Bacterial protein domains with a novel Ig-like fold bind human CEACAM receptors

Nina M. van Sorge[1%], Daniel A. Bonsor[2], Liwen Deng[3], Erik Lindahl[4], Verena Schmitt[5], Mykola Lyndin[5,6], Alexej Schmidt[7], Olof R. Nilsson[8], Jaime Brizuela,[9] Elena Boero[1], Eric J. Sundberg[2,10], Jos A.G. van Strijp[1], Kelly S. Doran[3$], Bernhard B. Singer[5$], Gunnar Lindahl[8,11*] & Alex J. McCarthy[1,9*]

[$] equal contributions

* Correspondence and Materials Requests to be made to:-
Alex J. McCarthy (Email: a.mccarthy@imperial.ac.uk, Phone: +44 (0)20 7594 3868)
Gunnar Lindahl (Email: gunnar.lindahl@tmb.lth.se, Phone: +46 735 342255)

[1]University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands.
[2]Institute of Human Virology, University of Maryland School of Medicine, University of Maryland, Baltimore, MD, USA
[3]Department of Immunology & Microbiology, University of Colorado Anschutz Medical Campus, Aurora, CO, United States.
[4]Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden.
[5]Institute of Anatomy, University Hospital, University Duisburg-Essen, Essen, Germany.
[6]Department of Pathology, Sumy State University, Sumy, Ukraine.
[7]Department of Medical Biosciences, Pathology, Umeå University, SE-901 85 Umeå, Sweden
[8]Department of Laboratory Medicine, Division of Medical Microbiology, Lund University, Lund, Sweden
[9]MRC Centre for Molecular Bacteriology & Infection, Imperial College London, London, United Kingdom.

[10]Department of Biochemistry, Emory University School of Medicine, Atlanta, GA, United States

[11]Department of Chemistry, Division of Applied Microbiology, Lund University; SE-22100 Lund, Sweden

[%] Current Address: Department of Medical Microbiology and Infection Prevention, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, The Netherlands.

Abstract

CEACAMs are cell surface proteins with immunoglobulin (Ig) folds that regulate cell communication, cell signalling and immunity. Several human bacterial pathogens express adhesins that bind CEACAMs for enhanced adhesion to epithelial surfaces. Here, we report an adhesin-CEACAM interaction with unique properties. We found that β protein from the human Gram-positive bacterial pathogen *Streptococcus agalactiae* contains a domain with Ig superfamily (IgSF) fold that promotes binding to the N-terminal domain of human CEACAM1 and CEACAM5. A crystal structure of the complex showed that the IgSF domain in β represents a novel Ig-fold subtype called IgI3, in which unique features allow binding to CEACAM1. β-IgI3 homologs in other human Gram-positive bacteria including pathogens, are predicted to maintain the IgI3 fold. We show that one homolog interacts with CEACAM1, suggestive that the IgI3 fold could provide a broad mechanism to target CEACAMs.

Introduction

Human carcinoembryonic antigen-related cell adhesion molecules (CEACAMs) are a family of 12 cell surface receptors belonging to the immunoglobulin (Ig) superfamily (IgSF).[1] Several CEACAMs are expressed on epithelial cells, endothelial cells and leukocytes, [2,3] whilst other CEACAMs have a more restricted expression profile.[4] CEACAMs regulate immune responses through formation of homophilic and heterophilic interactions. These properties allow CEACAMs to regulate multiple physiological and pathophysiological processes including cell-to-cell communication, epithelial differentiation, apoptosis and regulation of pro-inflammatory reactions.[5–7] All CEACAMs, except CEACAM16, are anchored to the cell surface through either a single transmembrane domain or a glycophosphatidylinositol (GPI) anchor. Each CEACAM is composed of an N-terminal domain with V-set (IgV) fold and varying numbers of C2-set (IgC2) folds.[1] Homophilic and heterophilic interactions are mediated by dimerization of the N-terminal IgV folds.[2,8–10]

Several bacterial pathogens interact with human CEACAM receptors.[11] These pathogens are predominately human-specific and include *Neisseria meningitidis*, *N. gonorrhoeae*, *Haemophilus influenzae*, *Moraxella catarrhalis*, *Escherichia coli*, *Helicobacter pylori* and *Fusobacter spp*.[12–19] To date, all pathogens tested bind CEACAM1, and additionally bind CEACAM5 and/or CEACAM6. All bacteria reported to interact with human CEACAMs are Gram-negative organisms. Each of these bacteria expresses a distinct cell surface adhesin that binds human CEACAMs, a property that apparently has been acquired through a process of convergent evolution. Interactions with CEACAMs provide opportunities for the bacteria to enhance adhesion to and invasion of human epithelial surfaces via highly specific protein-protein interactions. However, the immune system has evolved to counteract pathogen exploitation of CEACAMs through CEACAM3, a phagocytic receptor exclusively expressed on neutrophils that can detect and eliminate some CEACAM-binding pathogens.[4,20,21] Of note, all bacterial ligands bind with high-affinity to the CEACAM dimerization interface of the N-terminal IgV domain.[15,18,22–26] However, the structural details of adhesin-CEACAM receptor interactions have only been resolved for HopQ of *H. pylori* and AfaE of *E. coli*.[24,25]

Here, we screened a set of human Gram-positive bacterial pathogens for their ability to interact with human CEACAM1. We report that strains in a subset of *Streptococcus agalactiae* (also known as Group B *Streptococcus*; GBS) interact with CEACAM receptors, the first demonstration that CEACAMs are targeted by a Gram-positive pathogen. The

streptococcal pathogen GBS is the leading cause of life-threatening bacterial infections in the neonatal period,[27] following maternal-to-infant transmission. We found that the GBS surface protein β contains an IgSF domain that binds specifically to human CEACAM1 and CEACAM5. This interaction enhanced adhesion of β-expressing GBS to CEACAM-positive epithelial cells. We employed structural methods to demonstrate that the IgSF domain of β has unique structural features and represents a previously unrecognised variant of the IgI fold, which we term the I3-set fold (IgI3). The unique features include the absence of cysteine residues, absence of *C′* and *C″* strands in the topology and a truncated *C* strand that is directly followed by a 1.5-turn α-helix. It is these features of the β-IgI3 domain that allow it to target the CEACAM1 N-domain dimerization interface. Homologs of β-IgI3 were identified in several human bacteria including pathogens and were predicted to maintain the IgI3 structural features. We demonstrate that one of these IgI3 variants also binds CEACAM1, despite considerable residue variation within the CEACAM binding interface, indicating that the interaction between bacterial IgI3 domains and CEACAMs may be of general importance.

Results

Group B *Streptococcus* binds CEACAM1 through the β surface protein

To date, all human bacterial pathogens known to engage human CEACAM1 are Gram-negative bacteria. Using glycosylated recombinant CEACAM1 (rCEACAM1) (Supplementary Fig. 1A) and a panel of Gram-positive bacterial species, we observed interaction of rCEACAM1 with a subset of GBS strains, but not to any of the other Gram-positive species in the screen including *Streptococcus* spp., *Enterococcus* spp. and *Staphylococcus* spp. (Fig. 1A). We found that CEACAM1 binding corelated with carriage of the *bac* gene, encoding the cell wall-anchored surface β protein that is commonly expressed in serotype Ia, Ib, II and V GBS strains.[28,29] (Fig. 1B). Moreover, β protein expression level, detected using anti-β-serum (Supplementary Fig. 1B), correlated with rCEACAM1-binding capacity (Fig. 1C). Deletion of *bac* in the GBS A909 reference strain abolished rCEACAM1 binding, whilst complementation with a vector carrying the *bac* gene re-established rCEACAM1 binding (Fig. 1D; Supplementary Fig. 1C). Mutation of other GBS surface adhesins or capsule, did not influence interaction with rCEACAM1 (Supplementary Fig. 1C). Additionally, we were able to pull down rCEACAM1 with GBS A909 (Fig. 1E), but not the Δ*bac* strain (Fig. 1F). Direct interaction of rCEACAM1 with purified β protein, but not with purified GBS surface proteins Rib or α, was demonstrated by Western blot analysis (Fig. 1G). Finally, β protein bound to CEACAM1-coated beads, but not control beads, in a concentration-dependent manner (Fig. 1H & 1I). Together, these data unequivocally show that the β protein specifically binds to human CEACAM1.

CEACAM1 binds to an IgSF domain in β protein

β protein interacts with several proteins of the human immune system, including factor H, Siglec-5, Siglec-7, Siglec-14 and IgA, through distinct domains (Fig. 2A).[30–34] To characterise the β protein domain interacting with CEACAM1, we expressed and purified recombinant β protein domains (B6N, IgABR, B6C, IgSF and β75KN) from *E. coli* (Fig. 2B) and tested their interaction with rCEACAM1. These domains include a part of β called IgSF. This is because it was previously proposed to adopt an IgSF fold, [29,35] which are composed of around 100 amino acids comprising two β sheets that pack face-to-face.[36] IgSF folds have been identified in various kingdoms including eukaryotes, prokaryotes, bacteria, viruses, fungi and plants. rCEACAM1 bound in a concentration-dependent manner to dynabeads (DB) coated with the biotinylated IgSF domain, but not to DB coated with other

β protein domains or biotinylated-HSA (Fig 2C). As a control, we included rSiglec-5, which interacted only with DB coated with the B6N domain as described previously (Fig. 2C).[32] To further test the interaction between CEACAM1 and the IgSF domain, we coated DB with rCEACAM1, which interacted with the IgSF domain, but not HSA, coupled to streptavidin (Fig. 2D). Thus, the IgSF domain in β specifically binds to human CEACAM1. Indeed, structural predictions of the β-IgSF domain suggest resemblance to a V-set Ig domain identified in SrpA, a cell-surface anchored protein in *Streptococcus sanguinis* (Fig. 2E).[37] However, rCEACAM1 did not bind to the surface of a SrpA-expressing *S. sanguinis* strain (Supplementary Fig. 2). We denoted the IgSF fold in β as β-IgSF.

CEACAM1 possesses four Ig-like domains (Supplementary Fig. 3A) including an N-terminal IgV-like domain and three IgC2 domains denoted A1, B1 and A2,[1] which is also the domain for homo- and heterophilic interactions implicated in the CEACAM-mediated functions. All bacterial ligands characterized to date target the IgV-like domain. To identify the domain of CEACAM1 that β-IgSF binds, we tested whether domain-specific CEACAM1 monoclonal antibodies (mAb) could inhibit the interaction of β-IgSF with CEACAM1-coated DB. In these assays, only blocking the N domain of CEACAM1 abolished the interaction with β-IgSF (Fig 2F; Supplementary Fig. 3B). Similarly, the *H. pylori* protein HopQ, which binds to the N-terminal IgV-like domain,[26] blocked the β-IgSF-CEACAM1 interaction (Supplementary Fig. 3C). To test for direct interaction with the CEACAM1 N-terminal domain, we coated DB with rCEACAM1-N or rCEACAM1-A1B1A2 (Supplementary Fig. 3D) and observed concentration-dependent binding of β-IgSF only to DB coated with rCEACAM1-N (Fig. 2G). In the reciprocal experiment, rCEACAM1-N, but not rCEACAM1-A1B1A2, bound to DB coated with β-IgC2 (Supplementary Fig. 3E).

<u>The β-IgSF domains binds with high-affinity to human CEACAM1 and CEACAM5</u>

Bacterial adhesins often bind to the N-terminal domain, not only of CEACAM1, but also CEACAM3, CEACAM5 and CEACAM6 reflecting the high sequence (~90% identity) and structural homology.[1,38] However, no bacterial adhesins to date bind CEACAM8. To ascertain the CEACAM binding profile for β-IgSF, we measured the affinity of β-IgSF for unglycosylated N-terminal domains of these five CEACAMs, purified from *E. coli*, by isothermal calorimetry (ITC) (Fig. 3A; Supplementary Table 3). β-IgSF bound with high affinity to rCEACAM1-N (K_D = 96±2 nM) and rCEACAM5-N (K_D = 152±27 nM), but did

not bind to rCEACAM3-N, rCEACAM6-N or rCEACAM8-N. The affinities of (β-IgSF)-(CEACAM1-N) and (β-IgSF)-(CEACAM5-N) interactions were comparable to those reported for other bacterial ligands. [24,39] As the binding of β-IgSF is weaker to the isolated N-terminal domain compared to glycosylated CEACAM1 (Fig. 2G), other CEACAM1 domains may stabilise the interaction, as reported for the HopQ-CEACAM1 interaction.[14] To confirm that the CEACAM binding profile of the protein is valid for GBS bacteria, we tested binding of the full-length glycosylated rCEACAMs to GBS by flow cytometry. The β-expressing GBS A909 strain recognised CEACAM1 and CEACAM5 only (Fig. 3B). Specificity for this subset was mostly confirmed in a wider panel of β-expressing GBS strains (Supplementary Fig. 4), though variation existed in CEACAM5 binding. Although some β-expressing GBS strains did not bind CEACAM5, the β-IgSF domain displayed 100% sequence conservation amongst 57 GBS proteins identified through BLAST search (data not shown), indicating that intrastrain variation in CEACAM5 binding is most likely due to differential β expression.

