Glycobiology, 2021, vol. 31, no. 9, 1192–1203
https://doi.org/10.1093/glycob/cwab043
Advance Access Publication Date: 13 May 2021
Original Article

OXFORD

Microbial Biology

A minimal sequon sufficient for *O*-linked glycosylation by the versatile oligosaccharyltransferase PglS

Cory J Knoot, Lloyd S Robinson and Christian M Harding[1]

VaxNewMo, 4340 Duncan Ave., Suite 202, St. Louis, MO 63110, USA

[1]To whom correspondence should be addressed: e-mail: christian.harding@vaxnewmo.com

Received 26 March 2021; Revised 29 April 2021; Accepted 7 May 2021

Abstract

Bioconjugate vaccines, consisting of polysaccharides attached to carrier proteins, are enzymatically generated using prokaryotic glycosylation systems in a process termed bioconjugation. Key to bioconjugation are a group of enzymes known as oligosaccharyltransferases (OTases) that transfer polysaccharides to engineered carrier proteins containing conserved amino acid sequences known as sequons. The most recently discovered OTase, PglS, has been shown to have the broadest substrate scope, transferring many different types of bacterial glycans including those with glucose at the reducing end. However, PglS is currently the least understood in terms of the sequon it recognizes. PglS is a pilin-specific *O*-linking OTase that naturally glycosylates a single protein, ComP. In addition to ComP, we previously demonstrated that an engineered carrier protein containing a large fragment of ComP is also glycosylated by PglS. Here we sought to identify the minimal ComP sequon sufficient for PglS glycosylation. We tested >100 different ComP fragments individually fused to *Pseudomonas aeruginosa* exotoxin A (EPA), leading to the identification of an 11-amino acid sequence sufficient for robust glycosylation by PglS. We also demonstrate that the placement of the ComP sequon on the carrier protein is critical for stability and subsequent glycosylation. Moreover, we identify novel sites on the surface of EPA that are amenable to ComP sequon insertion and find that Cross-Reactive Material 197 fused to a ComP fragment is also glycosylated. These results represent a significant expansion of the glycoengineering toolbox as well as our understanding of bacterial *O*-linking sequons.

Key words: bioconjugation, glycoconjugate vaccine, *O*-glycosylation, oligosaccharyltransferase, sequon

Introduction

Protein glycosylation is the most common type of posttranslational modification found in nature. Prokaryotic glycosylation was first reported in *Campylobacter jejuni* a little over two decades ago (Szymanski et al. 1999) and it was functionally transferred into *Escherichia coli* shortly thereafter (Wacker et al. 2002). Prokaryotic protein glycosylation is predominantly either *O*-linking or *N*-linking with *O*-linking systems attaching glycans to the side chains of serine or threonine residues and *N*-linking systems attaching glycans to asparagine side chains (Nothaft and Szymanski 2010; Schaffer and Messner 2017). Both *O*-linking and *N*-linking systems can be further grouped as oligosaccharyltransferase (OTase)-independent or OTase-dependent (Harding and Feldman 2019). OTase-independent glycosylation occurs in the cytoplasm and relies on dedicated glycosyltransferases to glycosylate cognate acceptor proteins. OTase-dependent glycosylation relies on an OTase to transfer a preassembled oligosaccharide en bloc to acceptor proteins in the periplasm. The OTase-dependent protein glycosylation pathway shares many similarities to O-antigen polysaccharide biosynthesis, starting with the transfer of a phosphorylated monosaccharide from a nucleotide-activated precursor to the lipid carrier undecaprenyl phosphate in the inner leaflet of the cytoplasmic membrane (Hug and

Feldman 2011; Valvano 2003). This lipid-linked monosaccharide is sequentially extended by the action of specific glycosyltransferases into a lipid-linked oligosaccharide, flipped to the periplasmic leaflet by a flippase (Raetz and Whitfield 2002) and subsequently transferred to an acceptor protein by an OTase. OTases are promiscuous and will transfer a variety of different glycans (Faridmoayer et al. 2008; Wacker et al. 2006), including long polysaccharides, from various bacterial species to acceptor proteins. This attractive property has led to the exploitation of OTases to transfer bacterial surface polysaccharides, like O-antigens and capsular polysaccharides (CPSs), to specific periplasmic carrier proteins, thereby generating polysaccharide–protein conjugates that are used as conjugate vaccines (Dow et al. 2020; Feldman et al. 2005). This glycoengineering process is termed bioconjugation and, to date, three different OTases named PglB, PglL and PglS have been characterized and used for bioconjugate vaccine development.

