, Hic/PspC10.1, Hic/PspC8.1)(**Figure 3B, upper panel).**

The C-terminal regions were more conserved in length, ranging from 236 (PspC5.1) to 348 aa (Hic/PspC8.1) and were clearly separated the PspC and the Hic members. The level of diversity between the C-terminal regions of variants within each group was low indicating that these domains are more highly conserved (**Figure 3B, lower panel).**

***Domain analyses of PspC and Hic variants*.** Using PspC3.1 with its five known domains as a blue print, the domain patterns of the other ten cluster variants was evaluated. This approach determined that three domains of PspC3.1, the signal peptide, the N-terminal Hypervariable Domain and Proline-Rich Domains are found in all PspC and Hic variants. All PspC variants use a Choline-Binding Domain, while Hic/PspC proteins have an LPsTG anchor (**Figure 1 and Figure 4**). Repeat Domains and the Random Coil Domain are found mainly in PspC proteins, but not in all variants. Additional sequence stretches were identified in some PspC or Hic variants that did not match with known domains of PspC3.1. These domains were evaluated separately to determine whether counterparts exist in other PspC and Hic variants or whether homologs exist in the protein data bank. This approach identified nine new domains, including one new domain in PspC3.1 and also three new sub variants of the Proline Rich Domain. This extended domain scenario shows that the individual PspC and Hic proteins harbour variable domain numbers, ranging from four (Hic/PspC8.1) to ten domains (PspC4.2**)(Figure 4).**

***Known domains of the N-terminal region*.** The known domains identified in the N-terminal region include:

***Signal peptide*.** A highly-conserved 37 aa long N-terminal signal sequence which directs the proteins for export and is cleaved upon processing is present in all PspC and Hic/PspC variants (**Supplementary Figure 3A**).

***Hypervariable domains*.** Mature PspC and Hic/PspC proteins expose N-terminal Hypervariable Domains, which are rich in charged residues. The length of the Hypervariable Domains ranged from 91 (PspC4.2) to 113 aa (PspC2.2), and as their name suggests they were highly variable in sequence. Only five residues, $\mathbf{T}_{11}$,$\mathbf{S}_{12}$,$\mathbf{I}_{59}$,$\mathbf{Y}_{63}$,$\mathbf{K}_{96}$ (numbering based on PspC3.1) are present in all proteins although additional residues are conserved in several variants. The N-terminal **Hypervariable Domains** appear to be specific for PspC/Hic protein variants (**Supplementary Figure 3B**) and they include a 12 amino acid long region, which in PspC3.1 was identified as Factor H binding region **(Figure 5A, Supplementary Figure 3C**).

Relationship analysis using a dendrogram identified three subtypes of the hypervariable domains. Subtype A (HVD-A**)** is present in PspC3.1, PspC5.1, and Hic/PspC11.1 HVD-B is present in the PspC2.2, PspC1.1, and PspC4.2, and HVD-C is present in PspC6.1, Hic/PspC7.1, Hic/PspC10.1, Hic/PspC9.1, and Hic/PspC8.1 (**Supplementary Figure 3C**).

***Repeat domains*.** All PspC-type proteins and Hic/PspC7.1 possess approximately 110 aa-long repeat domains (Repeat Domain). Five PspC (i.e. PspC3.1, PspC2.2, PspC6.1, PspC1.1, PspC5.1) possess a second Repeat Domain. The **Repeat Domains** have conserved sequences, they are rich in charged residues, and include conserved RNYPT motifs, which are binding sites for sIgA/pIgR **(Figure 5B, Supplementary Figure 4**). Related repeat domains were identified in PspK from

S. pneumoniae (H2BJK8) with 55 % homology to Repeat Domain1 and 71.6 % homology to Repeat Domain II, respectively. The solution structure of the Repeat Domain of PspC3.4 from strain TIGR has been solved (40). This domain folds into three antiparallel α-helices, and the YPT residues, representing the core sIgA/pIgR binding motif are positioned in a coiled-coil structured loop, which separates helix 1 and helix 2. This experimentally determined structure actually confirms and validates our *in vitro* structure prediction **(Figure 2A**).

***Random Coil Domain*.** Random coil domains are approximately 30 aa-long, show a coiled-coil structure and are relatively conserved in sequence. They are typically positioned downstream of the first Repeat Domain. No homologous were identified in the data bank **(Supplementary Figure 5**).

New Domains of the N-terminal Region

Sequence stretches in the PspC and Hic variants that did not match known domains of PspC3.1 were also identified. These sequences were used to search for counterparts in other PspC and Hic variants or for homologs in the protein data bank. This procedure identified nine new domains, including one new domain in PspC3.1 and also three new alternates of the Proline Rich Domain.

Serine-Rich Elements. Serine-Rich Elements with the overall motif S_nD/GS_2 were detected in five PspC and in all Hic/PspC variants. Nine protein variants harbor one, whereas PspC2.2 contains two; and PspC4.2 lacks such an element. These Serine-rich elements share a coiled-coil structure; but differ in position, type, and in sequence. Ser-rich elements following the Hypervariable Domain (PspC2.2, Hic/PspC7.1, Hic/PspC9.1, Hic/PspC8.1) or the unique Hic/PspC11.1 domain have

the consensus S_nD/GS_2 and are up to 24 aa long. The segments following the Random Coil Domain (PspC3.1, PspC2.2, PspC6.1, PspC1.1, PspC5.1, Hic/PspC10.1) have related S_2DS_2, units, which can be up to 18 aa long. The domain of Hic/PspC10.1 shows a variation to these common features **(Supplementary Figure 6A**). The biological role(s) of these elements are as yet unknown. In engineered proteins, related poly-serine-rich elements are integrated as flexible linkers that separate functional, individually folding domains (41). Interestingly the TKPET motif at the end of segments following the Hypervariable domains are highly related to the first seven residue long units found in Proline Rich Domains III and IV (see below).

Random Coil Extension Domains. Two separate new domains were identified in four proteins, which are positioned downstream of the Random Coil Domain-S_2DS_2 combination.

Random Coil Extension Domain 1. Two proteins, PspC1.1 and PspC5.1, contain a new domain following the Random Coil Domain-S_2DS_2 combination. These 83 aa long domains share almost identical sequences.

This domain includes several charged residues, and shares homology with domains in other proteins. Proteins containing such RICH type domains including secreted proteins such as PspC Q9KK19, SpsA O33742 and IgA Fc receptor binding protein P27951 from *Streptococcus agalactiae*. The domain is predicted to be involved in bacterial adherence or cell wall binding (42)**.**

Random Coil Extension Domain 2. PspC4.2 and Hic/PspC10.1 have different 114 (PspC4.2) or 126 aa-long (Hic/PspC10.1) segments following Random Coil Domain-S_2DS_2. These elements show moderate sequence homology among each other. The 126 aa domain of Hic/PspC10.1 harbors a N-terminal 37 aa extension, with the remaining portion being conserved in the PspC4.2 domain. The biological role of this unique segment is unclear. In PspC4.2 the region includes a long α-helical stretch and is followed by a ca 30 residue long coiled-coil stretch.

PspA-Like Domain. PspC1.1 and PspC5.1 have related, new domains following Repeat Domain II. These 131 or 130 aa-long domains are rich in charged residues, and exhibit 84.5% sequence homology with the A*/B element of *S. pneumoniae* PspA from strain DBL6A, which includes a lactoferrin-binding region (43,44). These data suggest that the newly identified domains in PspC1.1 and PspC5.1 bind lactoferrin (45,46).

PspC4.2 Specific Element. Domain pattern analysis identified an element in PspC4.2 which is positioned between the Hypervariable Domain and the Random Coil Domain. This 33 aa-long α-helical structured segment, lacks homology to other proteins in the databank role, thus its role remains unclear.

Repeat Type Domain. PspC4.2, Hic/PspC7.1, and Hic/PspC10.1 share related 92, 82, or 68 aa-long domains, which are distantly related (41.6% homology) to the Repeat Domains. These new Repeat Type Domain with a mostly α-helical structure intriguingly, lack an RNYPT binding motif and seem to be specific for PspC and Hic proteins.

A New Two Segmented Domain

A new two-domain segment was identified in PspC4.2 and the three Hic proteins, Hic/PspC7.1, Hic/PspC10.1, Hic/PspC9.1.

The upstream domain*.** The 24, 36, 40 or 37 aa-long upstream sections are rich in proline residues (Supplementary Figure 14), have a predicted coiled structure, and due to their location in the N-terminal region are termed ***Extracellular Proline Rich Segments. The high Proline content may suggest a function as chain breaker (47). These External Proline Rich Domains lack homology to other bacterial proteins, and thus seem unique for PspC proteins.

The downstream units show homology to the Fc binding part of protein C from* S. agalactiae*. The 89 or 78 aa long elements are rich in charged residues, lack proline residues, and have an α-helical structure. A blast search revealed 51.1% identity to a segment within the trypsin sensitive beta-antigen of *Streptococcus agalactiae* (strain P27951/Uniprot). This protein binds the Fc region of human IgA likely via two stretches (48). This sequence stretch is found in several bacterial immune evasion proteins. One group of Gram-positive bacteria share in their signal peptides a YSIRK motif. Also based on the many charged residues this domain (pfam05062) is also named RICH (Rich In Charged residues) and is identified in other secreted proteins of *S. pneumoniae* proteins including SpsA and the Fc binding part of human IgA from *Streptococcus agalactiae*. The function is proposed in bacterial adherence or cell wall binding.

***Hic/PspC11 Specific Element*.** Hic/PspC11.1 contains a unique 102 aa-long α-helical structured domain, which follows the Hypervariable Domain. Related

403 segments were identified in most Hic/PspC11 variants, however, not in other
404 pneumococcal or bacterial proteins Thus far, the function of this domain is unknown..
405
406 **Domain Composition of the C-terminal region**
407 The C-terminal regions of the analyzed PspC and Hic proteins are relatively
408 conserved in length (ranging from 237 aa (PspC5.1) to 348 aa (Hic/PspC8.1) and
409 each protein combines a modular Proline-Rich Domain with either the PspC specific
410 Choline-Binding Domain or the Hic characteristic LPsTG anchor (47 - 50). A general
411 pattern is emerging: PspC proteins link shorter Proline-Rich Domains (57 to 77 aa) to
412 longer Choline-Binding Domains (179 to 219 aa), while Hic proteins combine longer,
413 Prolin-Rich Domains (186 to 286 aa) with shorter LPsTG anchors (50 to 62 aa).
414
415 ***Proline-Rich Domains*.** Proline-Rich Domains have a modular structure and connect
416 the N-terminal region with the cell wall anchor. The proposed role as a bacterial cell
417 wall-spanning domain is consistent with the position prior to the anchor (49,50). Our
418 in silico analysis identified a modular composition and further distinct proline-rich
419 domains, which differ in length (57 to 286 aa), type, module composition, and
420 sequence.
421
422 ***Proline-Rich Domain I*.** Five PspC variants have highly related 59 to 77 aa long
423 domains, forming either three-segmented (PspC3.1, PspC6.1, PspC2.2), or two-
424 segmented domains (PspC1.1, PspC5.1)(**Supplementary Figure 7A**). The first N-
425 terminal segments have Proline dominated PAPA- and PAPAP motifs, and can be up
426 to 46 aa long. The C-terminal segments include PAPAP or PAPTP-forming motifs,
427 are up to 19 aa long, and have a coiled-coil structure. The middle segment found
428 only in the three segmented domains is conserved in length (23 aa), in sequence,

exhibits characteristic flanking Q-residues, and is rich in charged residues. Interestingly this segment has a predicted α-helical structure and lacks Prolines. Such Proline-Rich segments are also found in PspA (50,51).