Expression of human CEACAMs on epithelial cells enhances GBS adhesion

To analyze whether β-expressing GBS can utilize human CEACAMs for cellular adhesion, we tested the binding of strain A909 to well-characterised CEACAM-expressing HeLa cells (Fig. 4A). Consistent with rCEACAM binding, a higher percentage of the A909 inoculum was recovered from the CEACAM1- and CEACAM5-expressing HeLa cells in comparison to all other HeLa cell lines (Fig. 4A). Increased binding of GBS to human CEACAM1-expressing HeLa cells was confirmed by confocal microscopy (Fig. 4B). To ensure that the CEACAM-observed binding was not influenced by the cell line, we assessed GBS adhesion to CHO cells.[40] Also in this case, a higher percentage of the GBS inoculum was recovered from the CEACAM1- and CEACAM5-expressing CHO cells in comparison to all other CHO cell lines (Fig. 4C). For unknown reasons, there was considerable variation in GBS adhesion to the CEACAM1- and CEACAM5-expressing CHO cells. This variation could not be attributed to unstable CEACAM expression, as our cell lines expressed CEACAM at consistent levels (Supplementary Fig. 5A and 5B).

Adhesion of the GBS strain A909 to the CEACAM1-expressing CHO cell was abolished by mutation of the *bac* gene, and the phenotype could be recapitulated by a complemented mutant (Fig. 4D). This result was confirmed by confocal microscopy (Supplementary Fig. 5C). In agreement with the results obtained with pure proteins (Fig. 2F), the binding of β-

expressing GBS to CEACAM1-expressing CHO cells was inhibited by blocking the CEACAM1-N domain with a specific mAb (Supplementary Fig. 5D). Moreover, pre-incubation of A909 with rCEACAM1-N, but not rCEACAM8-N, impaired adhesion to CEACAM1-expressing CHO transfectants (Supplementary Fig. 5E). This indicates that the CEACAM1-N and β-IgSF domains were responsible for the adhesion phenotype.

To rule out the possibility that differences in adhesion of A909 wildtype and Δ*bac* strains to CHO cells reflected interstrain variation in growth rates during adhesion at 37°C, we developed an alternative assay in which we assessed adhesion of FITC-labelled GBS strains to detached cell lines during incubation at 4°C. A higher percentage of CEACAM1-expressing CHO cells were FITC-positive upon incubation with A909 in comparison to control cells (Fig. 4E and 4F). However, this was not observed for CEACAM5-expressing CHO cells, likely because the N domain of CEACAM5 was cleaved during the detachment process as demonstrated by the inability of an anti-CEACAM5-N mAb to bind to these cells (Supplementary Fig. 5A and 5B). As expected, adhesion of FITC-labelled A909 to the CEACAM1-expressing CHO cell was abolished by mutation of *bac* (Fig. 4G; Supplementary Fig. 5F). Collectively, the data on cellular adhesion therefore indicate that β-expressing GBS can utilize human CEACAM receptors for cellular adhesion.

Crystallography reveals that the IgSF domain in β adopts a novel Ig fold, the IgI3 fold

To gain insights into β-IgSF binding mechanisms, we aimed to solve its structure. However, we could only determine the structure in complex with CEACAM1-N at a resolution of 3.0 Å (Fig. 5A). The asymmetric unit contains two molecules of the (β-IgSF)-(CEACAM1-N) complex. The β-IgSF domain has the characteristic features of an Ig fold, principally a pair of β sheets built of anti-parallel β strands that surround a hydrophobic core (Fig. 5B). IgSF domains can be classified into (variable) V-set, (constant) C-set or (intermediate) I-set, with differentiation based on the number and placement of β-strands between the conserved cysteine residue disulphide bridge (Fig. 5C).[41,42] The V-set Ig domain contains ten β strands with four strands found on one sheet (*ABED*) and six strands on the other *(A′ GFCC′C″*). C-set Ig domains lack the *A′* and *C″* strands, and are further grouped into C1 or C2 based on presence or absence of the *D* strand, respectively. I-set Ig domains lack the *C″* strand and are classified into I1 or I2 based on presence or absence of a *D* strand, respectively. The β-IgSF domain has two β-sheets labelled *ABED* and *A′GFC*, with sheets connected by the *BC*, *EF*, *CD* and *AA′* loops (Fig. 5B). Therefore, β-IgSF has an I-set fold topology that most

closely resembles an I1-set domain (Fig. 5C). However, β-IgSF lacks cysteines and disulfide bridges that are characteristic for I-set folds.[41] Furthermore, the β-IgSF domain possesses a truncated *C* strand that is directly followed by 1.5-turn α-helix (Fig. 5C). These features were not observed in the structurally characterised IgI1 domains in the DALI or PDBeFOLD databases that β-IgSF most closely resembled, including macrophage colony stimulating factor 1 (MCS-F) and intracellular adhesion molecules 3 (ICAM-3) (Fig. 5D, 5E & 5F). Therefore, the topology of β-IgSF domain represents a previously unrecognised IgI fold subtype, denoted here as I3-set Ig (IgI3) domain. Accordingly, the domain in β will now be referred to as β-IgI3. The unique features of this domain are i) the absence of cysteine residues, ii) absence of *C′C″*, and iii) a truncated *C* that is directly followed by a 1.5-turn α-helix. Of note, the unique β-IgI3 region stretching from *C* to *D* strand possesses protruding hydrophobic residues, such as F42, located between *C* and α-helix, L46 located in the α-helix, and V53 located in the *CD* loop (Fig 5G).

<u>Identification of interacting residues in the (β-IgI3)–(CEACAM1-N) complex</u>

CEACAMs form their homophilic and heterophilic interactions through interactions promoted by the *A′GFCC′* face in the N-terminal V-set domains (Fig. 5A and Fig. 6A). Bacterial ligands also bind to the *A′GFCC′* face.[15,18,22–26] The *H. pylori* ligand HopQ uses a coupled folding and binding mechanism, [24] and simulated docking indicated that the *E. coli* ligand AfaE binds to the CEACAM dimerization interface through the *BE* strands and *DE* loop of an incomplete Ig fold.[25,43] The structure determination by X-ray crystallography (Fig. 6B; Supplementary Table 4) of the complex of β-IgI3 and CEACAM1-N at 3 Å resolution shows that β-IgI3 binds to the *A′GFCC′* face of CEACAM1 through residues located in the *C* to *D* strand region including the α-helix (Fig. 6C). Fig. 6B shows the interacting residues on the surfaces of the two molecules. Specifically, L46 in the α-helix of β-IgI3 interacts by van der Waals forces with F29 and L95 (distances 3.11Å and 3.61Å) located in the *C* and *G* strands of CEACAM1, respectively (Supplementary Fig. 6A; Supplementary Table 5). In addition, the protruding β-IgI3 residue V53 is within close contact of I91 (distance 3.16Å, *F* strand) and S32 (distance 3.36Å, *C* strand) of CEACAM1, D55 is within close contact of Y34 (*C* to *C'* loop) of CEACAM1, and F42 is in close contact of L95 (distance 2.87Å, *G* strand) of CEACAM1 (Supplementary Table 5). No conformational changes are observed in CEACAM1 upon binding of β-IgI3 (Supplementary Fig. 6B).

Based on the co-crystal data, we hypothesised that β-IgI3 residue L46 was critical for contacting CEACAM1-N via F29 and L95 (Fig. 6C, Supplementary Fig. 6A & 6B). Additionally, we hypothesised that β-IgI3 residue V53 was critical for contacting CEACAM1-N residue I91, and β-IgI3 residue F42 was critical for contacting CEACAM1-N residue L95. We generated alanine mutations in β-IgI3 at these positions as well as at several other sites that contact CEACAM1. ITC binding studies of these β-IgI3 mutants to 'wild-type' unglycosylated CEACAM1-N reveal that the mutant β-IgI3^{L46A} failed to bind to rCEACAM1-N. Three additional mutants, β-IgI3^{L52A}, β-IgI3^{V53A} and β-IgI3^{D55A}, had reduced affinity to bind rCEACAM1-N (K_D = 234±16, 562±44, 690±52 nM, respectively) (Supplementary Fig. 7A; Supplementary Table 6). In contrast, β-IgI3^{F42A} bound with higher affinity (K_D = 16±15 nM). In addition, we tested the ability of unglycosylated rCEACAM1-N to bind DB coated with the β-IgI3 variants (Fig. 6D). In this analysis, rCEACAM1-N displayed significantly reduced binding to DB coated with β-IgI3^{L46A} and β-IgI3^{V53A}, and significantly enhanced binding to DB coated with β-IgI3^{F42A}. Together, these data indicate that the L46, L52, V53 and D55 residues of β-IgI3 are critical for the binding to CEACAM1. These residues were observed to be conserved in 57 β protein sequences (data not shown).

For CEACAM1-N, the crystal structure of the complex indicated that β-IgI3 interacts with the dimer interface, contacting several residues which are important for CEACAM1 homodimerization and for binding to other bacterial adhesins. We hypothesised that residues F29, I91 and L95 of CEACAM1 are critical for contacting β-IgI3 (Fig. 6C, Supplementary Fig. 6A & 6B), and generated alanine mutants at these and several other positions in the homodimerization interface. In ITC analysis with unglycosylated CEACAM1-N proteins, both CEACAM1-N^{F29A} and CEACAM1-N^{I91A} failed to bind β-IgI3 (Supplementary Fig. 7B; Supplementary Table 6). Additionally, CEACAM1-N^{Q89A} also lacked ability to bind β-IgI3. Two mutants, CEACAM1-N^{Q44A} and CEACAM1-N^{L95A}, resulted in a 10-fold decrease in binding affinities (K_D = 996±116 and 1350±460 nM, respectively) whilst two further mutants, CEACAM1-N^{V96A} and CEACAM1-N^{N97A}, displayed only a modest decrease (4-fold) in binding affinities (K_D = 370±4 and 490±120 nM, respectively). In addition, we tested the ability of β-IgI3 coupled to streptavidin to interact with the unglycosylated wildtype and mutants rCEACAM1-N-coated DB (Fig. 6E). There were significant reductions in binding of CEACAM1-N^{F29A}, CEACAM1-N^{Q89A} and CEACAM1-N^{I91A}, as well as for CEACAM1-N^{Q44A} and CEACAM1-N^{L95A}, confirming that residues F29, Q89 and I91 are major targets

of β-IgI3 binding. I91 has also been identified as a critical CEACAM1 residue for interaction with *M. catarrhalis*,[18] *Neisseria spp.*,[22,23] *H. influenzae*,[16] and *Fusobacterium spp*.[15] Though Q89 in CEACAM1 is critical for interaction with β-IgI3 and other bacterial ligands, Q89 was not in close contact with any β-IgI3 residues. It is possible that mutation of CEACAM1 residue Q89 forms a gap that I91 attempts to fill that subsequently prevents its interaction with in β-IgI3 residue V53. In summary, contact of residue L46 in β-IgI3 with CEACAM1 residue F29, residue V53 in β-IgI3 with CEACAM1 residue I91 provide the critical interactions. Additional stability is gained through β-IgI3 residues F42 and D55.

Comparison of β-IgI3- & HopQ-bound CEACAM1 structures

Comparison with the HopQ-CEACAM1 complex, the first shown structure of a bacterial adhesin bound to CEACAM1, reveals that the same set of residues involved in dimerization of CEACAM1 are contacted by both β-IgI3 and HopQ (Supplementary Fig. 6C). However, HopQ does not have Ig-like fold, but contains several α-helices, and its structure is completely unrelated to that of β-IgI3.[24] HopQ uses an intrinsically disordered loop, that folds into a β hairpin and a small helix upon binding to CEACAM1. The β hairpin extends the CEACAM1 *A'GFCC'C''* face whilst the small helix straddles the face (Supplementary Fig. 6C). In contrast, β-IgI3 interacts through the small α helix and loop, with a single residue, L46, critical for β-IgI3 to fit into a pocket on CEACAM1-N formed by F29 and I91 (Fig. 6C). Mutation of any of these residues in β-IgI3 or CEACAM1-N causes a hole in the protein-protein interface or collapse of the pocket, respectively, abolishing binding as observed with our ITC experiments.