PglB is a general *N*-linking OTase from *C. jejuni* and was the first bacterial OTase to be characterized and used in the production of glycoengineered bioconjugates in *E. coli* (Feldman et al. 2005; Szymanski et al. 1999). PglB naturally transfers polysaccharides that have a C2-acetamido sugar at the reducing end (Wacker et al. 2006). While the natural glycan substrate versatility of PglB is the narrowest of all OTases, the *N*-linking sequon, D/E-X_1-N-X_2-S/T (where N is glycosylated and neither X_1 or X_2 are proline) (Kowarik et al. 2006), is the shortest. The *N*-linking sequon of bacteria is similar to that recognized by Stt3, the catalytic subunit of the eukaryotic *N*-linking OTase complex (Kowarik et al. 2006). PglL (also known as PglO) is a general *O*-linking OTase first characterized from *Neisseria* that transfers glycans with either a C2-acetamido sugar or galactose at the reducing end (Faridmoayer et al. 2007). In contrast to PglB, there is no known conserved sequon for PglL, with glycosylation preferentially occurring in regions of low complexity that are rich in smaller amino acids such as alanine, serine and proline (Vik et al. 2009). Recently, an optimized PglL sequon, WPAAASAP (where S is glycosylated), was derived from PilE, one of the natural pilin substrates for PglL (Pan et al. 2016); however, the hydrophilic amino acid sequences DPRNVGGDLD and QPGKPPR were required to flank the optimized sequon in order for PglL to efficiently glycosylate proteins containing this tag. PglS, previously referred to $PglL_{ComP}$, is an *O*-linking OTase that specifically glycosylates only one protein, ComP, a bacterial pilin protein of *Acinetobacter* species (Harding et al. 2015). Importantly, PglS is the only known OTase capable of naturally transferring glycans with glucose at the reducing end in addition to glycans containing either galactose or a C2-acetamido sugar at the reducing end (Harding et al. 2019).

In the last decade, PglB, and to a lesser extent PglL, have been used to develop bioconjugate vaccines against *Staphylococcus aureus*, *Shigella dysenteriae* and *flexneri*, extraintestinal pathogenic *E. coli*, *Salmonella* species and others (Hatz et al. 2015; Huttner et al. 2017; Reglinski et al. 2018; Sun et al. 2018; van den Dobbelsteen et al. 2016; Wacker et al. 2014). However, the inability of PglB and PglL to naturally transfer polysaccharides with glucose at the reducing end hinders these enzymes from being used to make bioconjugate vaccines against several prominent bacterial threats (Harding et al. 2019). For instance, ~75% of *Streptococcus pneumoniae* (pneumococcus) capsules, >50% of *Klebsiella pneumoniae* capsules and all 10 *Streptococcus agalactiae* group B (GBS) capsules contain glucose as their reducing end sugar (Berti et al. 2014; Geno et al. 2015; Pan et al. 2015). The natural ability of PglS to transfer polysaccharides with glucose at the reducing end therefore lends itself well to the development of broad pneumococcal, *K. pneumoniae* and GBS vaccines that target the CPSs of these pathogens. Indeed, using PglS with native ComP or an engineered ComP fusion as acceptor proteins, we recently reported the production of bioconjugate vaccines against current nonvaccine serotypes of pneumococcus (Harding et al. 2019) as well as hypervirulent *K. pneumoniae* (Feldman et al. 2019).

Although the polysaccharide substrate versatility of PglS makes it an attractive OTase for the production of next-generation bioconjugate vaccines, the minimal ComP sequon sufficient for PglS-dependent glycosylation has not yet been identified. Previous bioconjugate vaccines developed using PglS have relied on using either the full-length native ComP protein, which is membrane-associated, as the carrier protein or an N-terminally truncated 117 amino acid ComP variant that was translationally fused at the C-terminus of the exotoxin A (EPA) from *Pseudomonas aeruginosa*. A shorter, more modular ComP sequon that is efficiently glycosylated by PglS would be preferable since the previous iterations containing the 117 amino acid ComP fragment are glycosylated only when they are translationally fused to the C-terminus of the carrier protein, thereby limiting applications. With PglB, for instance, knowledge of the short *N*-linking sequon has allowed multiple glycosylation sites to be engineered into the surface of carrier proteins, resulting in singly and multiglycosylated bioconjugates (Ihssen et al. 2010). It has also enabled more sophisticated in vitro studies involving different PglB peptide substrate variants and their effects on peptide binding and catalysis (Gerber et al. 2013). As such, identification of a short ComP sequon could provide insights into the structural determinants of acceptor protein specificity in the PglS OTase family, facilitate comparisons to the sequons recognized by other *O*-linking OTases like PglL and help guide improvements in glycoengineering design.

In this study, we report on the identification of short, modular ComP fragments containing the *O*-linking site that are able to be efficiently glycosylated by PglS in vivo. We tested over 100 fragments of the *Acinetobacter soli* strain CIP 110264 ComP protein translationally fused to EPA to identify a 25 amino acid region termed $ComP_{C1}$ that is sufficient for glycosylation when added to the N- or C-terminus of a carrier protein. In addition, we show that the $ComP_{C1}$ fragment can be modularly fused to a genetically inactivated diphtheria toxin, a carrier protein that has not been previously used with PglS. Next, we further reduced the size of the ComP sequon by sequentially truncating a $ComP_{C1}$ fragment introduced internally in the EPA amino acid sequence on a solvent-exposed surface loop. This led to the identification of an 11 amino acid ComP sequon that is sufficient for robust glycosylation by PglS in *E. coli*. We also find evidence that sequons as short as five to six amino acids allow for low-level glycosylation by PglS. The short size and modularity of the 11 amino acid ComP sequon will allow for the engineering of multiglycosylated carrier proteins that were not accessible with the longer ComP fragments used in the initial generation of PglS-derived bioconjugates. Importantly, it also advances our understanding of the sequons recognized by diverse *O*-linking OTases.