Proline-rich domain II. PspC4.2 has a unique 57 aa-long Proline-Rich Domain. This new domain includes 19 Prolines and has an internal repeated segment with the sequence TPQV**P**K**P**EA**P**K. To date, this new domain has been identified only in PspC proteins)(**Supplementary Figure 7B**).

Proline-rich domain III. Hic/PspC7.1 harbors a unique 186 aa-long Proline-Rich Domain which includes an N-terminal 7 aa element followed by five almost identical 31 aa long repeats (KK**P**SA**P**K**P(**G/D)MQ**P**S**P**Q**P**EGKK**P**SV**P**AQ**P**GTED). Each repeat has **nine** Prolines and two KKPS(A/V)P motifs (denoted by white letters). The 31 aa repeats are followed by a truncated 24 aa-long repeat element (**Supplementary Figure 7C, D**).

Proline-rich domain IV. Four Hic variants harbor 247 to 286 aa long, Proline-Rich Domains containing 23, 19, or 26 modules. The modules vary in type and sequence, including multiple 11 aa modular repeats, which are followed by one truncated repeat, and a 16 aa long extension (**Supplementary Figure 8A,8B,8C**). Hic/PspC10.1 and Hic/PspC9.1 contain 14 and 16 modular units, respectively with the sequence (L/P)E**K**PKPEVKP**Q**. Hic/PspC8.1 and Hic/PspC11.1 contain 23 copies of (L/P)E**T**PKPEVKP**E** elements (variant residues in white letters on black background). They are followed by one shortened module and have distal nearly identical 16 aa-long C-terminal units, which at position 15 show a **T/P** variation)(**Supplementary Figure 8D, 8E, 8F**).

455

456 ***Cell wall attachment***. PspC proteins use longer modular Choline-Binding Domains
457 for cell wall attachment and by contrast, Hic proteins have shorter, 50–62 aa-long
458 anchors that include a sortase-dependent LPsTG motif (53,54).

459

460 **PspC-type protein variants possess choline-binding anchors.** PspC type,
461 variants have modular C-terminal Choline-Binding Domains and their length ranges
462 from 178 (PspC5.1) to 248 aa (PspC1.1). Most modules are 20 aa and apparently
463 two units can attach to one choline component (52). Related Choline-Binding
464 Domains are found in up to 15 other *S. pneumoniae* proteins, including the immune
465 evasion protein PspA, the autolysins LytA, LytB LytC, and CbpL (52). In the literature
466 these modular composed Choline-Binding Domains are sometimes termed choline-
467 binding modules. However given the domain composition of full length PspC and Hic
468 variants, we prefer to term such smaller, repetitively assembled subunits as modules.
469 Apparently both PspC and Hic variants use modular composited domains within their
470 C-terminal putatively interior regions.

471

472 ***Hic variants have C-terminal sortase signals.*** The five analyzed Hic variants share
473 C-terminal 50–62 aa anchors which display a specific pentapeptide LPsTG motifs.
474 The transpeptidase, sortase A cleaves within this conserved motif between Thr and
475 Gly, and subsequently the protein is covalently linked via the Thr to lipid II (P3
476 precursor) and a penicillin binding protein (55, 56)(**Figure 5E**). The domain
477 distribution of related PspC and Hic variants show differences which are indicative for
478 extra and intracellular positions, and reveal diverse structural composition among
479 PspC and Hic variant proteins.

Discussion

This analysis of domains within the six PspC and five Hic Hic variants identified 13 N-terminal and three C-terminal domains, including the seven known and nine new domains, and furthermore recognized three new alternates of the Proline-Rich Domain. The mature PspC and the Hic proteins are heterogeneous proteins. They generate intriguing diversity by combining different domains, by varying domain types, and by assembling domains in different numbers. Domain variability is increased by assembling variant modular elements and by sequence variations. This reflects antigenic variation, functional specialization and furthermore the different anchors are indicative for a different surface distribution (17,18). Three domains, the Signal peptides, the Hypervariable Domains and Proline-Rich Domains are found in all analyzed variants (Table I). Eleven domains are found in several (but not all) variants, and two domains are unique to single proteins. This extensive domain characterization shows differences between the analyzed PspC and Hic variants, reveals variable domain assembly features, and shows a different composition of the N and C-terminal regions.

Variability among PspC and Hic-variants. PspC, and Hic-type S pneumonia show related domains in their N-terminal regions, but differ more in their C-terminal regions. Also the proteins have different C-terminal anchors. PspC proteins with the Choline-Binding Domains contact multiple choline-moieties in a non-covalent manner. In contrast the LPsTG anchors attach the proteins covalently to the peptidoglycan (54). The type of C-terminal anchor not only influences cell wall attachment, but length and composition of the Proline-Rich Domains and furthermore seems to influence selection, composition, and number of the N-terminal domains.

505 These different domain pattern distributions are indicative for distinct surface
506 positioning and different roles in immune evasion.
507
508 ***Variability of N vs C-terminal regions*.** Broadly speaking, each PspC and Hic
509 protein has two major parts: the N-terminal, outside presented region, which includes
510 immune evasion and adhesion domains, and the C-terminal anchoring region.
511
512 The **N-terminal** regions of the analyzed PspC and Hic proteins vary in length,
513 and domain number ranging from 155 aa with two domains (Hic/PspC8.1) to 610 aa
514 with eight domains (PspC4.2). These regions share structural features, including long
515 α-helical structures, and a high proportion of charged residues. The Hypervariable
516 Regions are located most distant from the cell surface and show the highest degree
517 of variation. This diversity reflects differences in immune control and antigenic
518 variability, which is relevant for evading immune recognition by antibodies. Six of the
519 N-terminal domains are unique to PspC and Hic variants, others like the PspA
520 Related Domain and the region with homology to the IgA binding β antigen are found
521 in other pneumococcal or bacterial immune evasion proteins.
522
523 **The C-terminal** regions are distinct from the N-terminal regions. They are more
524 conserved in length, ranging from 236 aa (PspC5.1) to 348 aa (Hic/PspC8.1), have
525 more polar and amphipathic residues and PspC proteins have also more Tyr
526 residues. The Proline-Rich Domains, preceding the PspC and Hic-specific anchors,
527 show a modular composition, have mostly coiled-coil structures and differ in length.
528 Proline-Rich Domains of PspC proteins are shorter than those of Hic proteins. Given
529 the proposed location at the interface between cell wall and capsule, such diversity
530 could reflect different binding dynamics, strength of cell wall integration, or

531 morphological differences due to capsule thickness (54-59). Similarly the anchor
532 domains in the C-terminus differ in length, composition, and type of cell wall
533 integration.
534
535 ***Protein orientation, and cell wall integration.*** PspC and Hic are membrane
536 integrated, surface proteins and we are understanding now which parts of the
537 proteins have exterior or interior location. The N-terminal region, by extending from
538 the capsule, is exposed to the outside world and can interact with human proteins
539 (Table II). The C-terminal region includes a capsule spanning segment and an
540 internal cell wall anchor.
541
542 Cell wall attachment via the C-terminal anchor orients the N-terminus to the
543 outside to allow interaction with host plasma proteins and cell receptors. An
544 illustration of the orientation, spatial organization of one PspC and one Hic variant
545 including mapped binding sits for human plasma regulators is presented in **Figure**
546 **5E.** PspC1.1 representing an eight domain choline-attached variant and the short
547 four domain Hic variant (Hic/PspC8.1) show different compositions both in the
548 proposed extra and intracellular regions. Due to variable length the N-terminal
549 regions extend to different distances from the surface. In a linear model, for example,
550 Factor H, when bound via the hypervariable domain inhibits C3b formation and
551 assists in C3b inactivation remote from the bacterial surface. Similarly, the Proline-
552 Rich Domains, due to their variable length and composition, and the specific anchors
553 can integrate the proteins in the cell wall envelope with different depth and strength.
554
555 ***Tactical positioning and immune evasion.*** The two distinct anchors, show
556 different structures, mainly β-sheet composed Choline-Binding Domains vs coiled-

557 coiled and α-helical structured LPSTG anchors. This not only mediates non-covalent
558 vs. covalent attachment, but is also indicative of a more flexible vs. fixed cell wall
559 attachment, for a different surface distribution and probably also exposure to the
560 host. Indeed, for *S. pneumoniae* strain BNH418, a different spatial localization of a
561 PspC and a Hic variant was shown by super resolution microscopy (60). The PspC-
562 protein, with the Choline-Binding Domain localized to the division septum and bound
563 Factor H as a result controlled C3b opsonization. In contrast, the LPsTG anchored
564 Hic protein was localized to the bacterial poles. Such distinct surface localization can
565 apparently influence the site of complement control, adhesion to host cells and can
566 reflect a tactical positioning of these important immune evasion proteins. This
567 structure and sequence based analyses suggests that differences between PspC
568 and Hic variants extend beyond cell wall anchors to other domains and to domain
569 pattern usage.
570
571 When comparing prevalence and distribution of PspC vs. Hic variants among 349
572 clinical *S. pneumoniae* isolates derived from adult patients with invasive
573 pneumococcal disease, 298 isolates (85.4%) encoded a PspC-variant, 22 strains
574 (6.3%) a Hic-variant, 19 isolates (5.4%) had a *pspC* and *hic* genes and only 10
575 isolates (2.9%) had neither *pspC* nor *hic* genes (61). In addition, invasive, PspC
576 expressing strains bound more Factor H, and Factor H binding and immune control
577 was more effective in encapsulated as compared to unencapsulated strains. Similarly
578 the PspC variants (i.e. PspC2 and PspC6) were more efficient in Factor H binding
579 and complement inhibition on the bacterial surface as compared to Hic variants
580 (Hic/Pspc9 and Hic/PspC11) (62,63).
581

582 ***Conclusions and Perspectives.*** Evaluating the domain composition of selected
583 PspC and Hic variants and an in-depth characterization of the domain composition
584 resulted in a better understanding of the structure and role of these pneumococcal
585 virulence determinants in immune evasion. Our approach identified further
586 differences between PspC and Hic proteins, which are beyond their distinct
587 membrane anchors. Such knowledge allows a comparison of full-length proteins
588 based on domain patterns, numbers and composition and can result in a better
589 comparison between PspC and Hic proteins. Similarly, individual domains can be
590 compared based on structure, modular composition and sequence.
591
592 Analyzing the additional >60,000 PspC and Hic proteins deposited in the NCBI
593 protein data bank or gene products from new clinical isolates, will likely identify
594 additional variants, new domains, novel domain combinations, and also new
595 subdomains. Understanding the composition of these diverse pneumococcal
596 virulence factors will better explain their role in immune evasion, provide important
597 information for molecular strain typing, and for vaccine design. Last but not least this
598 may also allow a correlation between PspC or Hic type variants with invasive
599 pneumococcal infections and with clinical outcome e.g. of young patients with
600 Pneumococcal Hemolytic Uremic Syndrome.