β-IgI3 homologs are broadly distributed in human Gram-positive bacteria

To gain a broader perspective on the possible role of the unique β-IgI3 structure in bacterial pathogenesis, we investigated the distribution of this domain in the bacterial kingdom through BLAST analysis. A sequence similar or related to β-IgI3 was identified in 296 proteins predicted to be expressed by human Gram-positive bacteria, including several human pathogens. These sequences formed several different clades upon Maximum Likelihood analysis (Fig. 7A). The IgI3 domain of β was present in a large clade, IgI3 clade I, which also included sequences from *Streptococcus oralis*. A second major clade, clade II, contained sequences from other human streptococcal pathogens including GBS, *S. pyogenes*, *S. dysgalactiae* and *S. mitis*. Additionally, sequences closely related to clade II

were detected in *S. mitis*, *S. milleri* and *S. intermedius*, forming clade III. Finally, many homologs of β-IgI3 were present in proteins from a diverse panel of Gram-positive human pathogens including *S. pneumoniae*, *S. pseudopneumoniae*, and *Gemella haemolysans*. In addition, a homolog was identified in *Gardnerella vaginalis*, a Gram-variable bacterium. The sequences of β-IgI3 homologs were all located in proteins predicted to be surface-localized and anchored in the bacterial cell wall.

We predicted the tertiary structure of 11 representative homolog sequences based on the β-IgI3 crystal structure. These structures maintained the overall IgI3 fold, including an α-helix and loop located between the truncated *C* strand and the *D* strand (Supplementary Fig. 8), except clade XV sequence from *G. vaginalis* that lacked the *C* strand. These data suggest that the IgI3 structure is broadly distributed in bacterial cell wall anchored surface proteins. We focused further analysis on clade II that maintains key IgI3 structures despite sharing only 40% amino acid identity with β-IgI3 (Fig 7B; Supplementary Fig. 9A, 9B & 9C). This clade includes the R28 protein, which is found in *S. pyogenes* and in GBS (where the same protein has been given the name Alp3). R28 is a cell-wall anchored protein that is unrelated to β protein except for the IgI3 domain.[29] To understand the CEACAM1 binding capacity of β-IgI3 homologs, we inspected whether the critical protruding residues located between *C* and *D* strands in β-IgI3 (Fig. 5G) were conserved across the homologs. Though the CEACAM1-binding residues of β-IgI3 (F42, L46, V53 and D55) are not conserved in R28-IgI3 (Fig. 7C), two of the R28-IgI3 residues found at corresponding positions were of hydrophobic nature as in β-IgI3. Interestingly, additional residue variation was evident among other IgI3 domains (Supplementary Fig. 11). Since this result suggested that IgI3 homologs are likely to differ in their properties, it became particularly relevant to analyze whether the divergent β-IgI3 homolog in the R28 protein binds CEACAM1.

<u>β-IgI3 homologs bind human CEACAMs</u>

To characterise whether the IgI3 domain in R28 interacts with CEACAM1, we purified rR28-IgI3 protein domain from *E. coli* and tested interaction with rCEACAM1. Beads coated with biotinylated R28-IgI3 and β-IgI3, but not HSA, interacted with rCEACAM1 (Fig. 7D). In the reverse assay, beads coated with rCEACAM1 interacted with R28-IgI3 and β-IgI3, but not HSA, coupled to streptavidin (Fig. 7E). Taken together with the structural predictions described above, these data indicate that IgI3 folds from a wider

range of Gram-positive bacterial pathogens can interact with human CEACAM1, but individual IgI3 domains may interact with human CEACAM1 differently.

To analyze whether β-IgI3 and R28-IgI3 interact differently with CEACAM1, we simulated the docking of IgI3 domains onto human CEACAM1 and examined the free energy calculations to ascertain the key IgI3 residues in the complex formation (Supplementary Fig. 9C). Residues with binding free energy contributions lower than -2.0 kcal/mol and greater than 2 kcal/mol are identified as key residues and unfavourable residues, respectively. Simulation of β-IgI3 and CEACAM1 docking correctly predicted residues F42, L46 and V53 were key determinants of CEACAM1 binding, whilst D55 was an unfavourable determinant (Fig 7F; Supplementary Fig. 9D). The binding free energy of the (R28-IgI3)-(CEACAM1-N) complexes was strong (-31.02±6.70 kcal/mol, n = 50 simulations). In contrast to β-IgI3, residues located in the α-helix of R28-IgI3 were not identified as key determinants of CEACAM1 binding (Fig 7F; Supplementary Fig. 9E). Instead, critical R28-IgI3 residues (K48, I54, I56 and K59) were located in the *CD* loop and the *D* strand only. Mapping of key residues onto the surface structure suggest that β-IgI3 and R28-IgI3 target CEACAM1 through different faces on the IgI3 fold (Supplementary Fig. 9D and 9E). Thus, IgI3 domains in different bacterial proteins probably interact with human CEACAM1 through mechanisms that are not identical but use the same structural fold.

Discussion

We show here that adhesins from human bacterial pathogens contain a domain with a previously unrecognised Ig-like fold of the I-set. Molecular analysis of this domain in streptococcal β protein demonstrated that it binds to the IgV-like N-terminal domains of human CEACAM1 and CEACAM5, and biochemical analysis demonstrated that the interaction is of high specificity and high-affinity. The bacterial Ig-like fold identified is characterized by an absence of cysteine residues and the presence of a truncated *C* strand that is directly followed by a 1.5-turn α-helix. As this Ig-like fold of the I-set has unique features, it was designated IgI3. The absence of cysteine residues has been reported in Ig folds, [36] but has not been documented in IgI folds to date. In addition, the partial replacement of the *C* strand with an 1.5-turm α-helix was not identified in any other Ig fold in a search of PDB using HHpred.[44] While the I-set fold is the most abundant Ig-fold in human IgSF proteins,[41] I-set folds are rarely found within bacterial proteins as observed through a search of the SMART database.[45] However, our bioinformatic analysis and modelling revealed that the key structural features of IgI3 folds are widely distributed across human Gram-positive bacterial pathogens.

The crystal structure of the complex demonstrated that β-IgI3 targets the dimerization interface of CEACAM1. This is consistent with data for other bacterial ligands of CEACAMs including Opa of *Neisseria spp*, UspA1 of *M. catarrhalis*, HopQ of *H. pylori* and Dr adhesins of *E. coli*. All these adhesins are structurally unique, implying that their ability to bind CEACAMs is an example of convergent evolution.[14,18,19,23,46,47] Opa proteins of *Neisseria spp* are membrane proteins that bind to the CEACAM dimerization interface through two hypervariable loops.[48] UspA1 of *M. catarrhalis* forms a trimeric coiled-coil that binds to CEACAM dimerization interface through a small bend.[18] HopQ of *H. pylori* uses a disordered loop to bind the CEACAM dimerization interface.[24] AfaE of *E. coli* uses an Ig-like domain to target the CEACAM dimerization interface,[25,39,43] though it does not use a complete Ig fold and not an IgI3 fold. These bacterial ligands all bind to the *CC'C"FG* face of CEACAM despite their diverse structures. Likewise, the streptococcal β protein uses the IgI3 domain to bind this CEACAM face, and interaction is achieved through unique modifications in the IgI3 fold. Specifically, the 1.5-turn α-helix in β-IgI3 interacts with the *C* and *F* strands of CEACAM1 using residue L46 for binding, whilst the β-IgI3 *CD* loop is also involved in critical interaction with the *F* strand. Therefore, the unique structural features of the IgI3 domain are responsible for forming CEACAM1 interactions.

The data reported here show, for the first time, that a Gram-positive bacterial pathogen can bind to various human CEACAMs via protein-protein interaction. As Gram-negative bacteria engage human CEACAMs for enhanced cellular adhesion,[12,14,19,23,47,49] it is likely that β-expressing GBS also adhere to human cells via CEACAMs, a conclusion supported by our *in vitro* adhesion assays. Thus, CEACAM1 and CEACAM5 expand the repertoire of human cell surface proteins that GBS may employ for adhesion.[50–52] Further work will be necessary to determine the consequences of the interaction between β-IgI3 and CEACAM1 beyond cell lines.

The β protein of GBS is commonly expressed by serotype Ia, Ib, II and V strains,[29] and increased expression has been associated with virulence.[53] CEACAM1 is an inhibitory receptor that downregulates immune cell activity and functions. As β protein binds CEACAM1, as well as the inhibitory receptors Siglec-5 and -7 that are expressed on leukocytes, we speculate that dual- or multi-engagement of these inhibitory receptors may induce potent immunosuppressive signals through induction of tyrosine phosphorylation.[31,32,34] Moreover, the ability of the β protein to bind other components of the human immune system, *viz*. the plasma proteins IgA and factor H, could enable it to act as a multitool protein that assists epithelial cell adhesion, suppresses immune cell activity and promotes immune evasion. Similarly, the UspA1 protein of *M. catharralis* has multiple binding and interacts with CEACAMs, fibronectin, complement factor C3 and the complement regulator C4BP.[18,54–56]

Interestingly, β-IgI3 homologs are broadly distributed in Gram-positive bacteria including several human pathogens, and structural modelling of representative sequences predicted that the homologs have a structure resembling β-IgI3. Further, there was low conservation of the critical β-IgI3 residues, suggesting that each sequence may possess different properties. Still, studies of one homologous domain with divergent sequence, R28-IgI3 in clade II, demonstrated that it can bind to CEACAMs. Intriguingly, simulated docking suggested that R28-IgI3 and β-IgI3 bind through alternative strategies. As R28-IgI3-like sequences were found in proteins from multiple human pathogens, CEACAM binding may represent an unrecognised adhesion strategy for a range of Gram-positive bacteria.

In summary, our data demonstrate that a domain with unique Ig-like fold, the IgI3 fold, promotes binding of GBS to human CEACAM receptors. Moreover, homologous domains were identified in a variety of Gram-positive bacterial pathogens, suggesting that many Gram-positive bacterial pathogens employ IgI3 domains to engage CEACAMs. Characterisation of

these domains may lead to a better understanding of pathogenicity mechanisms and could pave-the-way for the development of novel therapeutic approaches to prevent infections.

Methods

Bacteria growth conditions

Organisms used in this study are listed in Supplementary Table 1. Bacteria were cultured in Todd-Hewitt Broth (TH; *S. agalactiae*), TH Broth + 1% yeast (*S. pyogenes*), Tryptic Soy Broth (TSB; *S. aureus*) or Lysogeny Broth (LB; *E. coli*) and supplemented with antibiotics as listed in Supplementary Table 1.

Expression and purification of glycosylated and unglycosylated rCEACAMs

Glycosylated CEACAMs were expressed and purified from Expi29F cells (Life Technologies). In brief, gBlocks containing open reading frames (ORFs) encoding the extracellular domain of CEACAMs, and a C terminal LPETGGSHHHHHH tag, were cloned into the pcDNA3.4 expression vector (Invitrogen) (Supplementary Table 2). Recombinant CEACAM-His proteins were expressed in Expi293F cells (Life Technologies) and purified by affinity chromatography (ÄKTA Pure, GE Healthcare Life Sciences) using a Nickel column (GE Healthcare Life Sciences) as previously described.[57] Eluate was dialysed against 300 mM NaCl 50 mM Tris pH 7.8 at 4°C. Unglycosylated CEACAMs were expressed and purified from *E. coli* cultures. In brief, gBlocks containing ORFs encoding the N-terminal or A1B1A2 domains of CEACAM1, and a C terminal LPETGGS-6xHis tag, were cloned into the pRSET-C vector (Supplementary Table 2). Vectors were transformed into the Rosetta Gami (RG) *E. coli* strain. RG strains were cultured in LB supplemented with 100 μg/ml ampicillin and 1 mM D-glucose at 37°C overnight, and subcultured 1:20 into LB supplemented with 1 mM D-glucose at 37°C until OD_{600} = 0.4. Protein expression was induced by culturing for 4 hours at 37°C following the addition of 1 mM IPTG. After lysis of bacteria, proteins were purified using a Nickel column (GE Healthcare Life Sciences) and affinity chromatography (ÄKTA Pure, GE Healthcare Life Sciences). Eluates dialysed against 300 mM NaCl 50 mM Tris pH 7.8 at 4°C. Large scale expression of tag-less CEACAM1-N domain and point mutants for use in isothermal titration calorimetry (ITC) and crystallization were expressed using pET21d vectors in *E. coli* and purified as previously described. [24]

Expression and purification of bacterial proteins

β, α and Rib proteins were purified from GBS cultures as previously described.[58,59] Domains of the β proteins (B6N, IgABR, B6C, IgSF and β75KN) and R28 (R28-IgI3) were cloned and expressed in *E. coli* (Supplementary Table 2). The IgSF domain structure was found to be in a

region overlapping the previously published B6C (amino acids 226-434) and β75KN domains (amino acids 434-788). [34] Therefore, we have updated domain designations as follows:- B6C represents amino acids 226-390, IgSF represents amino acids 391-503, and β75KN represents amino acids 504-788. Briefly, gBlocks containing ORFs and a C-terminal LPETGGS-6xHis tag, were cloned into the pRSET-C vector using BamHI and NdeI. Vectors were transformed into the RG strain. RG strains were cultured, induced, harvested and lysed as described above for unglycosylated CEACAMs. Bacterial supernatants were filtered, supplemented with 10 mM imidazole and passed over a Nickel column (GE Healthcare Life Sciences). Proteins were purified by affinity chromatography (ÄKTA Pure, GE Healthcare Life Sciences), eluted with imidazole and were dialysed against 300 mM NaCl 50 mM Tris pH 7.8 at 4°C.