Results

Identification of a putative disulfide-flanked region of $ComP_{110264}$ sufficient for PglS-dependent glycosylation

Previously, we demonstrated that the PglS ortholog from *Acinetobacter baylyi* strain ADP1 glycosylates the ComP ortholog from *A. soli* strain CIP 110264 at a single serine residue located at position 82 (Harding et al. 2019). Using this knowledge, we engineered PglS to functionally glycosylate heterologous proteins by

translationally fusing a large fragment (117 amino acids) of ComP to the C-terminus of a known carrier protein. Specifically, the 117 amino acid $ComP_{110264}$ fragment was fused at the C-terminus of a genetically deactivated EPA between a flexible GGGS linker. This chimeric carrier protein also had an N-terminal DsbA signal sequence (ssDsbA) for translocation to the periplasm via the Sec-pathway as well as a C-terminal hexahistidine tag for detection. While efficiently glycosylated by PglS, we sought to significantly shorten the size of the ComP fragment required for PglS-dependent glycosylation.

To identify a shorter ComP sequon required for PglS-mediated glycosylation, we cloned 87 different truncated $ComP_{110264}$ fragments onto the C-terminus of EPA and screened each in a 96-well plate format for protein glycosylation. $ComP_{110264}$ fragments were designed to shift one amino acid N- to C-terminal relative to serine 82, which is the site of PglS glycosylation (Figure 1A and B). Since we did not know if the size of the sequon would be important for PglS glycosylation, we generated fragments of various lengths of 25, 30, 35, 40 or 45 amino acids from ComP. The fragments were PCR amplified, cloned onto the C-terminus of EPA and tested for bioconjugation by PglS. For these experiments we used the serotype 8 pneumococcal capsular polysaccharide (CPS8) expressed from the pB-8 plasmid as the glycan source (Kay et al. 2016). The CPS8 glycan is composed of repeating units of the linear tetrasaccharide [→4)-α-D-Glc*p*-(1 → 4)-α-D-Gal*p*-(1 → 4)-β-D-Glc*p*A-(1 → 4)-β-D-Glc*p*-(1→] with glucose as the reducing end sugar (Bentley et al. 2006). Previous studies demonstrated that PglS efficiently transfers CPS8 glycans containing at least seven repeat units to a fragment of ComP; however, bioconjugates containing >15 CPS8 repeat units were also observed (Harding et al. 2019). Bioconjugation was performed in the glycocompetent *E. coli* strain SDB1. SDB1 has deletions of *wecA*, which encodes the glycosyltransferase that initiates biosynthesis of the enterobacterial common antigen and the O-antigen polysaccharides, and *waaL*, encoding the ligase that transfers undecaprenyl-pyrophosphate-linked glycan precursors to the outer core of lipid-A (Garcia-Quintanilla et al. 2014). Collectively, these mutations facilitate the accumulation of heterologously expressed lipid-linked glycan precursors, like the CPS8 polysaccharide lipid-linked precursor, for exclusive use by PglS. SDB1 strains expressing the CPS8 glycan, PglS and a chimeric EPA–ComP construct from IPTG-inducible vectors were cultured in LB, induced at mid-log growth phase and grown overnight. Samples were harvested ~20 h after induction for western blot analysis on whole-cell or periplasmic extracts to assess EPA–ComP fusion protein expression and protein glycosylation. Western blots were probed using antibodies against EPA (anti-EPA) and the hexahistidine tag (anti-His). Probing with both antibodies allowed us to ascertain if the EPA protein and/or the C-terminal ComP tag remained intact. For this experiment, we did not probe with the anti-CPS8 antisera, as trace contamination from lipid-linked CPS8 polysaccharide precursors in the whole-cell or periplasmic extractions also react with anti-glycan antisera. This can confound western blot interpretation, making it difficult to discern protein-linked from lipid-linked polysaccharide.