Materials and method

***Selection of PspC and Hic variant proteins*.** Each of the selected six PspC and five Hic proteins represents one of the two cluster as initially defined by Ianelli *et al.* (38). The sequences were derived from the NCBI protein site (status: **Feb./2018**). The general PspC /Hic designation is based on the definition by Iannelli,et al. (38). The protein names, corresponding bacterial strain, protein size, GenBank Accession number and protein ID are shown in (**Supplementary Table I).**

Secondary structure evaluation*.** The structure (α–helical, coiled-coil and β–sheet) of each selected PspC and Hic protein was evaluated using the program presented via the RaptorX server (http://raptorx.uchicago.edu/). The PspC3.1 shows best matched template is ***2vyuA with p-value:3.39e-10 and secondary structure: 42% (α–helical, 43% coiled-coil and, 14% β–sheets. The other ten PspC Hic variants were evaluated in the same manner, and showed a similar secondary structure composition (Supplementary Figures 1- 10). Each of the six PpsC variants matched best to the same template: 2vyuA, and the five Hic variants (Hic/PspC7.1, Hic/PspC8.1, Hic/PspC9.1, Hic/PspC10.1, Hic/PspC11.1) matched to templates (1w9rA, 4k12B, 2m6uA, 6iaA, 2m6uA, respectively). Three-class secondary structure prediction results are shown in the form of histograms which were constructed using ggplot2 from the R/Bioconductor.

Phylogenetic analysis. The PspC and Hic protein sequences and amino acid composition were evaluated using MEGA7 (www.megasoftware.net). There were a total of 976 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (62. Kumar S., 2016). The CLUSTALW program and the BLOSUM amino acid matrix was used to compare the allelic variants of PspC, following which

627 phylograms were generated using the Neighbor-Joining method (Bootstrap
628 value:100). Each domain's phylogram was generated by the same method described
629 for the full-length protein sequences. Phylogenetic trees are modified in MEGA7.
630
631 ***Domain blast analysis*.** The software BLASTp was used to conduct homology
632 searches of the GenBank database available at the National Center for
633 Biotechnology Information website (http://www.ncbi.nlm.nih.gov/). Furthermore the
634 software BLAST targeting database UnipRotKB reference proteomes plus Swiss-Prot
635 was used to find regions of local similarity between sequences
636 (https://www.uniprot.org/blast/). All the domains in this work have been done a blast.

Acknowledgments

The work of the authors is supported by the Collaborative Research Center, FungiNet (projects C6 (PFZ) and C4 (CS)) Deutsche Forschungsgemeinschaft (DFG). SL acknowledges a fellowship from the German Academic Exchange Service (DAAD) and from the ILRS, International Leibniz Research School for Biomolecular Interaction, Jena, Germany. Alfredo Sahagún-Ruiz was funded by a scholarship from PASPA-DGAPA, National Autonomous University of Mexico (UNAM), and from Mexican National Science and Technology Council (CONACYT) for a sabbatical stay at Department of Infection Biology, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute, Jena, Germany. SH received funding by the Deutsche Forschungsgemeinschaft DFG HA 3125/5-2.

References

1. Weiser JN, Ferreira DM, Paton JC. Streptococcus pneumoniae: transmission, colonization and invasion. Nat Rev Microbiol. 2018;16: 355-367. doi: 10.1038/s41579-018-0001-8.
2. Seth-Smith H. Pneu tricks. Nat Rev Microbiol. 2011;9: 230. doi: 10.1038/nrmicro2547.
3. Henriques-Normark B, Blomberg C, Dagerhamn J, Bättig P, Normark S. Nat Rev Microbiol. 2008;6: 827-837. doi: 10.1038/nrmicro2011.
4. Kadioglu A, Weiser JN, Paton JC, Andrew PW. Nat Rev Microbiol. 2008;6: 288-301. doi: 10.1038/nrmicro1871.
5. UNICEF. *Pneumonia*. (2018). https://www.unicef.org/publications/files/Pneumonia_The_Forgotten_Killer_of_Children.pdf
6. WHO Int; Home - Newsroom - Fact sheets - Detail – Pneumonia https://www.who.int/biologicals/areas/vaccines/pneumo/en/
7. Martinón-Torres F, Salas A, Rivero-Calle I, Cebey-López M, Pardo-Seco J, Herberg JA, et al. EUCLIDS Consortium. Life-threatening infections in children in Europe (the EUCLIDS Project): a prospective cohort study. Lancet Child Adolesc Health. 2018;2(6): 404–414. doi:10.1016/S2352-4642(18)30113-5.
8. Geno KA, Gilbert GL, Song JY, Skovsted IC, Klugman KP, Jones C, et al. Pneumococcal Capsules and Their Types : Past , Present , and Future. Clin Microbiol Rev. 2015;28: 871–899. doi:10.1128/CMR.00024-15.
9. Subramanian K, Henriques-Normark B, Normark S. Emerging concepts in the pathogenesis of the Streptococcus pneumoniae: From nasopharyngeal

colonizer to intracellular pathogen. Cell Microbiol. 2019;21: e13077. doi: 10.1111/cmi.13077.

10. Keller LE, Robinson DA, McDaniel LS. Nonencapsulated streptococcus pneumoniae: Emergence and pathogenesis. MBio. 2016;7: e01792. doi: 10.1128/mBio.01792-15.
11. Weiser JN, Ferreira DM, Paton JC. Streptococcus pneumoniae: transmission, colonization and invasion. Nat Rev Microbiol. 2018;16: 355–367. doi:10.1038/s41579-018-0001-8.
12. Zipfel PF, Hallström T, Hammerschmidt S, Skerka S. The complement fitness factor H: role in human diseases and for immune escape of pathogens, like pneumococci. Vaccine. 2008;26Suppl8: I67-74. doi: 10.1016/j.vaccine.2008.11.015.
13. Fernie-King B, Seilly DJ, Davies A, Lachmann PJ. Subversion of the innate immune response by micro-organisms. Ann Rheum Dis. 2002;61: Suppl 2:ii8-12. doi: 10.1136/ard.61.suppl_2.ii8.
14. Zipfel PF, Hallström T, Riesbeck K. Human complement control and complement evasion by pathogenic microbes-tipping the balance. Mol Immunol. 2013;56: 152-160. doi: 10.1016/j.molimm.2013.05.222.
15. Rooijakkers SH, van Strijp JA. Mol Immunol. 2007;44: 23-32. doi:10.1016/j.molimm.2006.06.011.
16. Lambris JD, Ricklin D, Geisbrecht BV. Complement evasion by human pathogens. Nat Rev Microbiol. 2008;6: 132-142. doi: 10.1038/nrmicro1824.
17. Engholm DH, Kilian M, Goodsell DS, Andersen ES, Kjærgaard RS. A visual review of the human pathogen Streptococcus pneumoniae. FEMS Microbiol Rev. 2017;41: 854-879. doi: 10.1093/femsre/fux037.

699 18. Jedrzejas MJ. Pneumococcal Virulence Factors: Structure and Function.
700 Microbiol Mol Biol Rev. 2001;65: 187–207. doi: 10.1128/MMBR.65.2.187-
701 207.200.
702 19. Pérez-Dorado I, Galan-Bartual S, Hermoso JA. Pneumococcal surface
703 proteins: when the whole is greater than the sum of its parts. Mol Oral
704 Microbiol. 2012;27: 221-245. doi: 10.1111/j.2041-1014.2012.00655.x.
705 20. Chen A, Mann B, Gao G, Heath R, King J, Maissoneuve J, et al. Multivalent
706 Pneumococcal Protein Vaccines Comprising Pneumolysoid with
707 Epitopes/Fragments of CbpA and/or PspA Elicit Strong and Broad
708 Protection. Clin Vaccine Immunol. 2015;22: 1079-1089. doi:
709 10.1128/CVI.00293-15.
710 21. Yuste J, Khandavilli S, Ansari N, Muttardi K, Ismail L, Hyams C, et al. The
711 effects of Pspc on complement-mediated Immunity to Streptococcus
712 pneumoniae Vary with Strain Background. Infec Immun. 2010;78: 283–292.
713 doi: 10.1128/IAI.00541-09.
714 22. Hammerschmidt S, Talay SR, Brandtzaeg P, Chhatwal GS. SpsA, a novel
715 pneumococcal surface protein with specific binding to secretory
716 Immunoglobulin A and secretory component. Mol Microbiol. 1997;25: 1113–
717 1124. doi: 10.1046/j.1365-2958.1997.5391899.x.
718 23. Rosenow C, Ryan P, Weiser JN, Johnson S, Fontan P, Ortqvist A, et al.
719 Contribution of novel choline-binding proteins to adherence, colonization
720 and immunogenicity of Streptococcus pneumoniae. Mol Microbiol. 1997;25:
721 819–829. doi:10.1111/j.1365-2958.1997.mmi494.x.
722 24. Dave S, Brooks-Walter A, Pangburn MK, McDaniel LS. PspC, a
723 Pneumococcal Surface Protein, Binds Human Factor H. J Infect Immu.
724 2001;69: 3435–3437. doi: 10.1128/IAI.69.5.3435-3437.2001.

725 25. Jarva H, Janulczyk R, Hellwage J, Zipfel PF, Björck L, Meri S.
726 Streptococcus pneumoniae evades complement attack and
727 opsonophagocytosis by expressing the pspC locus-encoded Hic protein that
728 binds to short consensus repeats 8-11 of factor H. J Immunol. 2002;168:
729 1886-1894. doi: 10.4049/jimmunol.168.4.1886.
730 26. Zhang JR, Mostov KE, Lamm ME, Nanno M, Shimida S, Ohwaki M, et al.
731 The polymeric immunoglobulin receptor translocates pneumococci across
732 human nasopharyngeal epithelial cells. Cell. 2000;102: 827–837.
733 doi:10.1016/s0092-8674(00)00071-4.
734 27. Cheng Q, Finkel D, Hostetter MK. Novel Purification Scheme and Functions
735 for a C3-Binding Protein from Streptococcus pneumoniae. Biochem.
736 2000;39: 5450–5457. doi:10.1021/bi992157d.
737 28. Janulczyk R, Iannelli F, Sjöholm AG, Pozzi G, Björck L. Hic, a novel surface
738 protein of Streptococcus pneumoniae that interferes with complement
739 function. J Biol Chem. 2000;275: 37257–37263. doi:
740 10.1074/jbc.M004572200.
741 29. Pangburn MK, Dave S, McDaniel LS, Carmicle S, Hammerschmidt S. Dual
742 Roles of PspC, a Surface Protein of Streptococcus pneumoniae, in Binding
743 Human Secretory IgA and Factor H. J Immunol. 2014;173: 471–477. doi:
744 10.4049/jimmunol.173.1.471.
745 30. Binsker U, Kohler TP, Krauel K, Kohler S, Habermeyer J, Schwertz H, et al.
746 Serotype 3 pneumococci sequester platelet-derived human
747 thrombospondin-1 via the adhesin and immune evasion protein Hic. J Biol
748 Chem. 2017;292: 5770–5783. doi: 10.1074/jbc.M116.760504.
749 31. Lu L, Ma Z, Jokiranta TS, Whitney AR, DeLeo FR, Zhang JR. Species-
750 Specific Interaction of Streptococcus pneumoniae with Human Complement