Binding of glycosylated and non-glycosylated recombinant CEACAM to bacteria

Six x 10^6 of mid-logarithmic phase bacteria were incubated with rCEACAM (1, 3, 5, 6 or 8), rCEACAM1-N or rCEACAM1-A1B1A2 at 4°C for 1 hour with shaking. After washing in PBS + 0.1% bovine serum albumin (BSA), rCEACAM on the bacterial surface was detected by incubation with FITC-conjugated anti-HIS monoclonal antibody (mAb) at 4°C for 1 hour with shaking. Bacterial fluorescence was measured by flow cytometric analysis following washing and fixation in 1% paraformaldehyde (PFA). To detect rCEACAM binding to GBS using pull-down Western blot analysis, bacterial pellets were resuspended in 1x sample buffer and heated to 95°C for 10 mins. Lysates were separated by SDS-PAGE in a 12.5% polyacrylamide gel at 270V, blotted onto nitrocellulose membranes and probed with anti-CEACAM1 (clone C51X/8) mAb. Membranes were probed using rabbit anti-mouse-IgG-HRP and developed using ECL substrate.

Measurement of β protein expression by GBS strains

Rabbit antiserum was raised against β protein as previously described. [58] Six x 10^6 of mid-logarithmic phase bacteria were incubated with heat-inactivated 0.1% rabbit anti-β serum or normal rabbit serum at 4°C for 1 hour with shaking, washed and incubated in the presence of PE-conjugated goat anti-rabbit-IgG at 4°C for 1 hour with shaking. Fluorescence of bacteria was measured by flow cytometric analysis after washing and fixation in 1% PFA.

Detection of CEACAM1 and β protein interaction

Western blot analysis

Purified GBS proteins, or β protein domains (B6N, IgABR, B6C, IgSF or β75KN), were separated by non-reducing SDS-PAGE, transferred to nitrocellulose membranes, blocked with 4% non-fat dried milk overnight at 4°C, probed with 10 μg/mL rCEACAM1-His (4°C) and horse radish peroxidase (HRP)-conjugated mouse anti-His-IgG, and developed using Pierce ECL Western Blot Substrate Reagent (ThermoFisher Scientific).

Binding of bacterial proteins to CEACAM1-coated dynabeads

rCEACAM-His was attached to nickel dynabeads (DB; Dynabead His-tag pull-down & isolation, Invitrogen) following standard protocols. Presence of CEACAM1 was confirmed by incubating DB in the presence of 5 μg/mL anti-CEACAM or isotype control mAb, and detection with FITC-conjugated mouse anti-His-IgG mAb. To detect binding of purified GBS proteins, 40 μg of DB was incubated in the presence of 10 μg/mL of β protein and probed with 0.1% mouse antiserum or normal mouse serum, and secondary goat anti-mouse-IgG-PE mAb. Biotinylated proteins (including β-IgSF-biotin or HSA-biotin amongst others) were incubated with PE-conjugated Streptavidin on ice for 10 mins. This generates biotin-streptavidin complexes on the remaining 3 binding pockets of streptavidin. To detect binding of proteins coupled with streptavidin (β-IgSF-biotin or HSA-biotin formed with Streptavidin-PE), 20 μg of DB was incubated with varying concentrations of protein-streptavidin. All dilutions and DB washes employed PBS + 0.1% BSA, all incubations were performed for 1 hour at 4°C, and all fixations used 1% PFA. Fluorescence of DB was measured by flow cytometry. For inhibition experiments, CEACAM-coated DB were pre-incubated with 5 μg/ml anti-CEACAM1-N (clone CC1/3/5-Sab, LeukoCom, Essen, Germany) or anti-CEACAM1-A1B1 mAb (clone B3-17, LeukoCom, Essen, Germany) or 10 μg/mL HopQ for 1 hour at 4°C.

Binding of CEACAM to bacterial protein coated dynabeads

Proteins with C-terminal biotin tags were attached to streptavidin-coated dynabeads (Dynabead M-280 Streptavidin, Invitrogen) following standard procedures. To detect binding of rCEACAM-His or rCEACAM-N-His, 3 x 10^5 DB were incubated in the presence of 9 μL of proteins (concentrations range 0.1 to 30 μg/mL), and probed with FITC-conjugated mouse anti-His-IgG. In the case of rSiglec-5-Fc, binding was detected using anti-IgG-AF488. All dilutions and DB washes employed PBS + 0.1% BSA, all incubations were performed for 1 hour at 4°C, and all fixations used 1% PFA. Fluorescence of DB was measured by flow cytometry.

<u>Isothermal titration calorimetry</u>

ITC measurements were performed on an iTC200 instrument (GE Healthcare). Typically 500 µM of unglycosylated CEACAM-N domains were loaded into the syringe of the calorimeter and 50 µM β-IgC2 protein was loaded into the syringe. All measurements were performed at 25 °C, with a stirring speed of 750 rpm., in 30 mM Tris-HCl, 150 mM NaCl, pH 7.5. Data were analyzed using the Origin 7.0 software.

Adhesion of GBS to epithelial cells

HeLa and Chinese Hamster Ovary (CHO) cell lines expressing CEACAMs were cultured as previously described, [40,47,60] and seeded into 24-well tissue culture plates. Tightly confluent monolayers were infected with mid-logarithmic phase (absorbance at $OD_{600} = 0.4$) GBS strains at a multiplicity of infection (MOI) of 10. Assays were commenced by centrifugation for 2 minutes at 700 rpm, and incubated at 37ºC with 5% CO_2 for 30 mins. Cells were washed five times with DMEM, detached with 0.25% trypsin, and lysed (PBS + 0.025% Triton-X). Adherent bacteria were enumerated using serial dilution plating and growth on Todd-Hewitt agar plates. In specific assays, cells were pre-incubated with 5 µg/mL mAb (anti-CEACAM1-N or anti-CEACAM1-A1B mAb) for 60 mins, or bacteria were pre-incubated with rCEACAM1-N for 60 minutes before co-culture. For confocal microscopy, experiments were performed with FITC-labelled GBS at a MOI of 10 for 30 minutes. After washing five times with PBS, monolayers were fixed with 4% PFA overnight prior to staining of nuclei with DaPI (Sigma Aldrich) and cell membranes with AF647-conjugated wheat germ agglutinin (ThermoFisher Scientific). To perform assays with detached CHO cell lines, adherent monolayers were detached by incubation at 37ºC for 5 mins with Accutase Cell Detachment Solution (BioLegend). After washing, cells were re-suspended to 3×10^6 cells/ml in DMEM containing 10% FCS. To measure CEACAM expression, cells were incubated with mouse mAb clone CC1/3/5-Sab (detects N domain of human CC1, CC3 and CC5) or 5C8C4 (detects A1B1A2B2A3B3 domain of human CC5; LeukoCom, Essen, Germany) for 30 minutes, washed and incubated with PE-conjugated rabbit anti-mouse-IgG. To measure GBS binding, cells were incubated with FITC-labelled GBS at a MOI of 10 for 30 minutes at 4ºC. After washing with PBS, cells were fixed in 4% PFA and fluorescence was measured by flow cytometry.

Crystallography of unglycosylated CEACAM1-N and β-IgSF

Initially, the β-IgSF & CEACAM1-N complex was prepared by mixing the proteins with a slight excess of CEACAM1-N (1:1.1 ratio). Protein was concentrated using an amicon Ultra-4 10,000 NMWL before purification by size exclusion chromatography using a Superdex 200

column (GE Healthcare) equilibrated with 20 mM Tris, 150 mM NaCl, pH 7.5. The complex was concentrated to 9 mg/mL and screened against the commercial Wizards I/II/III/IV (Rigaku) and JCSG^{+} (QIAgen) screens using a Crystal Gryphon Protein Crystallography System (Art Robbins Instruments). Crystals were found in several conditions but diffraction was never found beyond 8.0 Å in the best condition of 1.8 M Ammonium Sulfate, 0.1 M Sodium Citrate pH 5.5. β-IgSF contains both a C-terminal Sortase and 6xHis tag. This was removed by first incubating β-IgSF by itself with Carboxypeptidase A (Sigma). 2 mL of 6 mg mL^{-1} of β-IgSF was incubated with 1:100 w/w of Carboxypeptidase A for 24 hours at 4 °C before the addition of another 1:100 w/w of Carboxypeptidase A and incubation for 24 hours at room temperature. The reaction was terminated by the addition of EDTA to a final concentration of 5 mM. The protein was concentrated using an Amicon Ultra-4 10,000 NMWL before purification by size exclusion chromatography using a Superdex 200 column (GE Healthcare) equilibrated with 20 mM Tris, 150 mM NaCl, pH 7.5. CEACAM1-N was added in a slight excess (1:1.05 ratio) and purified by size exclusion as described above. The complex was concentrated and screened again at a concentration of 7.5 mg mL^{-1}. Crystals again formed in the condition of 1.8 M Ammonium Sulfate, 0.1 M Sodium Citrate pH 5.5, however their morphology were different. Crystals were cyroprotected in 30 % v/v glycerol. Anisotropic diffraction was typically observed with diffraction to ~3.0 Å in two directions and 4.0 Å in the other. Two datasets were collected; (i) beamline ID23-D at the Advance Photon Source and (ii) beamline 12-2 at the Stanford Synchrotron Radiation Lightsource. Data was processed and indexed by XDS, merged together with BLEND before scaling and conversion to structure factors using Aimless. Matthew's coefficient suggests five complexes in the asymmetric unit. Molecular replacement using MOLREP and CEACAM1-N (PDB entry 2gk2) as a search model in MOLREP only found two CEACAM1 molecules. Two copies of the complex would suggest a high solvent content (78 %). Refinement using REFMAC5 and visualization in Coot, showed that CEACAM1-N was well placed and residual electron density was present for the β-IgI3 protein. Density modification was performed using Parrot (solvent flattening, histogram matching and NCS averaging) to improve the maps. The β-IgSF backbone was built manually. A PDBeFold search of the backbone showed that it was an IgI-like fold, with the closest fold by R.M.S.D being the IgC2 domain of Perlecan (PDB entry 1gl4). The Ig fold of β-IgSF possesses a previous unrecognised modification (Fig 5B and 5C). We denote this as the I3-set domain, and the domain with β protein as β-IgI3. This was used to successfully position sidechains, lock the registry and provide extra restraints during refinement in REFMAC5 with

ProSMART. The (β-IgI3)-(CEACAM1-N) complex has been deposited to the PDB with the entry code, 6V3P. Atom contacts in the (β-IgI3)-(CEACAM1-N) complex interface were identified using NCONT (CCP4) with a cut-off of 4.0Å.[61]

Bioinformatics analysis

BLAST analysis of the β-IgI3 amino acid sequence was performed to identify homologs in bacterial proteins, and subsequently aligned using ClustalW. Phylogenetic analysis was performed using Maximum Likelihood approach and 1000 bootstrap replications in MEGA, [62] and the resulting tree displayed with interactive tree of life version 4. [63] The structure of IgI3 homologs was predicted using the β-IgI3 structure as input in SWISS-MODEL. [44] Only models passing a QMEAN score of < -4.00 were further analyzed. Docking of β-IgI3 structure or the R28-IgI3 predicted structure to CEACAM1-N was simulated 50 times using ZDOCK server,[64] in which CEACAM1-N residues F29, Q44, Q89, I91 and N95 were included as contact sites. The free energy contribution of each simulation was interpedently calculated using MM/GBSA analysis on the HAWKDOCK server.[65]

Data Availability

The co-crystal structure of β-IgI3 and CEACAM1-N is available at the PDB with ID: 6V3P. All other data supporting the findings of this study are available within the paper and its supplementary information files and are available from the corresponding author on reasonable request.