The complete western blot results for all 87 initial EPA–ComP variants are shown in Supplementary Figures S1–S8 with a selected subset shown in Figure 1C–E. Of the 87 EPA–ComP variants tested, we observed glycosylation in 37 different fusions (Supplementary Figures S1–S8). The size of the ComP fragment did not appear to influence PglS-dependent glycosylation as glycosylation was observed for EPA constructs containing all five different-sized ComP fragments tested (25, 30, 35, 40 and 45 amino acids). While size did not appear to be important for glycosylation, the presence of the Cys71 and Cys93 residues flanking Ser82 in $ComP_{110264}$ were found to be essential for these chimeric EPA–ComP proteins that contain the ComP fragment at the C-terminus. As seen in Figure 1C–E, ComP fragments lacking either Cys71 or Cys93 were not glycosylated. Only in those with both cysteine residues did we observe transfer of the CPS8 glycan. In addition, of the shortest nine fragments tested (fragments A1–H1 and A2), only three (fragments C1, D1 and E1) were glycosylated by PglS; all of which contain both Cys71 and Cys93 (Supplementary Figure S1). Of these 25 amino acid fragments, only the "C1" ComP fragment (EPA–$ComP_{C1}$) is shown in Figure 1C–E. The glycosylation efficiency (defined here as the observed CPS8-glycosylated EPA compared with the unglycosylated EPA) and average number of the CPS8 repeat units transferred by PglS were similar for all the predicted disulfide-containing fragments, although we did observe some experiment-to-experiment variability. Upon closer examination of the western blots, we observed that chimeric EPA–ComP variants (C2, D2, E3 and F3) predominantly migrated with a slightly lower molecular weight when compared with the unglycosylated EPA–ComP variants containing both Cys71 and Cys93 (Figure 1C). Moreover, the lower molecular weight chimeric EPA–ComP variants (C2, D2, E3 and F3) did not react with the anti-His antibody when compared with the anti-EPA signal (Figure 1D). Taken together, these observations indicate that the ComP fragments lacking both cysteine residues are unstable and likely prone to C-terminal degradation, thereby preventing glycosylation by PglS. We hypothesize that Cys71 and Cys93 are able to stabilize $ComP_{110264}$ by forming a covalent disulfide bridge.

Bioconjugation of CRM_{197} at the N- or C-terminus with a modular $ComP_{C1}$ tag

A variety of proteins from different organisms, typically inactivated bacterial toxins, have been used as carriers for conjugate and bioconjugate vaccines. Cross-Reactive Material 197 (CRM_{197}) is a genetically deactivated form of the diphtheria toxin that has been used extensively as the carrier protein in multiple conjugate vaccines for pneumococcus, *Neisseria meningitidis* and *Haemophilus influenza* type b (Berti and Adamo 2018). Using the knowledge gained above and the frequent use of CRM_{197} in conjugate vaccine formulations, we aimed to extend the PglS bioconjugation system to function with CRM_{197}. For these experiments, we translationally fused the 25-amino acid "C1" ComP fragment ($ComP_{C1}$) identified above to the C-terminus of CRM_{197} linked by a GGGS sequence (Figure 2A, top diagram). We chose $ComP_{C1}$ because it was among the smallest identified in the original screen and was glycosylated at similar levels as the other fragments. An SRP-dependent FlgI secretion sequence (ssFlgI) was added to the N-terminus for CRM_{197} for export to the periplasm (Goffin et al. 2017). Finally, a C-terminal hexahistidine tag was added to aid purification. *Escherichia coli* SDB1 cells expressing the CPS8 glycan along with PglS and the CRM_{197}–$ComP_{C1}$ carrier (expected size of 61.8 kDa) were cultured in shake flasks and harvested after 20 h of IPTG induction. The CRM_{197}–$ComP_{C1}$–CPS8 bioconjugate was purified with three successive rounds of chromatography. First, we employed Ni-affinity chromatography as the bioconjugates contain a C-terminal hexahistidine tag. Fractions containing bioconjugates were pooled and enriched for glycosylated bioconjugates using a MonoQ column and eluted with a linear salt gradient. A final polishing step to remove large aggregates and unglycosylated CRM_{197} was performed on a Superdex 200 Increase column resulting in approximately 70 μg of total glycoprotein