751 Factor H. J Immunol. 2008;181: 7138–7146. doi:
752 10.4049/jimmunol.181.10.7138.
753 32. Lu L, Ma Y, Zhang JR. Streptococcus pneumoniae recruits complement
754 factor H through the amino terminus of CbpA. J Biol Chem. 2006;281:
755 15464–15474. doi: 10.1074/jbc.M602404200.
756 33. Hyams C, Trzcinski K, Camberlein E, Weinberger. DM, Chimalapati S,
757 Noursadeghi M, et al. Streptococcus pneumoniae capsular serotype
758 invasiveness correlates with the degree of factor H binding and
759 opsonization with C3b/iC3b. Infect Immun. 2013;81: 354–363. doi:
760 10.1128/IAI.00862-12.
761 34. Kohler S, Hallström T, Singh B, Riesbeck K, Spartà G, Zipfel PF, et al.
762 Binding of vitronectin and factor H to hic contributes to immune evasion of
763 Streptococcus pneumoniae serotype 3. Thromb Haemost. 2015;113: 125–
764 142. doi: 10.1160/TH14-06-0561.
765 35. Haleem KS, Ali YM, Yesilkaya H, Kohler T, Hammerschmidt S, Andrew PW,
766 et al. The Pneumococcal Surface Proteins **PspA** and PspC Sequester Host
767 C4-Binding Protein To Inactivate Complement C4b on the Bacterial
768 Surface. Infect Immun. 2018;87: pii: e00742-18. doi: 10.1128/IAI.00742-18.
769 36. Dieudonné-Vatran A, Krentz S, Blom AM, Meri S, Henriques-Normark B,
770 Riesbeck K, et al. Clinical Isolates of Streptococcus pneumoniae Bind the
771 Complement Inhibitor C4b-Binding Protein in a PspC Allele-Dependent
772 Fashion. J Immunol. 2009;182: 7865–7877. doi:
773 10.4049/jimmunol.0802376.
774 37. Brooks-Walter A, Briles DE, Hollingshead SK. The pspC gene of
775 Streptococcus pneumoniae encodes a polymorphic protein, PspC, which

776 elicits cross-reactive antibodies to PspA and provides immunity to
777 pneumococcal bacteremia. Infect Immun. 1999;67: 6533–6542.
778 38. Iannelli F, Oggioni MR, Pozzi G. Allelic variation in the highly polymorphic
779 locus pspC of Streptococcus pneumoniae. Gene. 2002;284: 63-71.
780 doi:10.1016/S0378-1119(01)00896-4.
781 39. Meinel C, Spartà G, Dahse HM, Hörhold F, König R, Westermann M, et al.
782 Streptococcus pneumoniae from Patients with Hemolytic Uremic Syndrome
783 Binds Human Plasminogen via the Surface Protein PspC and Uses Plasmin
784 to Damage Human Endothelial Cells. J Infect Dis. 2018;217: 358–370. doi:
785 10.1093/infdis/jix305.
786 40. Luo R, Mann B, Lewis WS, Rowe A, Heath R, Stewart ML, et al. Solution
787 structure of choline-binding protein A, the major adhesin of Streptococcus
788 pneumoniae. EMBO J. 2005;24: 34–43. doi: 10.1038/sj.emboj.7600490.
789 41. Van Rosmalen M, Krom M, Merkx M. Tuning the Flexibility of Glycine-
790 Serine Linkers to Allow Rational Design of Multidomain Proteins. Biochem.
791 2017;56: 6565–6574. doi: 10.1021/acs.biochem.7b00902.
792 42. Keun Kim H, Thammavongsa V, Schneewind O, Missiakas D. Recurrent
793 infections and immune evasion strategies of Staphylococcus aureus. Curr
794 Opin Microbiol. 2012;15: 92-99. doi: 10.1016/j.mib.2011.10.012.
795 43. Håkansson A, Roche H, Mirza S, McDaniel LS, Brooks-Walter A, Briles DE.
796 Characterization of Binding of Human Lactoferrin to Pneumococcal Surface
797 Protein A. Infect Immun. 2001;69: 3372–3381. doi: 10.1128/IAI.69.5.3372-
798 3381.2001.
799 44. Hammerschmidt S, Bethe G, Remane PH, Chhatwal GS. Identification of
800 pneumococcal surface protein A as a lactoferrin- binding protein of
801 Streptococcus pneumoniae. Infect Immun. 1999;67: 1683–1687.

802 45. Senkovich O, Cook WJ, Mirza S, Hollingshead SK, Protasevich II, Briles
803 DE, et al. Structure of a Complex of Human Lactoferrin N-lobe with
804 Pneumococcal Surface Protein A Provides Insight into Microbial Defense
805 Mechanism. J Mol Biol. 2007;370: 701–713. doi:
806 10.1016/j.jmb.2007.04.075.
807 46. Xu Q, Zhang JW, Chen Y, Li Q, Jiang YL. Crystal structure of the choline-
808 binding protein CbpJ from Streptococcus pneumoniae. Biochem Biophys
809 Res Commun. 2019;514: 1192–1197. doi:10.1016/j.bbrc.2019.05.053.
810 47. Kanchi PK, Dasmahapatra AK. Polyproline chains destabilize the
811 Alzheimer's amyloid-β protofibrils: A molecular dynamics simulation study. J
812 Mol Graph Model. 2019;93: 107456. doi:10.1016/j.jmgm.2019.107456.
813 48. Jerlström PG, Chhatwal GS, Timrnis KN. The IgA-binding β antigen of the c
814 protein complex of Group B streptococci: sequence determination of its
815 gene and detection of two binding regions. Mol Microbiol. 1991;5: 843-849.
816 doi: 10.1111/j.1365-2958.1991.tb00757.x.

817 49. Girgis MM, Abd El-Aziz AM, Hassan R, Ali YM. Immunization With Proline
818 Rich Region of Pneumococcal Surface Protein A Has No Role in Protection
819 Against Streptococcus Pneumoniae Serotype 19F. Microb Pathog.
820 2020;138: 103761. doi: 10.1016/j.micpath.2019.103761.

821 50. Mukerji R, Hendrickson C, Genschmer KR, Park SS, Bouchet V, Goldstein
822 R, et al. The diversity of the proline-rich domain of pneumococcal surface
823 protein A (PspA): Potential relevance to a broad-spectrum vaccine.
824 Vaccine. 2018;36: 6834-6843. doi: 10.1016/j.vaccine.2018.08.045.
825 51. McDaniel LS, Ralph BA, McDaniel DO, Briles DE. Localization of protection-
826 eliciting epitopes on PspA of Streptococcus pneumoniae between amino

827 acid residues 192 and 260. Microb Pathog. 1994;17(5): 323-337.
828 doi:10.1006/mpat.1994.1078.
829 52. Maestro B, Sanz JM. Choline Binding Proteins from Streptococcus
830 pneumoniae : A Dual Role as Enzybiotics and Targets for the Design of
831 New Antibiotics. 2016;5: pii:E21. doi:10.3390/antibiotics5020021.
832 53. Hakenbeck R, Madhour A, Denapaite D, Brückner R. Versatility of choline
833 metabolism and choline-binding proteins in Streptococcus pneumoniae and
834 commensal streptococci. FEMS Microbiol. 2009;3(3): 572-586. doi:
835 10.1111/j.1574-6976.2009.00172.x.
836 54. Marraffini LA, Dedent AC, Schneewind O. Sortases and the art of anchoring
837 proteins to the envelopes of gram-positive bacteria. Microbiol Mol Biol Rev.
838 2006;70: 192-221. doi: 10.1128/MMBR.70.1.192-221.2006.
839 55. Pallen MJ, Lam AC, Antonio M, Dunbar K. An embarrassment of sortases-
840 A richness of substrates? Trends Microbiol. 2001;9: 97–101. doi:
841 10.1016/s0966-842x(01)01956-4.
842 56. Daniels CC, Coan P, King J, Hale J, Benton KA, Briles DE, et al. The
843 Proline-Rich Region of Pneumococcal Surface Proteins A and C Contains
844 Surface-Accessible Epitopes Common to All Pneumococci and Elicits
845 Antibody-Mediated Protection against Sepsis. Infect Immun. 2010;78:
846 2163–2172. doi: 10.1128/IAI.01199-09.
847 57. McDaniel LS, McDaniel DO, Hollingshead SK, Briles DE. Comparison of the
848 PspA sequence from Streptococcus pneumoniae EF5668 to the previously
849 identified PspA sequence from strain Rx1 and ability of PspA from EF5668
850 to elicit protection against pneumococci of different capsular types. Infect
851 Immun. 1998;66(10): 4748–4754. doi: 10.1128/IAI.66.10.4748-4754.1998.

852 58. Georgieva M, Kagedan L, Lu YJ, Thompson CM, Lipsitch M. Antigenic
853 Variation in Streptococcus pneumoniae PspC Promotes Immune Escape in
854 the Presence of Variant-Specific Immunity. mBio. 2018;9: 2.pii: e00264-18.
855 doi: 10.1128/mBio.00264-18.
856 59. Desvaux M, Dumas E, Chafsey I, Hébraud M. Protein cell surface display in
857 Gram-positive bacteria: From single protein to macromolecular protein
858 structure. FEMS Microbiol Lett. 2006;256: 1-15. doi: 10.1111/j.1574-
859 6968.2006.00122.x.
860 60. Pathak A, Bergstrand J, Sender V, Spelmink L, Aschtgen MS, Muschiol S,
861 et al. Factor H binding proteins protect division septa on encapsulated
862 Streptococcus pneumoniae against complement C3b deposition and
863 amplification. Nat Commun. 2018;9: 3398. doi: 10.1038/s41467-018-05494-
864 w.
865 61. van der Maten E, van den Broek B, de Jonge MI, Rensen KJW, Eleveld MJ,
866 Zomer AL, et al. Streptococcus pneumoniae PspC Subgroup Prevalence in
867 Invasive Disease and Differences in Contribution to complement Evasion.
868 Infect Immun. 2018;86: pii:e00010-18 doi: 10.1128/IAI.00010-18.
869 62. Chang B, Nariai A, Sekizuka T, Akeda Y, Kuroda M, Oishi K, et al. Capsule
870 switching and antimicrobial resistance acquired during repeated
871 Streptococcus pneumoniae pneumonia episodes. J Clin Microbiol. 2015;53:
872 3318–3324. doi:10.1128/JCM.01222-15.
873 63. Anderson D, Fakiola M, Hales BJ, Pennell CE, Thomas WR, Blackwell JM.
874 Genome-wide association study of IgG1 responses to the choline-binding
875 protein PspC of Streptococcus pneumoniae. Genes Immun. 2015;16: 289–
876 296. doi:10.1038/gene.2015.12.

877 64. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetic
878 Analyses version 7. Mol Biochem Evol. 2016;33(7): 1870-1874. doi:
879 10.1093/molbev/msw054

Supporting Information Captions

Figure Legends

Figure 1: Diversity of PspC and Hic cluster variants.