References

1. Gray-Owen, S. D. & Blumberg, R. S. CEACAM1: Contact-dependent control of immunity. *Nature Reviews Immunology* **6**, 433–446 (2006).
2. Bonsor, D. A., Günther, S., Beadenkopf, R., Beckett, D. & Sundberg, E. J. Diverse oligomeric states of CEACAM IgV domains. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 13561–13566 (2015).
3. Kuespert, K., Pils, S. & Hauck, C. R. CEACAMs: their role in physiology and pathophysiology. *Current Opinion in Cell Biology* **18**, 565–571 (2006).
4. Schmitter, T., Agerer, F., Peterson, L., Münzner, P. & Hauck, C. R. Granulocyte CEACAM3 Is a Phagocytic Receptor of the Innate Immune System that Mediates Recognition and Elimination of Human-specific Pathogens. *Journal of Experimental Medicine* **199**, 35–46 (2004).
5. Khairnar, V. *et al.* CEACAM1 promotes CD8+ T cell responses and improves control of a chronic viral infection. *Nature Communications* **9**, 2561 (2018).
6. Khairnar, V. *et al.* CEACAM1 induces B-cell survival and is essential for protective antiviral antibody production. *Nature Communications* **6**, 6217 (2015).
7. Helfrich, I. & Singer, B. B. Size matters: The functional role of the CEACAM1 isoform signature and its impact for NK cell-mediated killing in melanoma. *Cancers* **11**, E356 (2019).
8. Bonsor, D. A., Beckett, D. & Sundberg, E. J. Structure of the N-terminal dimerization domain of CEACAM7. *Acta Crystallographica Section:F Structural Biology Communications* **71**, 1169–1175 (2015).
9. Watt, S. M. *et al.* Homophilic adhesion of human CEACAM1 involves N-terminal domain interactions: structural analysis of the binding site. *Blood* **98**, 1469–79 (2001).
10. Zhou, H., Fuks, A., Alcaraz, G., Bolling, T. J. & Stanners, C. P. Homophilic Adhesion between Ig Superfamily Carcinoembryonic Antigen Molecules Involves Double Reciprocal Bonds. *Journal of Cell Biology* **122**, 951–960 (1993).
11. Tchoupa, A. K., Schuhmacher, T. & Hauck, C. R. Signaling by epithelial members of the CEACAM family - Mucosal docking sites for pathogenic bacteria. *Cell Communication and Signaling* **12**, 27 (2014).
12. Königer, V. *et al.* Helicobacter pylori exploits human CEACAMs via HopQ for adherence and translocation of CagA. *Nature Microbiology* **2**, 16188 (2016).
13. Virji, M., Makepeace, K., Ferguson, D. J. P. & Watt, S. M. Carcinoembryonic antigens (CD66) on epithelial cells and neutrophils are receptors for Opa proteins of pathogenic neisseriae. *Molecular Microbiology* **22**, 941–50 (1996).
14. Javaheri, A. *et al.* Helicobacter pylori adhesin HopQ engages in a virulence-enhancing interaction with human CEACAMs. *Nature Microbiology* **2**, 16189 (2016).
15. Brewer, M. L. *et al.* Fusobacterium spp. target human CEACAM1 via the trimeric autotransporter adhesin CbpF. *Journal of Oral Microbiology* **11**, 1565043 (2019).
16. Hill, D. J. *et al.* The variable P5 proteins of typeable and non-typeable Haemophilus influenzae target human CEACAM1. *Molecular Microbiology* **39**, 850–862 (2001).
17. Chen, T. & Gotschlich, E. C. CGM1a antigen of neutrophils, a receptor of gonococcal opacity proteins. *Proc Natl Acad Sci U S A* **93**, 14851–14856 (1996).
18. Conners, R. *et al.* The Moraxella adhesin UspA1 binds to its human CEACAM1 receptor by a deformable trimeric coiled-coil. *EMBO Journal* **27**, 1779–1789 (2008).
19. Tchoupa, A. K., Lichtenegger, S., Reidl, J. & Hauck, C. R. Outer membrane protein P1 is the CEACAM-binding adhesin of Haemophilus influenzae. *Molecular Microbiology* **98**, 440–455 (2015).

20. Adrian, J., Bonsignore, P., Hammer, S., Frickey, T. & Hauck, C. R. Adaptation to Host-Specific Bacterial Pathogens Drives Rapid Evolution of a Human Innate Immune Receptor. *Current Biology* **29**, 616–630 (2019).
21. Heinrich, A. *et al.* Moraxella catarrhalis induces CEACAM3-Syk-CARD9-dependent activation of human granulocytes. *Cellular Microbiology* **18**, 1570–1582 (2016).
22. Villullas, S., Hill, D. J., Sessions, R. B., Rea, J. & Virji, M. Mutational analysis of human CEACAM1: The potential of receptor polymorphism in increasing host susceptibility to bacterial infection. *Cellular Microbiology* **9**, 329–346 (2007).
23. Virji, M. *et al.* Critical determinants of host receptor targeting by Neisseria meningitidis and Neisseria gonorrhoeae : identification of Opa adhesiotopes on the N-domain of CD66 molecules. *Molecular Microbiology* **34**, 538–551 (1999).
24. Bonsor, D. A. *et al.* The Helicobacter pylori adhesin protein HopQ exploits the dimer interface of human CEACAMs to facilitate translocation of the oncoprotein CagA. *The EMBO Journal* **37**, e98664 (2018).
25. Korotkova, N. *et al.* Binding of Dr adhesins of Escherichia coli to carcinoembryonic antigen triggers receptor dissociation. *Molecular Microbiology* **67**, 420–434 (2008).
26. Moonens, K. *et al.* Helicobacter pylori adhesin HopQ disrupts trans dimerization in human CEACAMs. *The EMBO Journal* **37**, e98665 (2018).
27. Seale, A. C. *et al.* Estimates of the Burden of Group B Streptococcal Disease Worldwide for Pregnant Women, Stillbirths, and Children. *Clinical Infectious Diseases* **65**, S200–S219 (2017).
28. Hedén, L.-O., Frithz, E. & Lindahl, G. Molecular characterization of an IgA receptor from group B streptococci: sequence of the gene, identification of a proline-rich region with unique structure and isolation of N-terminal fragments with IgA-binding capacity. *European Journal of Immunology* **21**, 1481–1490 (1991).
29. Lindahl, G., Stålhammar-Carlemalm, M. & Areschoug, T. Surface proteins of Streptococcus agalactiae and related proteins in other bacterial pathogens. *Clinical Microbiology Reviews* **18**, 102–127 (2005).
30. Russell-Jones, G. J., Gotschlich, E. C. & Blake, M. S. A surface receptor specific for human IgA on group B streptococci possessing the Ibc protein antigen. *Journal of Experimental Medicine* **160**, 1467–75. (1984).
31. Fong, J. J. *et al.* Siglec-7 engagement by GBS β-protein suppresses pyroptotic cell death of natural killer cells. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 10410–10415 (2018).
32. Carlin, A. F. *et al.* Group B Streptococcus suppression of phagocyte functions by protein-mediated engagement of human Siglec-5. *Journal of Experimental Medicine* **206**, 1691–1699 (2009).
33. Areschoug, T., Stålhammar-Carlemalm, M., Karlsson, I. & Lindahl, G. Streptococcal β protein has separate binding sites for human factor H and IgA-Fc. *Journal of Biological Chemistry* **277**, 12642–12648 (2002).
34. Nordström, T. *et al.* Human Siglec-5 inhibitory receptor and immunoglobulin A (IgA) have separate binding sites in streptococcal β protein. *Journal of Biological Chemistry* **286**, 33981–33991 (2011).
35. Bateman, A., Eddy, S. R. & Chothia, C. Members of the immunoglobulin superfamily in bacteria. *Protein Science* **5**, 1 (1996).
36. Halaby, D. M. & Mornon, J. P. E. The Immunoglobulin Superfamily: An Insight on Its Tissular, Species, and Functional Diversity. *Journal of Molecular Evolution* **46**, 389–400 (1998).
37. Bensing, B. A. *et al.* Structural basis for sialoglycan binding by the streptococcus sanguinis SrpA adhesin. *Journal of Biological Chemistry* **291**, 7230–7240 (2016).

38. Popp, A., Dehio, C., Grunert, F., Meyer, T. F. & Gray-Owen, S. D. Molecular Analysis of Neisserial Opa Protein Interactions with the CEA Family of Receptors: Identification of Determinants Contributing to the Differential Specificities of Binding. *Cellular Microbiology* **1**, 169–181 (1999).
39. Korotkova, N. *et al.* Binding of Dr adhesins of Escherichia coli to carcinoembryonic antigen triggers receptor dissociation. *Molecular Microbiology* **67**, 420–434 (2008).
40. Hollandsworth, H. M. *et al.* Anti-carcinoembryonic antigen-related cell adhesion molecule antibody for fluorescent visualization of primary colon cancer and metastases in patient-derived orthotopic xenograft mouse models. *Oncotarget* **11**, 429–439 (2020).
41. Wang, J. H. The sequence signature of an Ig-fold. *Protein and Cell* **4**, 569–572 (2013).
42. Wang, J.-H. & Springer, T. A. Structural specializations of immunoglobulin superfamily members for adhesion to integrins and viruses. *Immunological Reviews* **163**, 197–215 (1998).
43. Anderson, K. L. *et al.* An Atomic Resolution Model for Assembly, Architecture, and Function of the Dr Adhesins. *Molecular Cell* **15**, 647–657 (2004).
44. Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology* **430**, 2237–2243 (2018).
45. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research* **46**, D493–D496 (2018).
46. Pettigrew, D. *et al.* High resolution studies of the Afa/Dr adhesin DraE and its interaction with chloramphenicol. *Journal of Biological Chemistry* **279**, 46851–46857 (2004).
47. Bos, M. P., Kuroki, M., Krop-Watorek, A., Hogan, D. & Belland, R. J. CD66 receptor specificity exhibited by neisserial Opa variants is controlled by protein determinants in CD66 N-domains. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 9584–9589 (1998).
48. Billker, O., Popp, A., Gray-Owen, S. D. & Meyer, T. F. The structural basis of CEACAM-receptor targeting by neisserial Opa proteins. *Trends in Microbiology* **8**, 258–260 (2000).
49. Gutbier, B. *et al.* Moraxella catarrhalis induces an immune response in the murine lung that is independent of human CEACAM5 expression and long-term smoke exposure. *American Journal of Physiology - Gastrointestinal and Liver Physiology* **309**, 250–261 (2015).
50. Deng, L. *et al.* The group B streptococcal surface antigen I/II protein, BspC, interacts with host vimentin to promote adherence to brain endothelium and inflammation during the pathogenesis of meningitis. *PLoS Pathogens* **15**, (2019).
51. Bolduc, G. R. & Madoff, L. C. The group B streptococcal alpha C protein binds α1β1-integrin through a novel KTD motif that promotes internalization of GBS within human epithelial cells. *Microbiology* **153**, 4039–4049 (2007).
52. Samen, U., Eikmanns, B. J., Reinscheid, D. J. & Borges, F. The surface protein Srr-1 of Streptococcus agalactiae binds human keratin 4 and promotes adherence to epithelial HEp-2 cells. *Infection and Immunity* **75**, 5405–5414 (2007).
53. Nagano, N., Nagano, Y., Nakano, R., Okamoto, R. & Inoue, M. Genetic diversity of the C protein β-antigen gene and its upstream regions within clonally related groups of type la and lb group B streptococci. *Microbiology* **152**, 771–778 (2006).
54. Tong Tan, T., se Nordström, T., Forsgren, A. & Riesbeck, K. The Respiratory Pathogen Moraxella catarrhalis Adheres to Epithelial Cells by Interacting with

Fibronectin through Ubiquitous Surface Proteins A1 and A2. *Journal of Infectious Disease* **192**, 1029–138 (2005).

55. Nordström, T., Blom, A. M., Tan, T. T., Forsgren, A. & Riesbeck, K. Ionic Binding of C3 to the Human Pathogen Moraxella catarrhalis Is a Unique Mechanism for Combating Innate Immunity . *The Journal of Immunology* **175**, 3628–3636 (2005).
56. Nordström, T., Blom, A. M., Forsgren, A. & Riesbeck, K. The Emerging Pathogen Moraxella catarrhalis Interacts with Complement Inhibitor C4b Binding Protein through Ubiquitous Surface Proteins A1 and A2 . *The Journal of Immunology* **173**, 4598–4606 (2004).
57. Zhao Y *et al.* The Orphan Immune Receptor LILRB3 Modulates Fc Receptor-Mediated Functions of Neutrophils. *Journal of Immunology* **204**, 954–966 (2020).
58. Lindahl, G., Akerström, B., Vaerman, J.-P. & Stenberg, L. Characterization of an IgA receptor from group B streptococci: specificity for serum IgA. *European Journal of Immunology* **20**, 2241–2247 (1990).
59. Stålhammar-Carlemalm, M., Stenberg, L. & Lindahl, G. Protein Rib: A Novel Group B Streptococcal Cell Surface Protein that Confers Protective Immunity and Is Expressed by Most Strains Causing Invasive Infections. *Journal of Experimental Medicine* **177**, 1593–603 (1993).
60. Daniel, S. *et al.* Determination of the Specificities of Monoclonal Antibodies Recognizing Members of the CEA Family Using a Panel of Transfectants. *International Journal of Cancer* **55**, 303–310 (1993).
61. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography* **67**, 235–242 (2011).
62. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular biology and evolution* **33**, 1870–1874 (2016).
63. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* **47**, W256–W259 (2019).
64. Pierce, B. G. *et al.* ZDOCK server: Interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* **30**, 1771–1773 (2014).
65. Weng, G. *et al.* HawkDock: a web server to predict and analyze the protein-protein complex based on computational docking and MM/GBSA. *Nucleic Acids Research* **47**, W322–W330 (2019).
66. Areschoug, T., Linse, S., Stålhammar-Carlemalm, M., Hedén, L.-O. & Lindahl, G. A Proline-Rich Region with a Highly Periodic Sequence in Streptococcal Protein Adopts the Polyproline II Structure and Is Exposed on the Bacterial Surface. *Journal of Bacteriology* **184**, 6376–6383 (2002).
67. Webb, B. & Sali, A. Comparative protein structure modeling using MODELLER. *Current Protocols in Bioinformatics* **2016**, Unit-5.6 (2016).