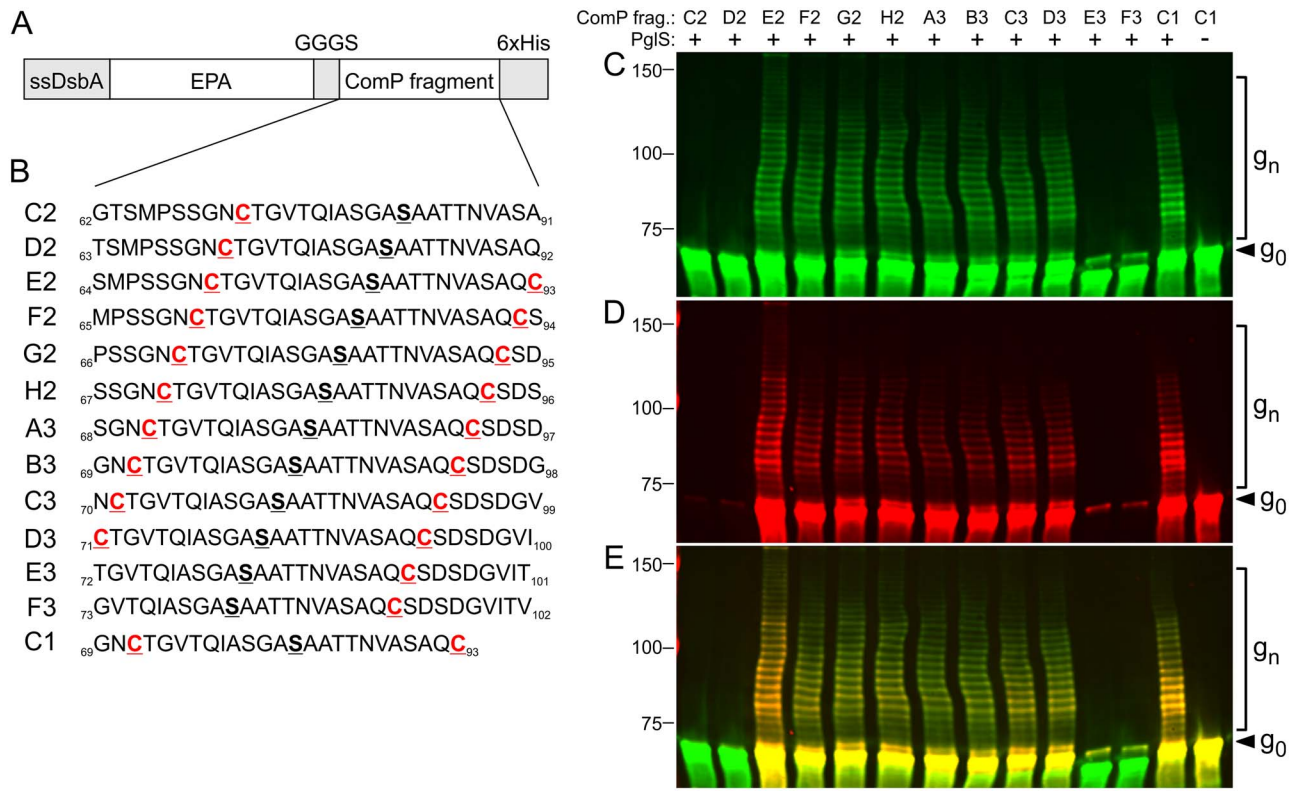

Fig. 1. Design strategy and bioconjugation experiments used to identify $ComP_{110264}$ fragments capable of being glycosylated by PglS. (**A**) Schematic of EPA–$ComP_{110264}$ fusion proteins used in bioconjugation experiments. Domains and tags are not drawn to scale. "ssDsbA" corresponds to the DsbA Sec secretion signal. GGGS is a flexible linker between EPA and the $ComP_{110264}$ fragment. (**B**) Subset of the 87 $ComP_{110264}$ fragments tested for bioconjugation. The bold, underlined serine is the site of glycosylation. Cysteine residues found to be important for stability and glycosylation of $ComP_{110264}$ in these experiments are shown in red. (**C**) Western blot of periplasmic extracts from *E. coli* SDB1 expressing PglS, the CPS8 glycan and an EPA–ComP variant. Anti-EPA channel. Each lane contains a different EPA–ComP variant with the ComP fragment corresponding to the sequence shown in panel **B**. (**D**) Anti-His channel. (**E**) Merge of **C** and **D**. Equivalent amounts of periplasmic extract based on OD_{600} were loaded per lane. To the right of panels **C**–**E**, g_0 marks unglycosylated EPA–ComP and g_n marks protein glycosylated with different numbers of CPS8 repeat units. Protein mass markers (in kDa) are indicated to the left of panels **C**–**E**.

per liter LB in shake flasks. As seen in Figure 2B–D, western blotting on the purified samples using anti-CRM_{197} and pneumococcal CPS8 antisera demonstrated that the CRM_{197}–$ComP_{C1}$ carrier was glycosylated with CPS8. Digestion of the purified bioconjugates with Proteinase K prior to separation on sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) resulted in a complete loss of the CRM_{197}- and polysaccharide-specific signals, indicating that the CPS8 glycans were covalently attached to CRM_{197}–$ComP_{C1}$ protein.

Next, we tested whether the $ComP_{C1}$ glycotag could be moved to another site in the CRM_{197} fusion. As such, we designed a new construct placing $ComP_{C1}$ N-terminal to the CRM_{197} coding region (Figure 2A, bottom diagram). The FlgI secretion signal was placed immediately N-terminal to $ComP_{C1}$ and CRM_{197} was C-terminally tagged with hexahistidine. *Escherichia coli* SDB1 cells expressing the CPS8 glycan along with PglS and the $ComP_{C1}$–CRM_{197} carrier were cultured in shake flasks and harvested after 24 h. As seen in Figure 2E, western blot analysis of periplasmic extracts probing with an anti-His antibody showed that the $ComP_{C1}$–CRM_{197} was also glycosylated by PglS. The average number of CPS8 repeat units and glycosylation efficiency of both fusions was comparable. These experiments demonstrate that the $ComP_{C1}$ glycotag can be used with multiple carrier proteins and can be placed at the N- or C-terminus of a fusion construct.