PspC and Hic proteins were initially considered to represent one protein class that based on the different surface anchors can be divided into two major clusters. **A:** PspC variants with choline-binding domains representing the PspC group, and Hic variants having sortase LPsTG motifs for cell wall anchoring the second namely Hic group. For each group additional cluster or subgroups were identified. For the analysis one variant from each cluster was selected, i.e. for PspC group: PspC1.1, PspC2.2, PspC3.1, PspC4.2, PspC5.1, PspC6.1; and for the Hic group: Hic/PspC7.1, Hic/PspC8.1, Hic/PspC9.1, Hic/PspC10.1, Hic/PspC11.1. **B: Overall sequence homology among the selected cluster variants.** Amino acid homology of the indicated full-length protein variants was compared to PspC3.1, which was used as reference. The sequence variation shows differences for the six selected PspC and the five Hic variants. This difference is indicative for compositional variation among the two major protein groups.

Figure 2: Structural regions and domain position of PspC3.1.

***In silico* structure analyses of PspC3.1 dissects distinct structural region. A: The structure of the well-characterized PspC3.1 (strain D39) was evaluated *in silico*.** The N-terminal part of the molecule shows a long stretch composed mainly of α-helices (red columns) (aa 1-410), being followed by a 72 aa long coiled-coil structured segment (grey area) and by a 219 aa long segment with β–sheet folds (blue columns). The numbers on top represent the amino acid position within the protein. The signal peptide (positions 1-37) which is cleaved upon processing is shown by the box with grey background and blue lines. The vertical grey bar

separating the N-terminal α helical from the coiled coil structured region may represent the position of the bacterial cell wall. **B: Structural regions and domain composition of PspC3.1**. The mainly α–helical structured region (position 38 to 410) is termed the N-terminal region. The remained of the protein that includes the 72 aa coiled-coil structured and the 219 aa mainly β-sheet segment is termed the C-terminal region (left panel). To correlate structural regions with the domain composition, the know domains of PspC3.1 were aligned (right panel). The hypervariable domain, repeat domain I, random coil domain, repeat domain II aligned with the N-terminal, mainly α-helical region. In the C-terminal part the Proline-Rich Domain lined up with the coiled-coil structured region and the Choline-Binding Domain with the β-sheet region. The grey horizontal line separates the N and C-terminal regions and likely marks the border of the cell wall and capsule facing the outside environment. **C: Amino acid composition of N and C-terminal regions.** The amino acid composition was evaluated for each region separately. The N-terminal region is rich in charged residues (48%), has low degree of polar and amphipathic residues (24%), and contains a low fraction of Tyr residues (left panel). The C-terminal region contained a lower fraction of charged residues (22%), had more polar amphipathic amino acids (38%) and more Tyr residues (8%).

Figure 3: Differences in the N and C-terminal regions of the PspC and Hic variants.

A: The N and C-terminal regions of PspC and Hic type proteins differ in amino acid composition. The amino acid composition of the N and C-terminal regions was evaluated for each of the six PspC and the five Hic variants. The N-terminal regions of the PspC and Hic variants are rich in charged residues (35-45%), have a low degree of polar and amphipathic residues, and contain a low percentage of Tyr

932 residues. The PspC variants had also a high portion of charged residues (28-
933 27%)(upper panel). The C-terminal regions of the PspC variants had a lower fraction
934 of charged residues (16 % or less), more polar and hydrophilic residues (32-36% and
935 more Tyr residues (8.3-9.1%). The composition of the C-terminal region of Hic
936 variants differed from that of PspC variants. The C-terminal regions of Hic variants
937 showed more charged residues, a lack of Tyr residues and less polar and
938 amphipathic residues (lower panel). **B: Homology alignment of the N and C-**
939 **terminal regions of PspC and Hic type proteins.** The homology alignment of the N
940 and C-terminal regions identifies two groups. For the N-terminal regions the first,
941 group A is dominated by PspC type proteins but also includes Hic/PspC11.1. The
942 second N terminal group B is dominated by Hic type proteins but also includes the
943 PspC4.2 variant. The C-terminal regions show a clear separation among the PspC
944 and Hic variants.
945
946 **Figure 4: Domain structure of the six PspC and th efive Hic variants**.
947 **A: PspC3.1 with the domain architecture** is shown on the left side. The PspC and
948 Hic variants differ in length and in domain number. The proteins each representing
949 one member of the previously identified clusters are arranged based on their overall
950 homology. To reflect the different lengths of the proposed outside and interior regions
951 the proteins are centered along the axis, which separates the N-terminal α-helical
952 region from the C-terminal region. N-terminal regions are shown on yellow and C-
953 terminal regions on a grey background. Proteins are drawn to scale. The signal
954 peptides and the most C-terminal segments of class II proteins, which are cleaved by
955 the transpeptidase sortase are not presented. Know domains as identified for
956 PspC3.1 are shown in filled color. New domains are patterned, and the names are
957 represented on grey background. The mapped binding sites for the human plasma

proteins Factor H in the Hypervariable Domain are shown by the purple bar and that of sIgA/pIgR in the Repeat Domains by green bars. Lactoferrin and IgA binding domains are proposed by homology with binding domains of *S. pneumoniae* protein PspA and by the IBC protein from *S. agalactiae*.

Figure 5: Sequence Variation and Conservation of Binding Domains and Surface Orientation of PspC1.1 and Hic/PspC8.1.

A: Sequence variation of the Factor H binding motif in the Hypervariable Domains of the six PspC and the five Hic variants. WebLogo was used to evaluate amino acid variation. **B: Sequence conservation in the binding sites for human sIgA/pIgR** in the Repeat Domains I and II. **C: Sequence variation among the Choline-Binding Modules 2 and 3 the PspC variants**. Residues relevant for the contact with choline are indicated by the arrows and include W_3 and W_{10} of one module, as well as Y_{11} of the following module. WebLogo was used to evaluate sequence variation in the second and third choline-binding modules of the PspC variants. **D: Sequence conservation in C-termini of Hic-type proteins** of the sortase recognition motif LPsTG of covalently anchored proteins. Following sortase cleavage after the T residue and attachment to Penicillin Binding Protein. **E: Structure and proposed orientation of PspC1.1 associated with phosphorylcholine (**PCho**), and sortase linked Hic/PspC8.1 variant**. The arrangement is based on the concept that PspC1.1 is non-covalently associated to the teichoic acids via its interaction with PCho. In contrast the Hic/PspC8.1 variant is which is covalently linked via the sortase anchor to peptidoglycan Penicillin binding protein (PBP). This attachment and orientation suggests that the Proline-Rich Domains may represent flexible cell wall and capsule spanning segment. Furthermore, the variable length of the C-terminal regions can indicate different types

of cell wall attachment, as well as discrete sizes and thickness of the cell wall and the capsule. The grey line represents the bacterial membrane, cell wall and the shaded grey region indicates the position of the capsule. The proposed exterior domains of the PspC and the Hic variant are shown in yellow or red color. The known, mapped binding domains for human plasma regulator Factor H in the Hypervariable Domains (PspC1.1 and Hic/PspC81) and the sIgA or cell surface receptor pIgR in the Repeat Domains I and II (PspC1.1) are indicated by purple and green bars. Attached Factor H mediates complement evasion and blocks complement mediated opsonophagocytosis and release of anaphylatoxins C3a and C5a. SIgA or pIgR bind to two sites in PspC1.1 and avoid opsonization by sIgA or mediate adhesion to human epithelial cells. The binding sites for additional human plasma protein like vitronectin are not mapped so far. The C-terminal region, with a proposed interior location are shown in green, blue or purple color and include the Proline Rich Domains followed by Choline-Binding Module (PspC1.1) or LPsTG mediated anchor (Hic/PspC8.1).

Table I: Human Regulators binding to PspC and Hic variants.

Binding of human plasma regulators to PspC and Hic proteins. The binding sites for Factor H has been mapped within the Hypervariable Domain of PspC3.1 and that of sIgA and the extracellular domain of pIgR to the RNYPT motif of Repeat Domains I and II. Binding of C3, C4BP, Plasminogen, Thrombospondin 1, vitronectin have been show to intact *S. pneumoniae* and to full length PspC and Hic proteins, but their binding sites have not been mapped to single domains so far. Interaction of Lactoferrin and IgA is proposed based on homology between PspC and Hic variants with the *S. pneumoniae* immune escape protein PspA, and the homology to the sIgA binding protein of *S. agalactiae*.

Table II: Domains Identified in the evaluated PspC and Hic variants. The domains are listed from N-terminal to the C-terminal region, and known as well as new domain and domain alternates are presented. Also domains which are specific for PspC/Hic variants are shown, as well as domains which are shared and found in other *S. pneumoniae* proteins and in other bacterial proteins. RD ?

SP signal peptide; HVD hypervariable Domain; RD Repeat Domain; RCD Random Coil Domain; SnD/GS2 Serine Rich segment; RCE Random Coil Extension; R-type repeat related Domain; EPRD Extracellular Proline Rich Domain; VS Variant specific; IgA IgA Binding Domain, PRD Proline-Rich Domain, CBP Choline-Binding Domain.

Supplementary Figure Legends

Supplementary Figure 1: Structural composition of PspC variants with choline binding anchors. (A) The structure of Psp2.2 was evaluated *in silico*. The N-terminus shows a long stretch composed mainly of α-helices (red columns) (aa 1-440), being followed by a 74 aa long coiled-coil structured segment (grey area) and by a 179 aa long segment with β–sheet folds (blue columns). The signal peptide (aa 1-37) which is cleaved upon processing is shown by the grey background and blue lines. The vertical grey bar separating the N-terminal α helical region and the C-terminal coiled coil structured region may represent the position of the bacterial cell wall. **(B) Structural composition of PspC2.2.** The structure of Psp2.2 was evaluated *in silico*. The N-terminus shows a long stretch composed mainly of α-helices (red columns) (aa 1-405), being followed by a 77 aa long coiled-coil structured segment (grey area) and by a 199 aa long segment with β–sheet folds (blue columns). The signal peptide (aa 1-37) which is cleaved upon processing is shown by the grey background and blue lines. The vertical grey bar separating the N-terminal α helical region and the C-terminal coiled coil structured region may represent the position of the bacterial cell wall. **(C) Structural composition of PspC1.1.** The structure of PspC1.1 was evaluated *in silico*. The N-terminus shows a long stretch composed mainly of α-helices (red columns) (aa 1-626), being followed by a 64 aa long coiled-coil structured segment (grey area) and by a 248 aa long segment with β–sheet folds (blue columns). The signal peptide (aa 1-37) which is cleaved upon processing is shown by the grey background and blue lines. The vertical grey bar separating the N-terminal α helical region and the C-terminal coiled coil structured region may represent the position of the bacterial cell wall. **(D) Structural composition of PspC5.1.** The structure of Psp5.1 was evaluated *in*

silico. The N-terminus shows a long stretch composed mainly of α-helices (red columns) (aa 1-632), being followed by a 59 aa long coiled-coil structured segment (grey area) and by a 178 aa long segment with β–sheet folds (blue columns). The signal peptide (aa 1-37), which is cleaved upon processing is shown by the grey background and blue lines. The vertical grey bar separating the N-terminal α–helical region and the C-terminal coiled-coil structured region may represent the position of the bacterial cell wall. **(E) Structural composition of PspC4.2.** The structure of Psp4.2 was evaluated *in silico*. The N-terminus shows a long stretch composed mainly of α-helices (red columns) (aa 1-610), being followed by a 57 aa long coiled-coil structured segment (grey area) and by a 199 aa long segment with β–sheet folds (blue columns). The signal peptide (aa 1-37) which is cleaved upon processing is shown by the grey background and blue lines. The vertical grey bar separating the N-terminal α-helical region and the C-terminal coiled-coil structured region may represent the position of the bacterial cell wall.