Acknowledgments

The authors thank Carla J.C. de Haas, Piet Aerts, Kok P.M. van Kessel (UMC Utrecht, Utrecht, The Netherlands), Birgit Maranca-Hüwel, Bärbel Gobs-Hevelke (University of Duisberg-Essen, Germany) and Margaretha Stålhammar-Carlemalm (Lund University, Sweden) for excellent technical support and invaluable help. We thank Paul Sullam (University of California, USA) for providing *S. sanguinis* strains. We also wish to thank

the support staff of Beamlines of 12-2 and ID23-D at the Stanford Synchrotron Radiation Lightsource and the Advanced Photon Source for their aid in data collection, respectively. We thank Irene M. Mavridis (Institute of Child Health, Greece) for critical reading of the manuscript. This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement 700862, Deutsche Forschungsgemeinschaft DFG grant SI-1558/3-1, The Swedish Research Council, The Foundation Olle Engkvist Byggmästare, and the National Institutes of Health R01 NS116716 (to K.S.D.)

Author Contributions

Conception & design of the work (N.M.V., L.D., D.A.B., E.J.S., J.A.G.V., K.S.D., B.B.S., G.L., A.J.M.). Acquisition, analysis or interpretation of data (N.M.V., L.D., D.A.B., J.B., V.S., M.L., A.S., O.R.N., E.B., E.L., J.A.G.V., K.S.D., B.B.S., G.L., A.J.M.). Drafting of manuscript (N.M.V., L.D., D.A.B., J.A.G.V., K.S.D., B.B.S., G.L., A.J.M.). All authors read and commented on the final version of the manuscript.

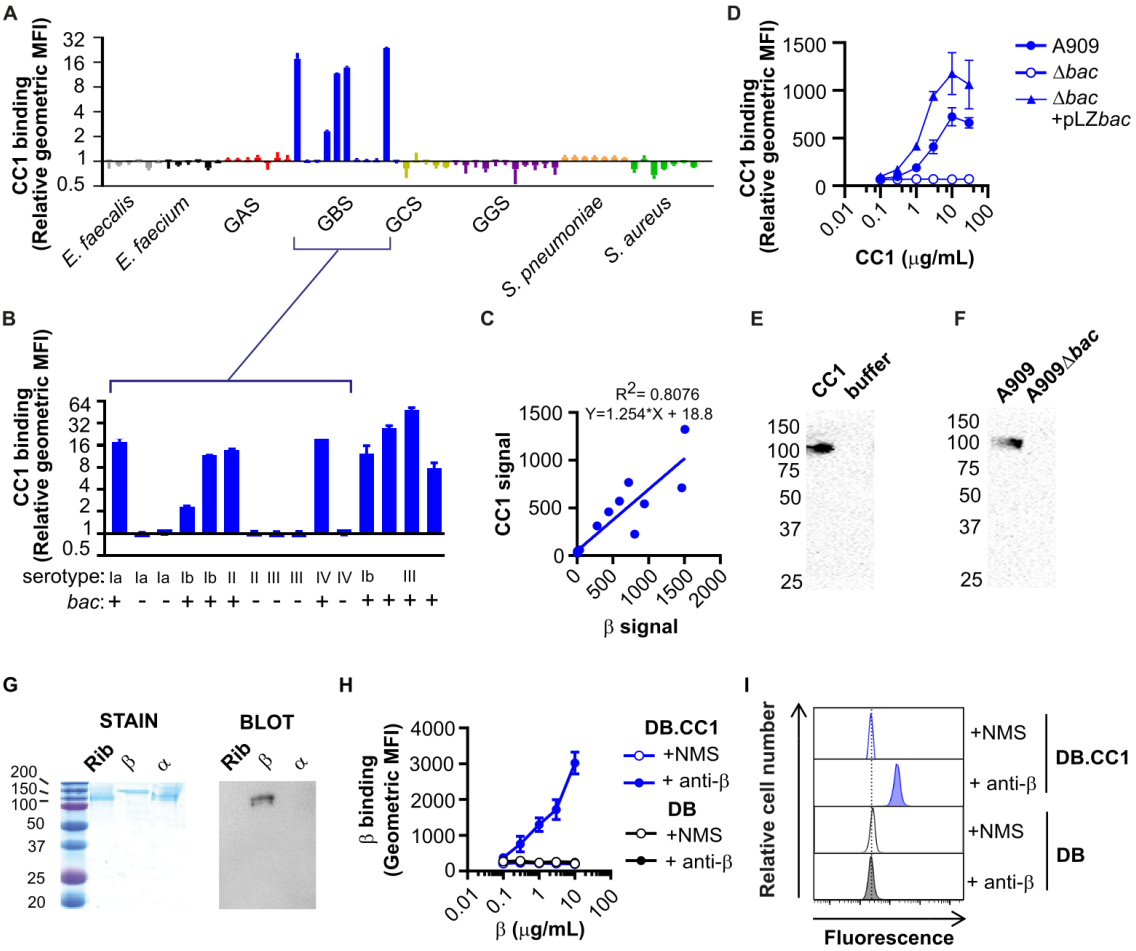

Figure 1: Human CEACAM1 binds to group B *Streptococcus* (GBS) via the surface expressed β protein. A and B) Binding of recombinant (r)CEACAM1 (CC1)-HIS (10 μg/mL) to A) a panel of Gram-positive bacteria, namely *Enterococcus faecalis*, *Enterococcus faecium*, Group A *Streptococcus* (GAS), Group B *Streptococcus* (GBS), Group C *Streptococcus* (GCS), Group G *Streptococcus* (GGS), *Streptococcus pneumoniae*, *Staphylococcus aureus*, B) an expanded collection of GBS isolates, with serotype and carriage of *bac* gene shown. Mean and standard deviation (SD) values are reported for *n* = 3 independent experiments. **C)** Correlation between rCC1-binding capacity and β protein expression in GBS strains, respectively. β protein expression was quantified using rabbit anti-β protein serum and a secondary PE-conjugated goat anti-rabbit-IgG. rCC1 binding to live bacteria was quantified by a secondary anti-His-FITC monoclonal antibody (mAb). Mean binding values from *n* = 3 independent experiments. **D)** Concentration-dependent binding of rCC1-HIS to wild type (WT), Δ*bac* and Δ*bac* + pLZ.*bac* complemented A909 strains. Mean and SD values are reported for *n* = 3 independent experiments. **E)** Binding of human rCC1-HIS (30 μg/mL) to GBS strain A909 strain was analyzed by pull-down experiments followed by Western blot analysis. **F)** Human rCC1-HIS (30 μg/mL) binding to GBS A909 and A909Δ*bac* strains analyzed by pull-down experiments followed by Western blot analysis. **G)** Western blot analysis of rCC1-HIS (30 μg/mL) binding to purified β protein, but not to

control GBS surface proteins α and Rib. Panel on the left show staining of proteins after separation by SDS-PAGE. **H** and **I)** Binding of purified β protein to dynabeads (DB) coated with rCC1 (DB.CC1) or control. Binding of β protein was detected using mouse anti-β serum or normal mouse serum (NMS), and PE-conjugated goat anti-mouse-IgG. H shows mean and SD values compiled from n = 3 independent replicates, and I shows representative flow cytometry plots using 10 μg/ml rCC1.

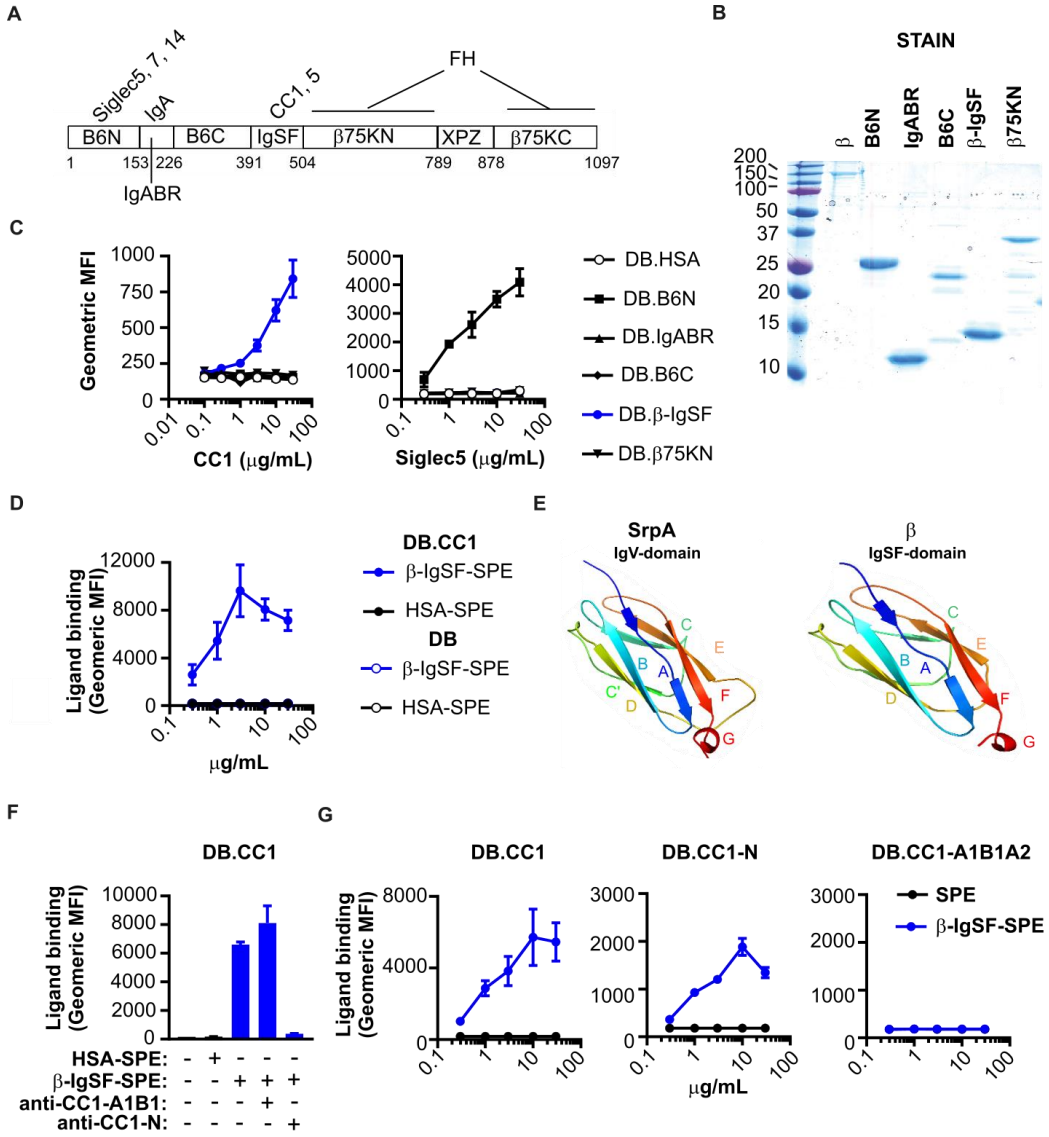

Figure 2: The IgSF domain in the β protein of GBS binds to the N-terminal domain of CEACAMs. **A)** Schematic representation of the β protein.[28,66] The first residue of each domain is indicated. XPZ is a proline-rich repeat region. Known human ligand binding partners are indicated above each respective domain. IgA binds to a region denoted IgABR. Factor H binds the β75K region, but does not bind via XPZ. **B)** Staining of recombinant β protein domains (B6N, IgABR, B6C, IgSF and β75KN) after separation by SDS-PAGE. Recombinant proteins were expressed in *E. coli*. Purified β protein is used as a control. **C)** Binding of rCC1-His or rSiglec5-Fc to dynabeads (DB) coated with β protein domains. Mean and SD values are reported. **D)** Binding of varying concentrations of β-IgSF or human serum albumin (HSA) coupled to streptavidin-PE (SPE), to DB coated with CC1 (DB.CC1) or control protein (DB). Mean and SD values are reported. **E)** Structure of the relevant (Siglec-like sialic acid-binding) domain of the highest-scoring template 5KIQ (SrpA binding region), and predicted three-dimensional model of the IgSF domain from the β protein, constructed from the Hhpred alignment using

MODELLER.[67] **F)** Inhibition of β-IgSF binding to DB.CC1 by anti-CC1-N mAb (clone CC1/3/5-Sab), but not by anti-CC1-A1B1 mAb (clone B3-17). β-IgSF-biotin was coupled to SPE. Mean and SD values are reported. **G)** Concentration-dependent binding of β-IgSF to DB coated with 100 μg/ml CC1 (DB.CC1), CC1-N (DB.CC1-N) or CC1-A1B1A2 (DB.CC1-A1B1A2). Assays utilized proteins coupled to SPE. Mean and SD values are reported.