Identification of an 11 amino acid $ComP_{110264}$ sequon sufficient for PglS glycosylation

While the preceding experiments show that Cys71 and Cys93 are required for the stability of $ComP_{110264}$ fragments when translationally fused at the C-terminus of EPA (Figure 1C–E), these data do not ascertain whether the two cysteine residues and the putative disulfide bridge formed between them are necessary for glycosylation by PglS. The *N*-linking sequon recognized by PglB has been engineered into multiple sites on surface loops of EPA, essentially being used as an "internal" glycotag (Ihssen et al. 2010). Therefore, in order to determine whether Cys71 and Cys93 of ComP were actually required for glycosylation, we chose a site in the EPA coding sequence to insert the entire 23 amino acid fragment of $ComP_{110264}$ spanning Cys71 to Cys93, which we refer to as the iGT_{CC} for **i**nternal **G**lyco**T**ag–**c**ysteine–**c**ysteine. Using the crystal structure of EPA to guide our design (PDB ID 1IKQ) (Wedekind et al. 2001), we inserted the $ComP_{110264}$ iGT_{CC} between residues Ala489 and Arg490 of EPA, which is in a β-turn structure on the surface of the catalytic domain (Figure 3A). As a control, we also made a variant of the iGT with the two cysteine residues replaced by serine residues, termed iGTss ("serine–serine"). Serine residues are hypothesized to contribute a similar steric bulk as the cysteine residues but are unable to oxidize and form a disulfide bond. Diagrams of these constructs and the sequences of the iGTs are shown in Figure 3B. The ability of PglS

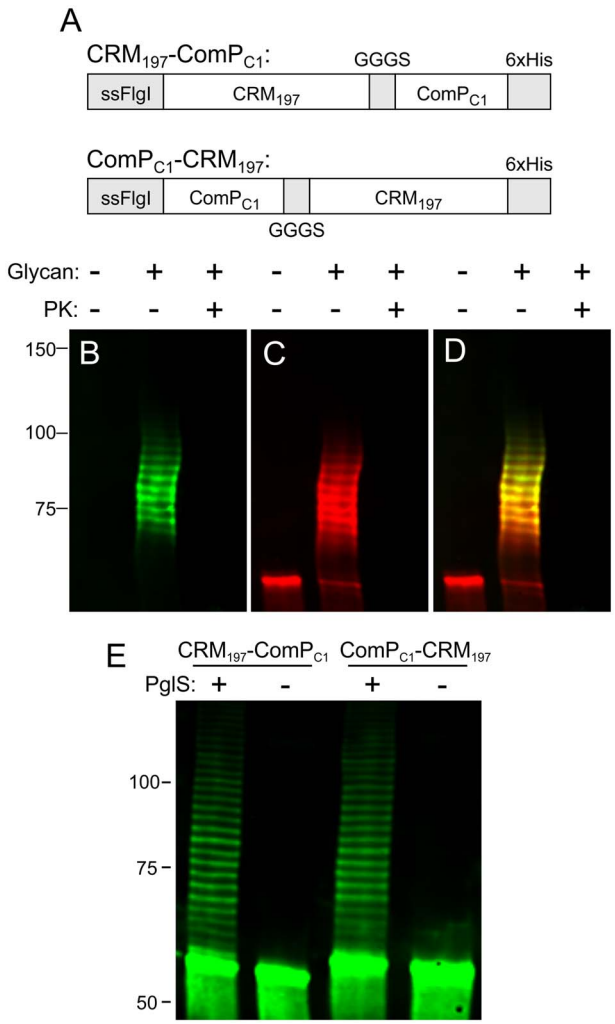

Fig. 2. Western blot of CRM_{197}–$ComP_{C1}$ bioconjugates with pneumococcal CPS8. (**A**) Diagrams of the CRM_{197}–$ComP_{C1}$ and $ComP_{C1}$–CRM_{197} fusion protein constructs. Domains and tags are not drawn to scale. "ssFlgI" corresponds to the FlgI SRP secretion signal. GGGS is a flexible linker between CRM_{197} and $ComP_{C1}$. (**B**) Purified CRM_{197}–$ComP_{C1}$–CPS8, anti-CPS8 channel. (**C**) Anti-CRM_{197} channel. (**D**) Merge of **B** and **C**. Loss of signal in the proteinase K (PK)-treated samples shows that the pneumococcal serotype 8 signal is CRM_{197}-linked and not the result of contamination from free polysaccharide or lipid-linked polysaccharide precursors. (**E**) Western blots showing the anti-His channel of periplasmic extracts of *E. coli* SDB1 expressing CRM_{197}–$ComP_{C1}$ or $ComP_{C1}$–CRM_{197} and the CPS8 glycan in the presence (+) or absence (−) of PglS. Equivalent amounts of periplasmic extract based on OD_{600} were loaded per lane. The $ComP_{C1}$ glycotag is modular and can be glycosylated at either the N- or C-terminus of CRM_{197} at similar efficiencies. Protein mass markers (in kDa) are indicated to the left of panels **B**–**E**.

to transfer CPS8 to the EPA_{iGTcc} or EPA_{iGTss} was assessed in a three-plasmid system as described above. As seen in Figure 3C–E, both the cysteine–cysteine and serine–serine variants of EPA_{iGT} were glycosylated, demonstrating that Cys71 and Cys93 (and the putative disulfide bond formed between them) are not required for glycosylation of the iGTs.