Supplementary Figure 2: Structural composition of Hic type variants with LPsTG anchors. (A) Structural composition of Hic/PspC7.1. The structure of HIC/Psp7.1 was evaluated *in silico*. The N-terminus shows a long stretch composed mainly of α-helices (red columns) (aa 1-533), being followed by a 186 aa long coiled-coil structured segment (grey area) and by a 50 aa long segment with an LPsTG motif. This segment has mostly α–helical structure. The signal peptide (aa 1-37) which is cleaved upon processing, is shown by the grey background and blue lines. The vertical grey bar separating the N-terminal α-helical region and the C-terminal mostly coiled-coil structured region may represent the position of the bacterial cell wall. **(B) Structural composition of Hic/PspC10.1.**The structure of HIC/Psp10.1

was evaluated *in silico*. The N-terminus shows a long stretch composed mainly of α-helices (red columns) (aa 1-502), being followed by a 204 aa long coiled-coil structured segment (grey area) and by a 57 aa long segment with an LPsTG motif. This segment has preceding the motif a coiled coil and α–helical structure following the motif. The signal peptide (aa 1-37) which is cleaved upon processing, is shown by the grey background and blue lines. The vertical grey bar separating the N-terminal α-helical region and the C-terminal mostly coiled-coil structured region may represent the position of the bacterial cell wall. **(C) Structural composition of Hic/PspC9.1.** The structure of HIC/Psp9.1 was evaluated *in silico*. The N-terminus shows a long stretch composed mainly of α-helices (red columns) (aa 1-279), being followed by a 247 aa long coiled-coil structured segment (grey area) and by a 57 aa long segment with an LPsTG motif. This segment has preceding the motif a coiled coil and α–helical structure following the motif. The signal peptide (aa 1-37) which is cleaved upon processing is shown by the grey background and blue lines. The vertical grey bar separating the N-terminal α–helical region and the C-terminal mostly coiled-coil structured region may represent the position of the bacterial cell wall. **(D) Structural composition of Hic/PspC8.1.** The structure of HIC/Psp8.1 was evaluated *in silico*. The N-terminus shows a long stretch composed mainly of α-helices (red columns) (aa 1-155), being followed by a 286 aa long coiled-coil structured segment (grey area) and by a 62 aa long segment with an LPsTG motif. This segment has preceding the motif a coiled coil and α–helical structure following the motif. The signal peptide (aa 1-37) which is cleaved upon processing is shown by the grey background and blue lines. The vertical grey bar separating the N-terminal α-helical region and the C-terminal mostly coiled-coil structured region may represent the position of the bacterial cell wall. **(E) Structural composition of Hic/PspC11.1.**

The structure of HIC/Psp11.1 was evaluated *in silico*. The N-terminus shows a long stretch composed mainly of α-helices (red columns) (aa 1-264), being followed by a 286 aa long coiled-coil structured segment (grey area) and by a 62 aa long segment with an LPsTG motif. This segment has preceding the motif a coiled coil and α–helical structure following the motif. The signal peptide (aa 1-37) which is cleaved upon processing, is shown by the grey background and blue lines. The vertical grey bar separating the N-terminal α-helical region and the C-terminal mostly coiled coil structured region may represent the position of the bacterial cell wall.

Supplementary Figure 3: Amino acid sequences of Signal Peptides and of the Hypervariable Domains.

A: Sequence Conversation of the Signal Peptide. Sequences of the N-terminal region of the six PspC and the five Hic/PspC variants are shown. Conserved residues are shown with a black background; residues which are present in most proteins are shown on grey background. Positively charged residues are shown in blue, and negatively charged residues in red characters. **B: Sequences of the Hypervariable Domains** of the six PspC and the five Hic variants. The Factor H binding sites which has been mapped for PspC3.1 is shown with green background. The hypervariable domains can be separated into three major groups, termed HVD-A, HVD-B and HVD-C.

Supplementary Figure 4: Sequence Conservation in Repeat Domains I and II

A: Sequences of the N-terminal region of the Repeat Domains of six PspC- and the HIC/PspC variant. See legend to Supplementary Figure 11 for explanation. The conserved binding domain for sIgA is shown with yellow background. **B:** Alignment of Repeat Domain II.

Supplementary Figure 5: Conserved Residues of the Random Coil Domains and S_nD/GS_2 Domains.

A: Sequences of the Random Coil Domain following the first Repeat Domain are shown. See legend to Supplementary Figure 3 for explanation. **B**: **Conserved Residues in S_nD/GS_2 Domains.** The upper panels show the segments that follow the Hypervariable regions and the lower panel the Proteins and segments that follow the random Coil domain.

Supplementary Figure 6: Conserved Amino Acid Residues of the New N-terminal Domains.

A: Sequence Homology of the New Random Coil Extension region I**; B:** Sequence Homology of the New Random Coil Extension region II **C:** Sequence Homology of the **Random Coil Domain; D: Homology of PspA like Domains.** The bottom row shows the sequence of PspA from strain EF6769 with includes a lactoferrin binding domain. E: **Homology of Repeat Relate** The bottom row shows the sequence of PspA from *S. agalactiae.* **F: Sequence of the PspC4.2 specific Element**
G: Sequence of the Hic/PspC11.1 Specific Domain. General: Conserved residues are shown with a black background; residues which are present in most proteins are shown on grey background. Positively charged residues are shown in blue, and negatively charged residues in red characters.

Supplementary Figure 7: Conserved Residues of the C-terminal Proline-Rich Domain I to Proline-Rich Domain III.

A: Proline-Rich Domain I. Proline Rich Doman I is used by five PspC proteins, PspC3.1, PspC2.2, PspC6.1 PspC1.1, PspC5.1. This domain has three major regions. The first and third domains have conserved Pro residues and the motifs

PAPAP and PAPAT are found in most domains. The C-terminal part is also rich in Pro residues. The middle segment, element II which is represented by flanking Q-residues, is found in PspC1.1, PspC2.2, and PspC6.1 but is absent in PspC1.1 and PspC5.1. **B: Proline-Rich Domain II.** PspC4.2 uses a separate repeat domain which has an 18-residue element duplicated. **C: Proline-Rich Domain III.** Hic/PspC7.2 has a Proline Rich Domain composed of six segments. The first segment is seven residues long, segments 2-5 include duplicated 31 residue long regions and the most C-terminal segment represents is a truncated 24 aa long version of these repeat units. **D:** The conserved and variant residues were identified by WEBLOGO.

Supplementary Figure 8: Conserved Amino Acid Residues of the C-terminal Proline-Rich Domains IV.

Proline-Rich Domain IV of Hic/PspC9.1 and Hic/pspC10.1. Variants of the Proline-Rich Doman IV is used by Hic/PspC9.1(A) and Hic/PspC10.1 (B). Both domains use a six residue long segment as first unit, which are followed, by four (Hic/PspC9.1) or two (Hic/PspC10.1) 11 amino acid long segments. Next segments 6-21 (Hic/PspC9.1) or segments 4-17 (HIC/PspC10.1) are 11 residues long and highly related to each other. The most C-terminal segment is 22 residues long unit. **C:** Sequence homology alignment of the conserved 11 residues long elements of PspC9.1 and Hic/PspC10.1. **C**: conserved and variant residues in Hic/PspC9.1 and Hic/PspC10.1 The Variation occurs at position 1 of the repeat units. WEBLOGO was used for alignment. **Proline-Rich Domain VI. Highly related variant of the** Proline Rich Doman VI are used by Hic/PspC8.1 **(D**) and Hic/PspC11.1 (**E**). Both domains use a seven residue long segment as first unit. Units 1-24 are highly related almost conserved 11 residue long repeated units. A variation occurs at position 1 of the repeat units. **F:** Homology alignment of the conserved 11 residues long elements of

Hic/PspC8.1 and Hic/PspC11.1 using WEBLOGO. A variation occurs at position 1 of the repeat units.

Supplementary Figure 9: Conserved amino acid residues of the C-terminal anchor sequences of the PspC and Hic variants.

A: Sequence Homology among the modules of the Choline Binding Domains in the six PspC variants, i.e. PspC3.1, PspC2.2, PspC6.1, PspC1.1, PspC5.1, PspC4.2 are shown. Based on sequence comparison the first modules, have a more different amino acid structure and show conserved sequence pattern which identified these domains as the first positon in a choline binding domain.. The following modules form a middle segment, where the modules are also relatively conserved to each other in sequence. This middle segment shows variation in number of modules, ranging from five to eight. Similarly the three most C-terminal modules of each domain show position specific features and they are conserved among the six PspC variants. These modules are termed third to last, second to last and last most C terminal module.**B:** Sequence alignment of the LPsTG anchor of the five HIC variants. The **LPsTG** motif which is relevant for sortase anchoring is shown in white letters on red background.

Supplementary Table I: Proteins representing the specific clusters which were selected for the detailed Sequence and structural Analyses.

Protein names, strain origin, length in aa, as well as GenBank and Protein Accession numbers are presented.