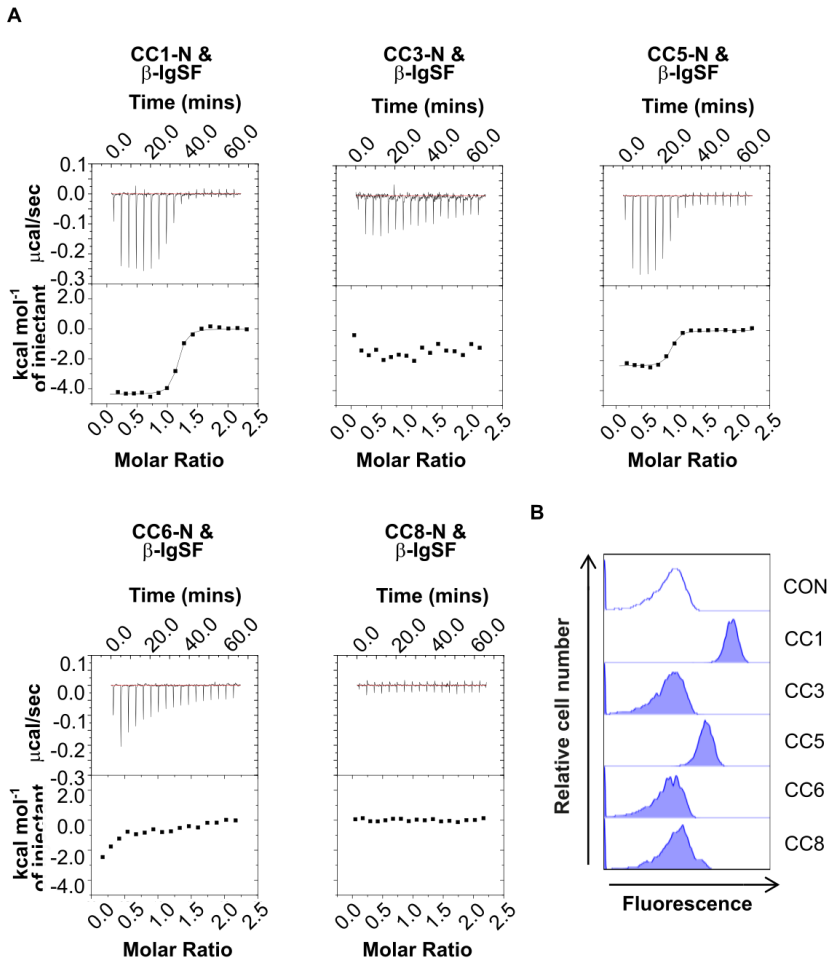

Figure 3. β-IgSF binds with high-affinity to human CEACAM1 and CEACAM5. A) Isothermal Titration Calorimetry (ITC) binding curves of CEACAM-N domains and β-IgSF. Experiments were performed using β-IgSF and N domains of (r)CEACAM1 (CC1), CEACAM3 (CC3), CEACAM5 (CC5), CEACAM6 (CC6) and CEACAM8 (CC8). **B)** Representative flow cytometry plots to analyze interaction of rCC1, CC3, CC5, CC6 or CC8 (10 μg/ml) with GBS A909 from n = 3 independent experiments.

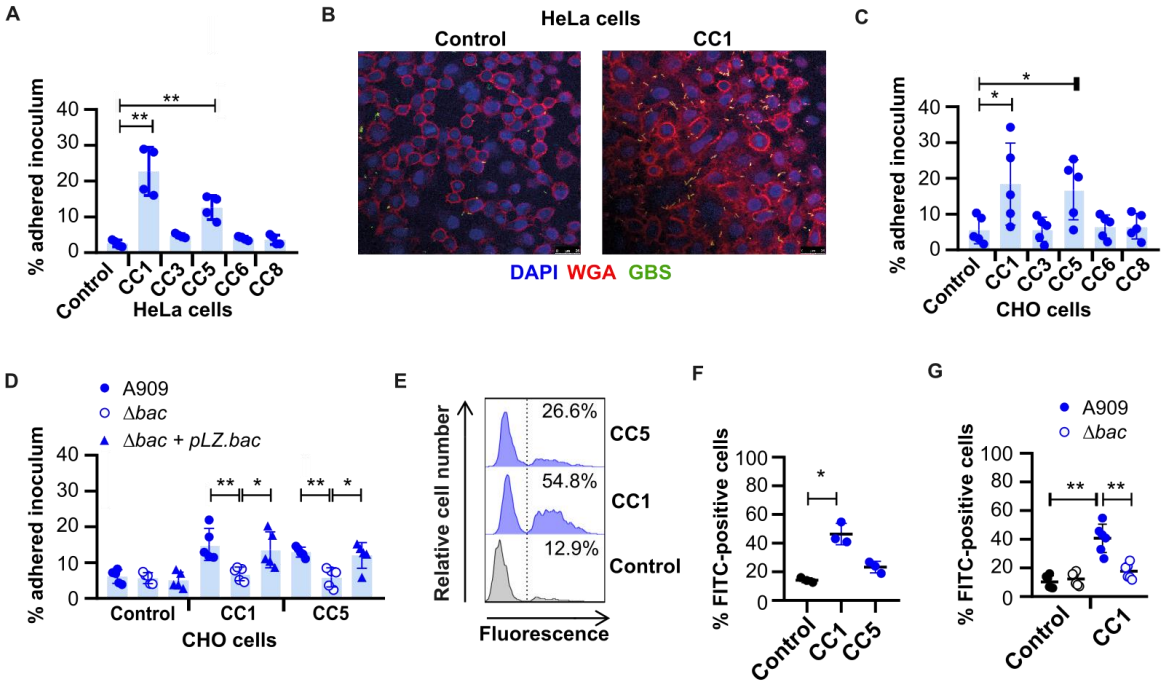

Figure 4: GBS exploit human CEACAM for enhanced adhesion to epithelial cells. A) Adherence of GBS A909 to human CEACAM1 (CC1)-, CEACAM3 (CC3)-, CEACAM5 (CC5)-, CEACAM6 (CC6)- or CEACAM8 (CC8)-expressing, or an empty vector control, HeLa transfectants at MOI of 10 for 30 mins. Mean and SD for *n* = 4 independent replicates are shown. **B)** Confocal imaging of FITC-labelled GBS A909 strain adhesion to human CC1-expressing or empty vector control HeLa transfectants. Cell membranes were stained with AF-647 conjugated wheat germ agglutinin (WGA)(red) and nuclei with DAPI (blue). **C)** Adherence of A909 to human CC1-, CC3-, CC5-, CC6- or CC8-expressing, or an empty vector control, CHO transfectants at MOI of 10 for 30 mins. Mean and SD from *n* = 5 independent replicates is displayed. **D)** Adhesion of WT, Δ*bac* and Δ*bac* + pLZ.*bac* complemented A909 strains to human CC1- or CC5-expressing CHO transfectants, or empty vector control cells. Mean and SD from *n* = 5 independent replicates is displayed. **E and F)** Adhesion of FITC-labelled A909 to human CC1-expressing, CC5-expressing or control CHO transfectants at an MOI of 10. Cell lines were detached and then incubated with FITC-labelled A909 for 30 mins at 4ºC. Fluorescence of cells was measured by flow cytometry analysis. The percentage of fluorescent cells was calculated for each population. Representative flow cytometry plots are shown E, and the integrated results displaying the mean and SD from 3 independent replicates are reported in F. **G)** Adhesion of FITC-labelled A909 strains to human CC1- expressing or control CHO transfectants at MOI of 10. Cell lines were detached and then incubated with FITC-labelled A909 strains for 30 mins at 4ºC. Fluorescence of cells was measured by flow cytometry analysis, where the percentage of fluorescent cells was calculated. Mean and SD for each population for *n* = 6 independent replicates is displayed. Data in A, C, D, F and G was analyzed by repeated One-way ANOVA with Sidak's multiple comparisons. **P*<0.05, ***P*<0.01.

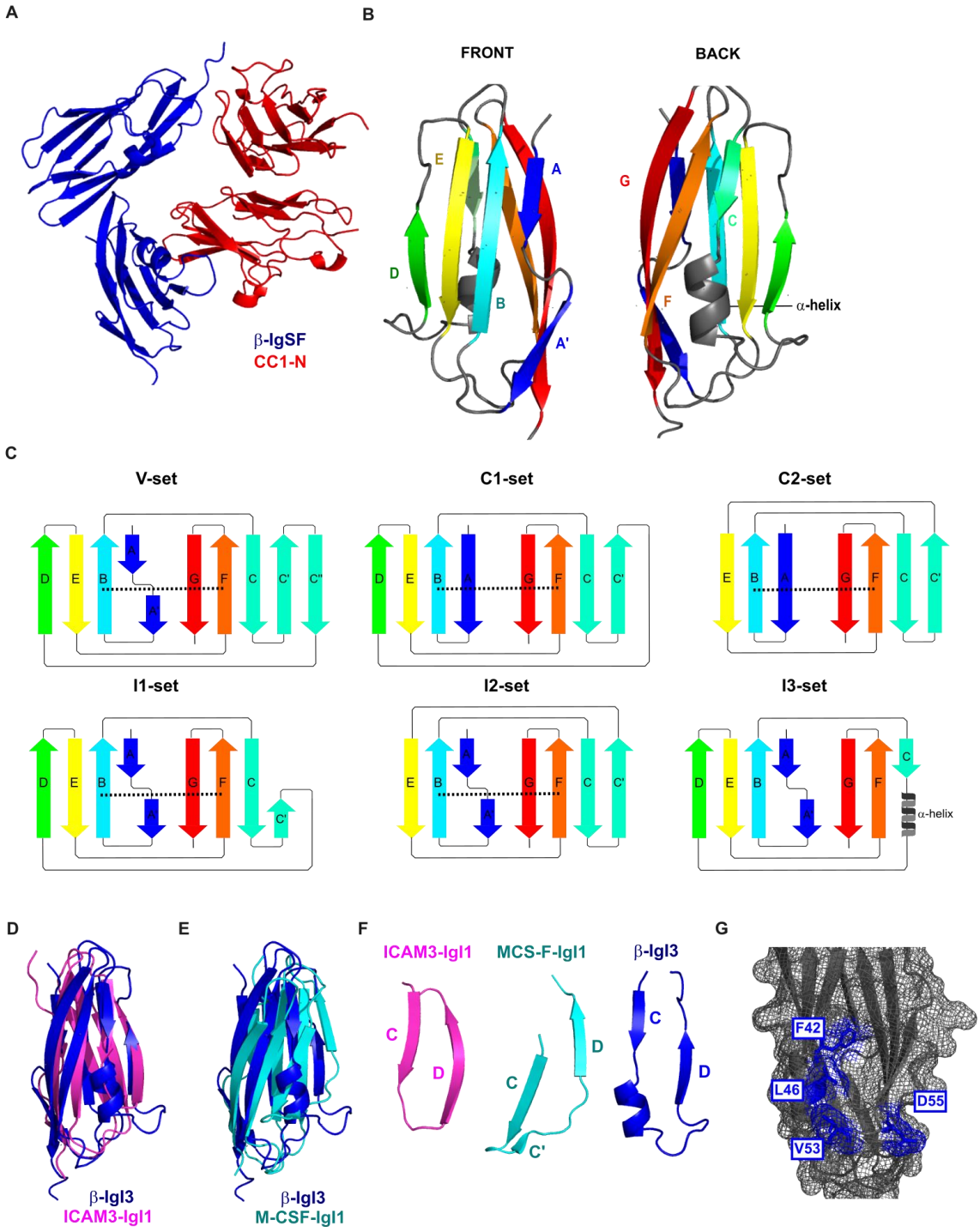

Figure 5: The Ig-like domain of β protein represents a new IgI fold subtype A) Dimer co-crystal structure of β-IgSF (blue) and CEACAM1-N (CC1-N; red). **B)** Molecular structure of β-IgSF, in which each β-strand is coloured differently. The 1.5-turn α-helix shown in grey is located between the *C* strand and the *D* strand. **C)** Illustration of the secondary structure organisation of the V-set, C1-set, C2-set, I1-set, I2-set and I3-set domains. Each secondary structure is composed of varying combinations of the β-strands *A, A′, B, C, C′, C″, D, E, F, G*. Disulfide bridges betwen the *B* strand and *F* strand are represented as a black dashed lines. The C-set domains lack the *A′* and *C″* strands, and are subtyped into C1-set and

C2-set based on presence and absence of the *D* strand, respectively. The I-set domains lack *C″* strand, and are subtyped into C1-set and C2-set based on presence and absence of the *D* strand, respectively. The IgSF fold from β protein has the I3-set domain, characterised by absence of disulfide bridge, presence of the *D* strand, and presence of a truncated *C* strand followed by a 1.5-turn α-helix. **D** and **E)** Superposition of the IgI3 domain from β protein (blue) onto IgI1 domains from ICAM3 (magenta, PDB entry 1T0P; D) or macrophage colony stimulating factor 1 (MCS-F; turquoise, PDB entry 3EJJ; E). **F)** Comparison of β-IgI3 (blue) with the most similar folds; the IgI1 domain of MCS-F (turquoise) and the IgI1 domain 1 of ICAM-3 (magenta,). **G)** The molecular surface mesh of β-IgI3 residues showing protruding hydrophobic residues located between the *C* and *D* strands in blue.