Since the cysteine residues were not necessary for glycosylation, we hypothesized that a shorter motif representing the minimal *O*-linking ComP sequon could be found within the 23-amino acid sequence spanning from Cys71 to Cys93. To investigate this, we

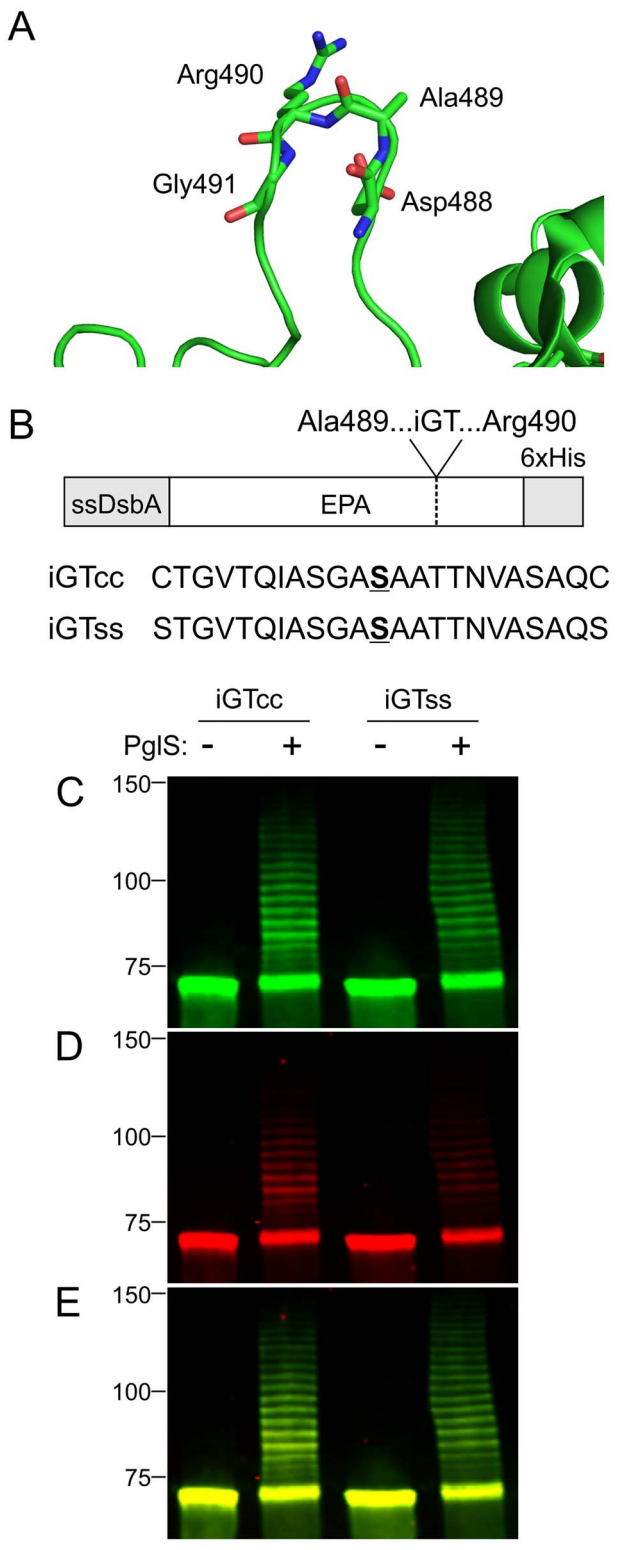

Fig. 3. Western blot analysis of EPA variants containing iGTs derived from $ComP_{110264}$. (**A**) Structure of the surface of EPA (PDB 1IKQ) showing the hairpin loop where the iGTs were inserted. Residues in the loop are labeled. Carbon atoms are shown in green, oxygen in red and nitrogen in blue. The EPA protein backbone is shown as a green strand. (**B**) Diagrams of the internally tagged EPA construct used for these experiments. The protein sequences of the two iGT variants inserted between EPA residues Ala489 and Arg489 are shown below the diagrams. These have either two terminal cysteines ("iGTcc")

generated shorter variants of the iGT_{CC} translationally inserted between EPA residues Ala489–Arg490 in order to identify which ComP residues were necessary for glycosylation. We alternated deleting a single amino acid from either side of the 23-amino acid iGT_{CC}, generating 22 truncated variants that each contained Ser82, the site of PglS glycosylation (Figure 4A and B). These variants were named after the number of deleted residues from either side of the iGT_{CC}, e.g. Δ3-4 corresponds to a deletion of three amino acids from the N-terminal side of iGT_{CC} and a four-amino acid deletion from the C-terminal side. The shortest variant generated was five amino acids long. These truncated EPA–iGT_{CC} variants were tested for bioconjugation with CPS8 and PglS in shake flasks under the same conditions as the preceding experiments. As a negative control, we included a construct expressing only the EPA coding sequence along with DsbA secretion and hexahistidine tags.