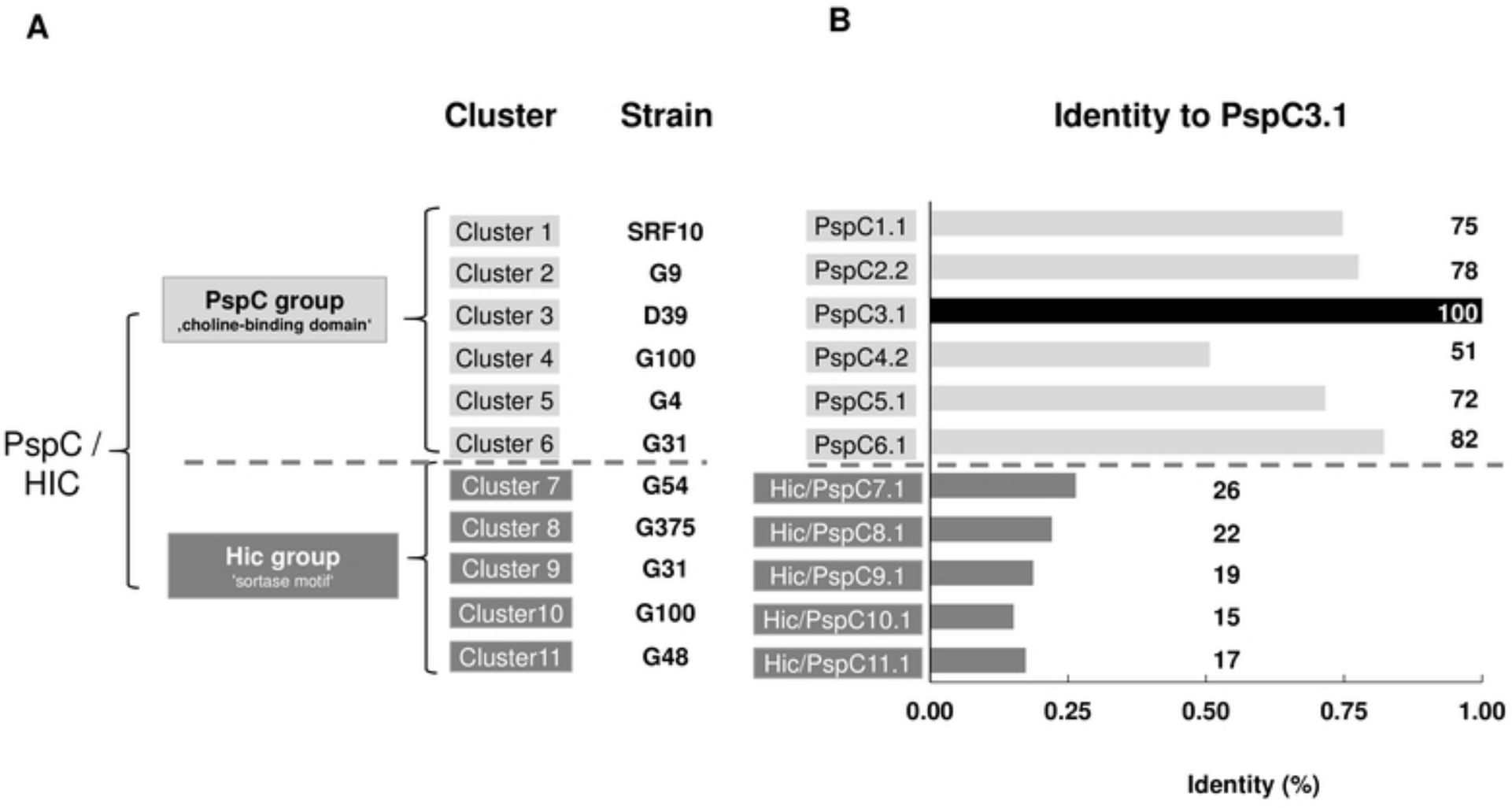
A
Cluster
Strain
PspC group
‚choline-binding domain'
Hic group
'sortase motif'
PspC / HIC
Cluster 1
SRF10
Cluster 2
G9
Cluster 3
D39
Cluster 4
G100
Cluster 5
G4
Cluster 6
G31
Cluster 7
G54
Cluster 8
G375
Cluster 9
G31
Cluster10
G100
Cluster11
G48
B
Identity to PspC3.1
PspC1.1
75
PspC2.2
78
PspC3.1
100
PspC4.2
51
PspC5.1
72
PspC6.1
82
Hic/PspC7.1
26
Hic/PspC8.1
22
Hic/PspC9.1
19
Hic/PspC10.1
15
Hic/PspC11.1
17
0.00
0.25
0.50
0.75
1.00
Identity (%)

Figure 1

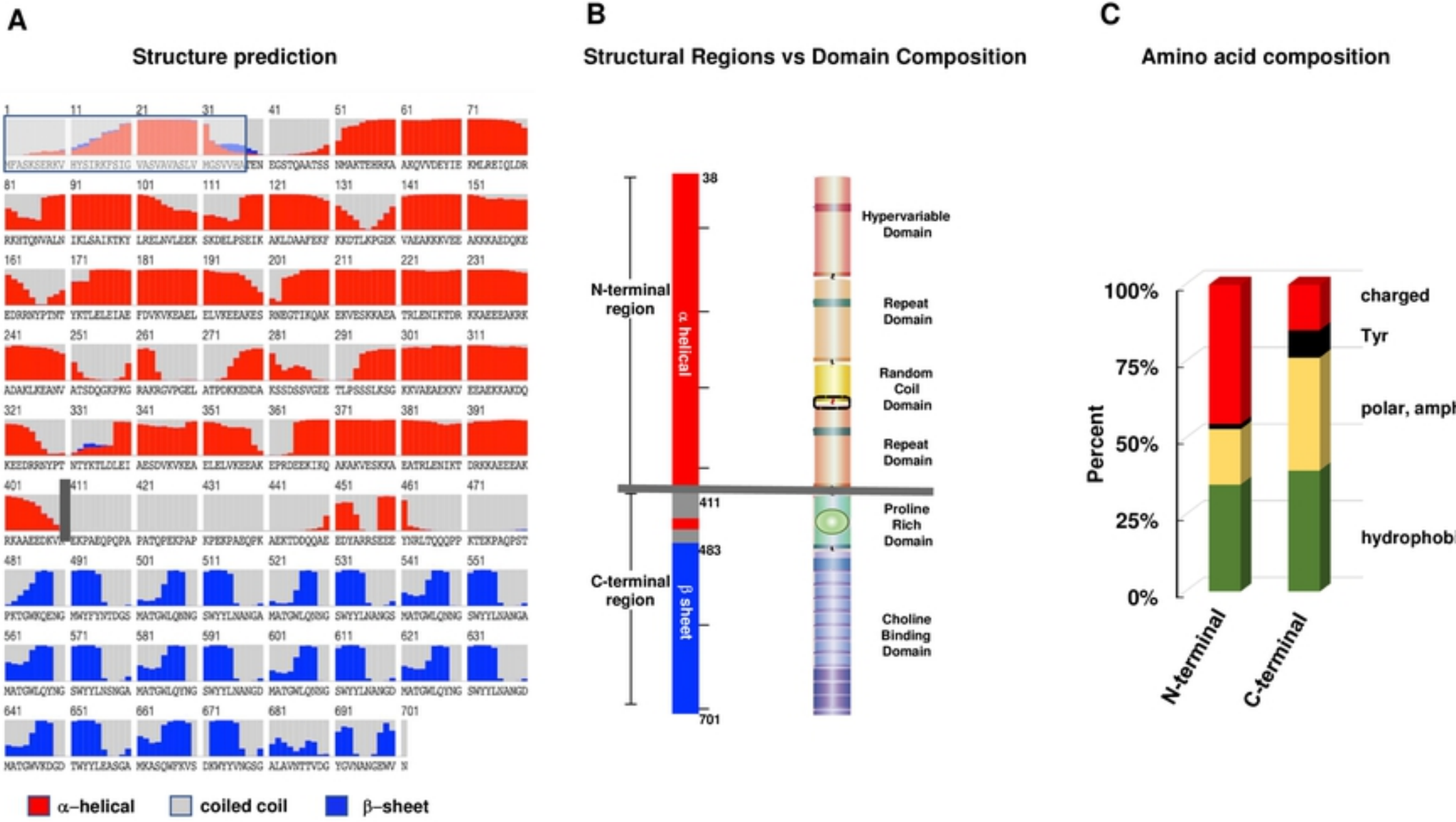
A
Structure prediction
α–helical
coiled coil
β–sheet
B
Structural Regions vs Domain Composition
N-terminal region
C-terminal region
α helical
β sheet
38
411
483
701
Hypervariable Domain
Repeat Domain
Random Coil Domain
Repeat Domain
Proline Rich Domain
Choline Binding Domain
C
Amino acid composition
Percent
100%
75%
50%
25%
0%
charged
Tyr
polar, ampl
hydrophob
N-terminal
C-terminal

Figure 2

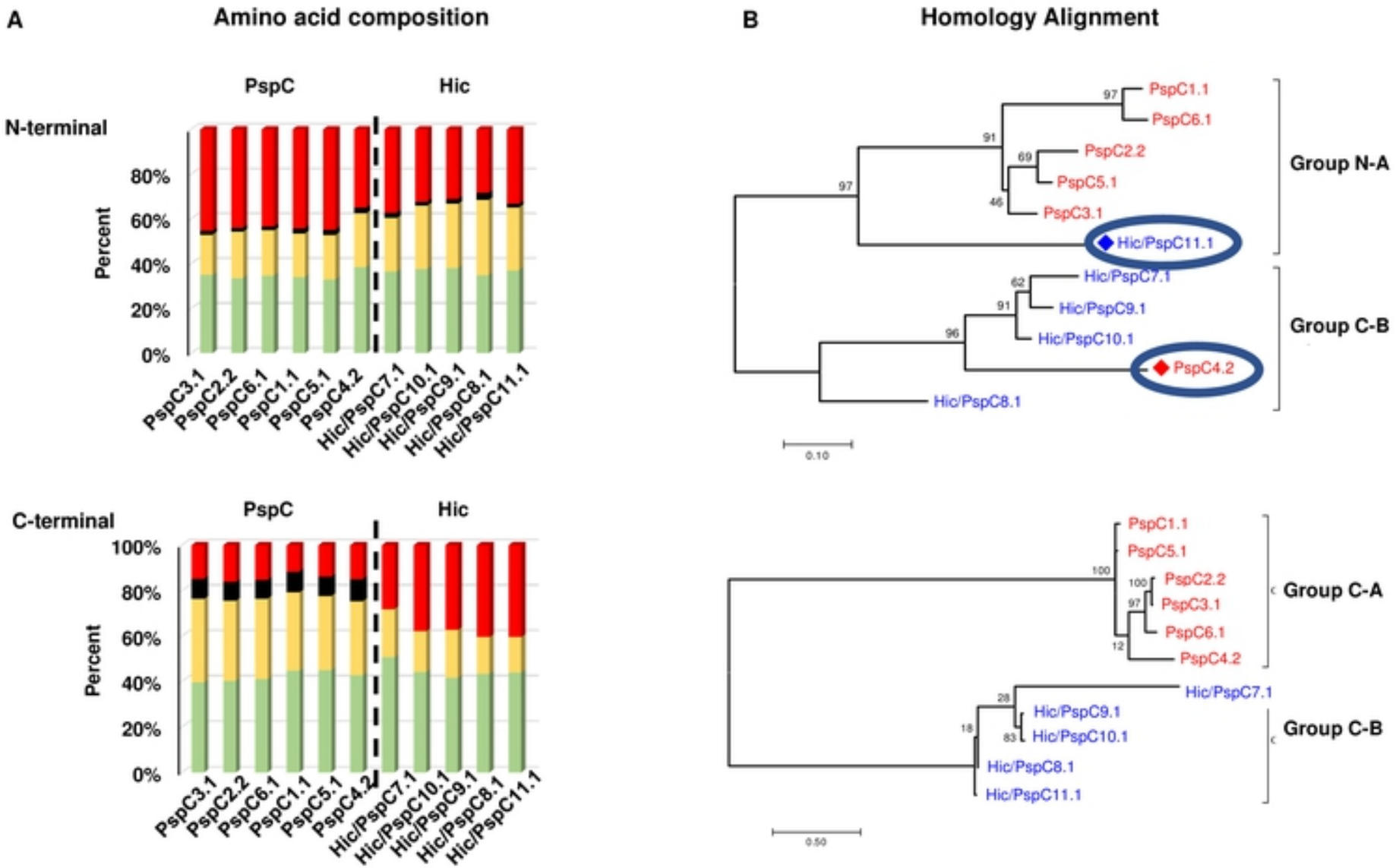
A
Amino acid composition
PspC
Hic
N-terminal
Percent
80%
60%
40%
20%
0%
PspC3.1
PspC2.2
PspC6.1
PspC1.1
PspC5.1
PspC4.2
Hic/PspC7.1
Hic/PspC10.1
Hic/PspC9.1
Hic/PspC8.1
Hic/PspC11.1
C-terminal
100%
B
Homology Alignment
Group N-A
Group C-B
Group C-A
0.10
0.50