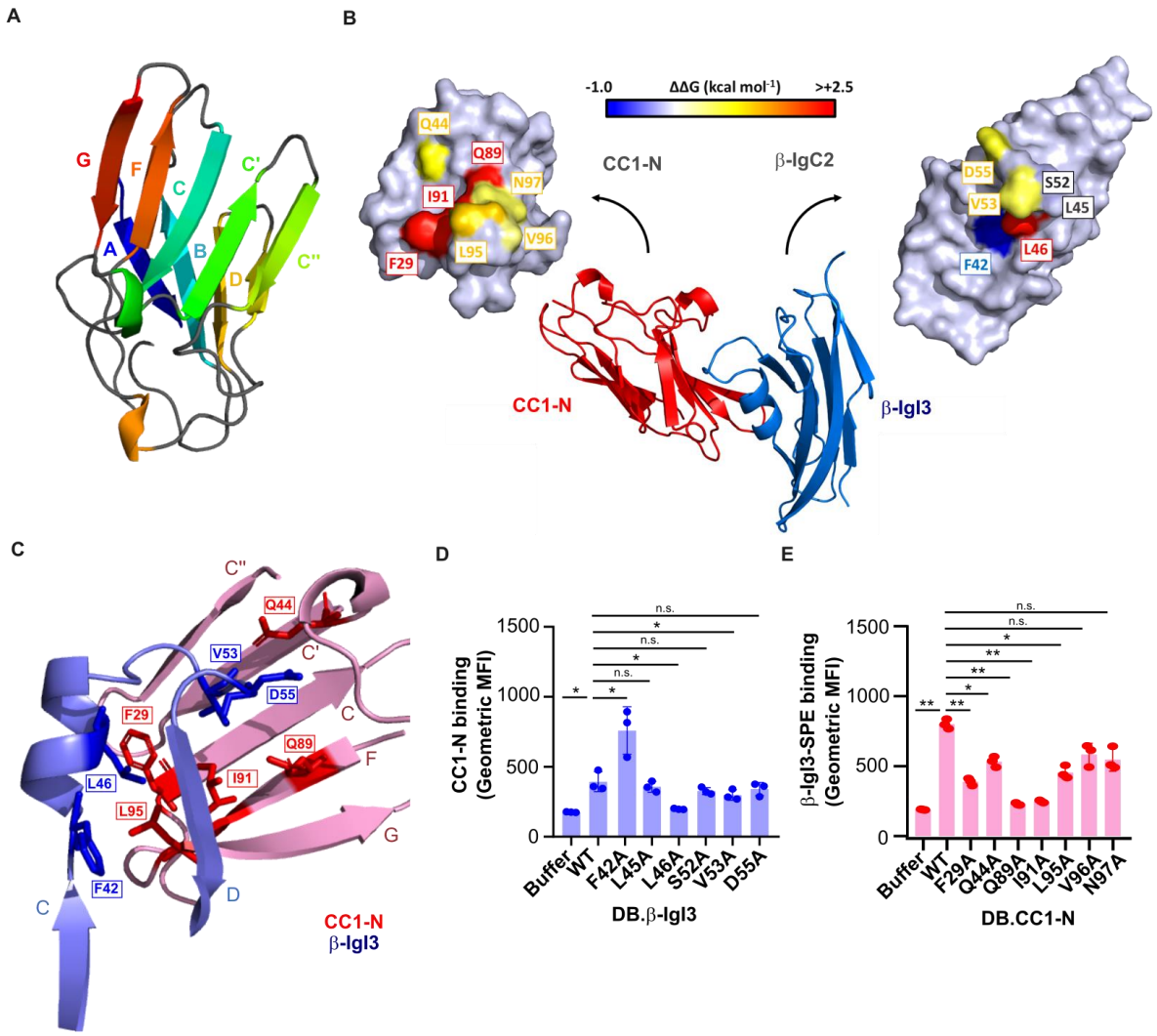

Figure 6. The IgI3 domain of β protein binds CEACAM1 through the unique α-helix and *CD* loop. A) Molecular structure of CEACAM1 (CC1)-N, in which each strand of the IgV structure is coloured differenty. **B)** Molecular structure of the β-IgI3 & CC1 N-terminal domain co-complex. Residues that form the β-IgI3-CC1 interface are highlighted. **C)** Close-up of the β-IgI3 (light blue) interface with CC1-N (pink), including highlighting of residues important for forming β-IgI3 (blue) & CC1-N (red) interactions. The *GFCC′C″* strands of CC1-N are shown and labelled, whilst the *C* to *D* strands of β-IgI3 are shown and labelled. **D)** Binding of rCC1-N to dynabeads (DB) coated with β-IgI3 wild type (WT) or mutants. Mutation of β-IgI3 residues 42, 46 and 53 lead to significant changes in CC1-N binding capacity. **E)** Binding of β-IgI3 coupled to Strepatividin-PE to DB coated with rCC1-N wildtype or variants. Mutation of CC1-N residues 29, 44, 89, 91 and 95 lead to significant changes in β-IgI3 binding capacity. Mean and SD values in D and E are displayed for $n = 3$ independent replicates. Statistical significance was calculated using *t* test for D and E, where *, $p<0.05$; **, $p<0.01$.

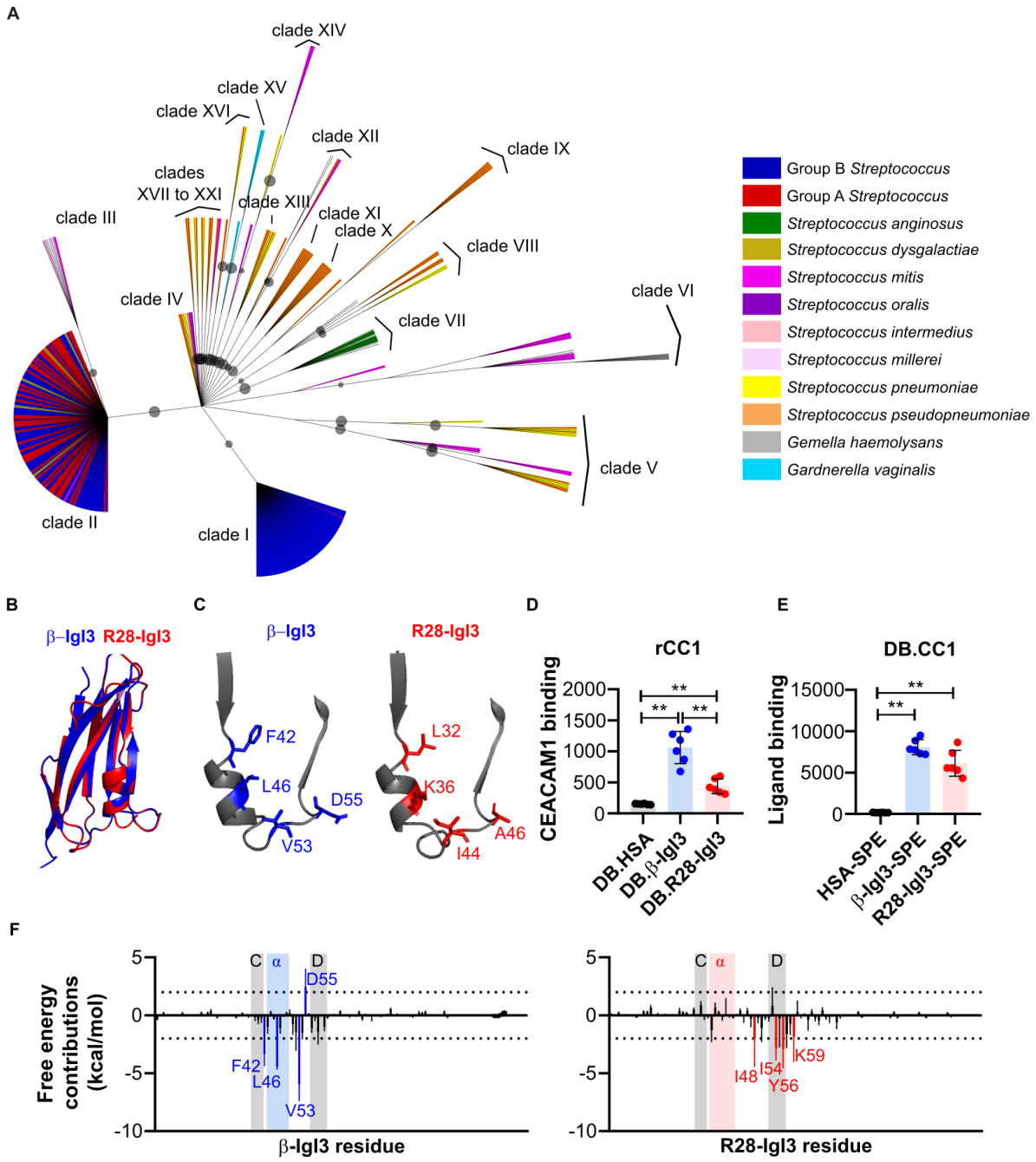

Figure 7: The unique IgI3 structure is found in other Gram-positive bacterial species that bind human CEACAM1. A) Amino acid based Maximum Likelihood tree of β-IgI3 homologs. Clades of related IgI3 domains are shown, with clade I representative of β-IgI3 domain sequences. Branches are coloured by bacterial species from which the sequence originates, except those with no species designated. Maximum likelihood values bootstrap values > 60 are shown as a circle, where larger circles indicate a higher bootstrap support value. **B)** Superposition of the IgI3 domain from β protein (blue) onto predicted IgI3 domain structure of clade II sequence of *S. pyogenes* R28 protein (Red). **C)** Close-up of the truncated *C* strand, α-helix and *D* strand of β-IgI3 (blue) and R28-IgI3 (red), indicating the corresponding R28-IgI3 residues present at sites in β-IgI3 that bind CEACAM1 (CC1). **D)** Binding of

rCC1-HIS to dynabeads (DB) coated with β-IgI3, R28-IgI3 or control. Mean and SD is shown for *n* = 6 independent replicates. **E)** Binding of β-IgI3, R28-IgI3 or human serum albumin (HSA), coupled to PE-conjugated streptavidin (SPE), to DB coated with CC1 (DB.CC1). Mean and SD is shown for *n* = 6 independent replicates. **F)** The per residue free energy contribution spectrums of β-IgI3/CC1-N interface and R28-IgI3/CC1-N interface. The residues with binding free energy contributions lower than -2.0 kcal/mol are identified as key (favourable) residues and coloured blue (for β-IgI3) or red (for R28-IgI3). Free energy contributions at -2.0 kcal/mol and 2.0 kcal/mol are indicated by dashed lines. The position of the truncated *C* strand, α-helix and *D* strand is shaded for each protein. The mean and standard deviation is representative of estimates from 50 docked models. Statistical significance was calculated using a One-way ANOVA with Sidak's multiple comparisons for D and E, where *, $p<0.05$; **, $p<0.01$.