Figure 3

Table I: Domain used by *S. pneumoniae* PspC and Hic proteins

#	Region			Domain	Sub domains	Class	n	Module	*Structure*		Comment Host Ligand
1			known	**SP**			**11**				
2	**N-term**		known	**HVD**	HVD-A, HVD-B, HVD-C		**11**		α helix	PspC/Hic specific	Factor H
3			known	**RD**	RD-I, RD-II		**7**		α helix		sIGA/pIgR
					RDII	PspC	**5**		α helix		
4			known	**RCD**			**8**		αhelix		
5		**1**	**new**	**S_nD/GS_2**	3 Positions		**10**		coiled coil		
6		**2**	**new**	**RCE1**		PspC	**2**		α helix		Lactoferrin
7		**3**	**new**	**RCE2**			**2**		A helix		
8		**4**	**new**	**PspA related**		PspC	**2**		α helix	in PspA	
9		**5**	**new**	**R-type**			**3**		α helix		IgA
10		**6**	**new**	**EPRD**			**4**		α helix		
11		**7**	**new**	**IgA**			**4**		α helix	*S. agalactiae*	
12		**8**	**new**	**VS4.2**		PspC	**1**		α helix	Specific	
13		**9**	**new**	**VS11.1**		Hic	**1**		α helix	Specific	
14	**C-term**		known	**PRD**	PRD-IA, PRD-IB	PspC	**5**	**Modular**	coiled coil	also in PspA	Cell wall spanning?
			new		PRD-II	PspC	**1**	**Modular**	coiled coil	?	
			new		PRD-III	Hic	**1**	**Modular**	coiled coil	?	
			new		PRD-IV	Hic	**4**	**Modular**	coiled coil	?	
15			known	**anchor**	CBD	PspC	**6**	**Modular**	β sheets	several	Anchor
16			known		LPsTG	Hic	**5**		coiled coil	many	Anchor

Table I

Table II: Host Regulators binding to *S. pneumoniae* PspC and Hic proteins

Host Regulator	Function	Binding Site
Factor H	Complement Regulation	HVD
sIgA/pIgR	Adhesion	Repeat Domains
C3	C3 Inactivation	Not mapped
C4BP	CP Inhibition	Not mapped
Plasminogen	Proenzyme; plasmin cleaves inactivates C3, C3b and fibrin	Not mapped
Thrombospondin-1	adhesive glycoprotein, cell-cell and cell-matrix interaction	Not mapped
Vitronectin	Complement control & adhesion	Not mapped
Lactoferrin	Fe metabolism	Proposed by homology
IgA	IgA Inactivation?	Proposed by homology

Table II

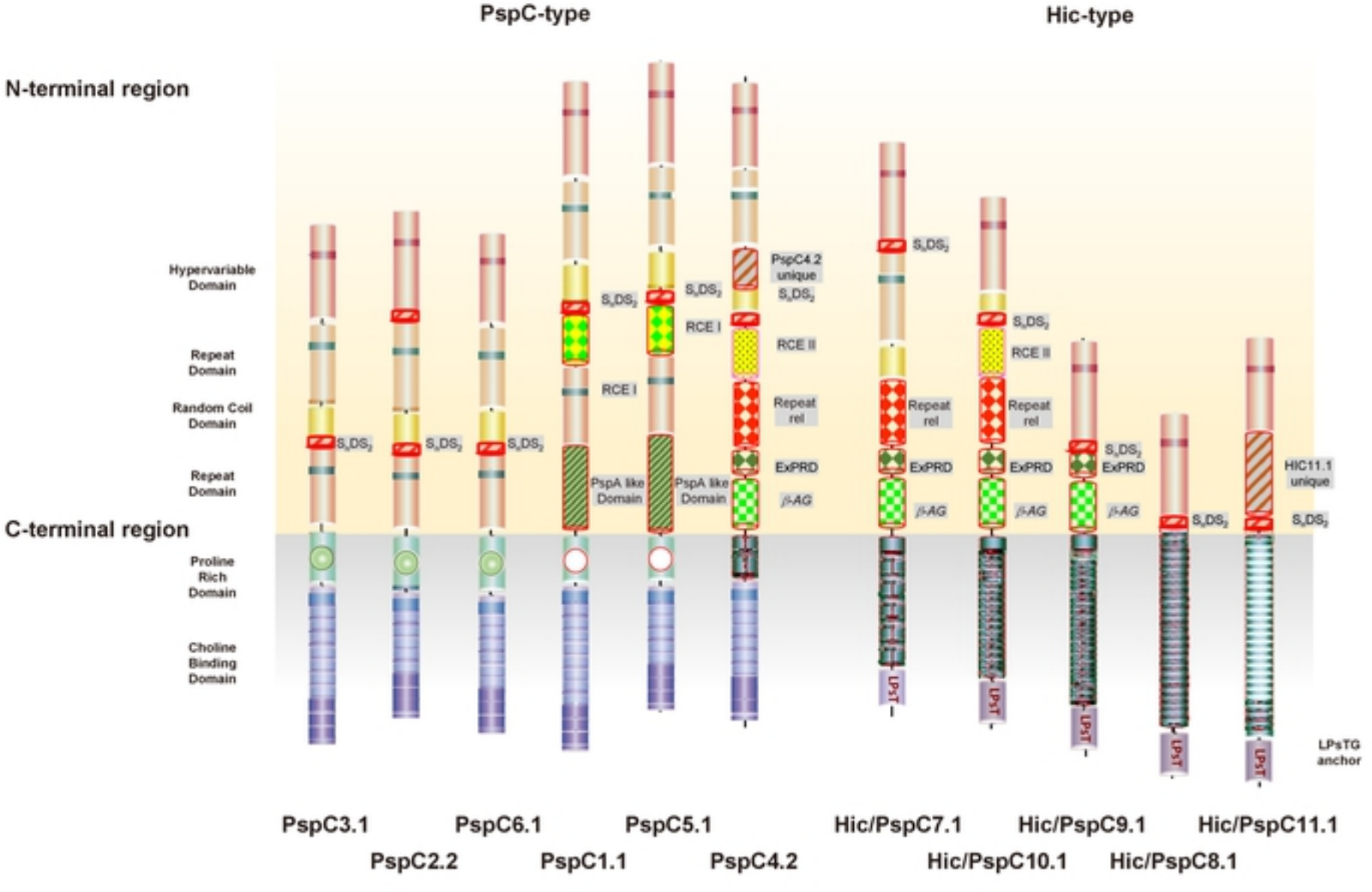
PspC-type
Hic-type
N-terminal region
Hypervariable Domain
Repeat Domain
Random Coil Domain
Repeat Domain
C-terminal region
Proline Rich Domain
Choline Binding Domain
S1DS2
RCE I
RCE II
PspA like Domain
PspC4.2 unique
Repeat rel
ExPRD
β-AG
HIC11.1 unique
LPsT
LPsTG anchor
PspC3.1
PspC6.1
PspC5.1
Hic/PspC7.1
Hic/PspC9.1
Hic/PspC11.1
PspC2.2
PspC1.1
PspC4.2
Hic/PspC10.1
Hic/PspC8.1

Figure 4

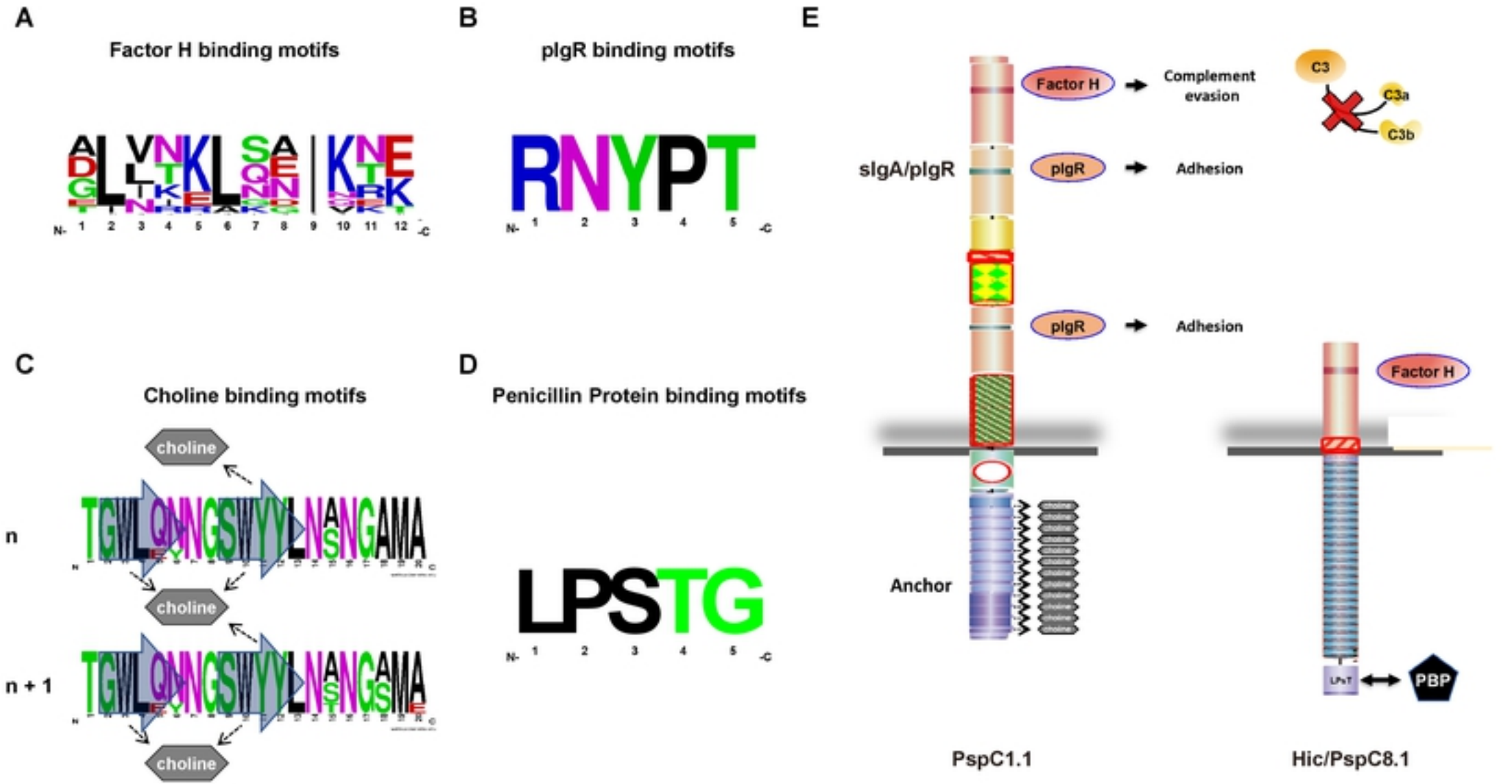
A
Factor H binding motifs
N- 1 2 3 4 5 6 7 8 9 10 11 12 -C
B
pIgR binding motifs
RNYPT
N- 1 2 3 4 5 -C
C
Choline binding motifs
choline
n
choline
n + 1
choline
D
Penicillin Protein binding motifs
LPSTG
N- 1 2 3 4 5 -C
E
Factor H
Complement evasion
C3
C3a
C3b
sIgA/pIgR
pIgR
Adhesion
pIgR
Adhesion
Anchor
Factor H
LPxT
PBP
PspC1.1
Hic/PspC8.1

Figure 5