As seen in Figure 4C, we observed robust glycosylation for all constructs that were at least 11 amino acids in length. Their glycosylation efficiency was comparable to the 23-amino acid iGT_{CC} tag, suggesting modest truncations on either side of Ser82 do not have a significant impact on the glycosylation efficiency of PglS. The shortest internal ComP fragment that was efficiently glycosylated was iGTΔ6-6 having the sequence IASGASAATTN (Figure 4C). Removal of either the N-terminal isoleucine residue (iGTΔ7-6) or C-terminal asparagine residue (iGTΔ6-7) dramatically reduced the glycosylation efficiency of the carrier protein. Variants smaller than iGTΔ6-6 mostly showed minimal glycosylation, the best of these being iGTΔ7-6 with sequence ASGASAATTN. Interestingly, a small amount of higher molecular weight laddering was also observed in the smallest ComP fragments, iGTΔ9-8 and iGTΔ9-9 (Supplementary Figure S9), suggesting that these six and five amino acid variants, respectively, were glycosylated by PglS at low levels. Since the iGTΔ6-6 variant was the smallest version that worked effectively as a glycotag, we wanted to validate that iGTΔ6-6 was being glycosylated with CPS8. To this end, we purified EPA iGTΔ6-6 protein from whole-cell lysates using a Ni-affinity chromatography and performed western blotting on the eluate using antisera specific to either protein or the CPS8 glycan. The results of these experiments clearly show that the iGTΔ6-6 was being glycosylated with CPS8 by PglS (Figure 4D–F). We conclude that $ComP_{110264}$ can be shortened from 117 amino acids to an internal sequon as short as 11 amino acids while maintaining robust glycosylation efficiency.

The preceding iGT truncation series was only tested at one internal site on EPA between residues Ala489 and Arg490. Last, we wanted to determine whether one of the smaller iGT variants could be moved to a second site on EPA and still be efficiently glycosylated by PglS. As such, we chose a second site between EPA residues Glu548 and Gly549 and incorporated the iGTΔ3-4 amino acid sequence (Supplementary Figure S10). Like the first site, the second is found on a surface-exposed hairpin loop in the catalytic domain of EPA. To our knowledge, these two sites have not been used before for sequon insertion and bioconjugation. We tested this alternately tagged variant for bioconjugation with CPS8 and PglS under the same conditions as the other truncations. This construct was also glycosylated with CPS8 (Supplementary Figure S10), showing that both sites on EPA are viable for insertion of the ComP sequon.

Identification of a conserved cysteine-flanked region in ComP orthologs

Having identified the minimal $ComP_{110264}$ sequence sufficient for glycosylation by PglS, we asked whether we could identify similar sequences in ComP orthologs. We were particularly interested in identifying cysteine-flanked regions in ComP orthologs and whether there were conserved motifs or trends in the intervening sequences. Using $ComP_{110264}$ as a template, we searched the genomes of other bacteria in the *Moraxellaceae* family using BlastP and assigned hits as ComP orthologs using two primary criteria. First, we checked for the presence of a *pglS* gene immediately downstream of the candidate *comP*, which is the typical genetic architecture of pilin-OTase pairs in *Acinetobacter*. Second, *Acinetobacter* species that encode a *comP*–*pglS* pair also encode a *pglL* gene elsewhere in the genome (Harding et al. 2015). In contrast, *Acinetobacter* species without a *pglS* or *comP* tend to encode a general *pglL* immediately downstream of a *pilA*–*tfpO* pair (Harding et al. 2015). We thus only assigned genes as *comP* in strains that carry both a *pglS* gene next to the putative *comP* and a *pglL* elsewhere in the chromosome.

Using these criteria, we identified six putative ComP orthologs besides $ComP_{ADP1}$, which has been previously characterized (Table I). These other ComP orthologs are found in *Acinetobacter* and *Psychrobacter* species. All of the ComP orthologs are of similar size, between 140 and 150 amino acids. Based on a multiple sequence alignment of these proteins, each contains a cysteine-flanked region similar to that in $ComP_{110264}$ (Figure 5). This conserved cysteine-flanked motif is not found in the *Acinetobacter nosocomialis* M2 PilA, suggesting that it is unique to ComP orthologs (Supplementary Figure S11). The cysteine residues corresponding to Cys71 and Cys93 from $ComP_{110264}$ were found in similar locations within all ComP orthologs, with the size of the cysteine-flanked region ranging from 19 to 23 amino acids in length. Generally, the sequences flanking the glycosylated serine (Ser82 of $ComP_{110264}$) contain smaller amino acids, primarily alanine, glycine, asparagine, aspartate and serine/threonine. This is consistent with the observation that *O*-linking OTases tend to target glycosylation regions containing amino acids of lower complexity, namely alanine, serine and proline (Vik et al. 2009). Unlike PglL, however, proline residues near the glycosylation site are absent in the ComP sequon (Pan et al. 2016). Intriguingly, some homologues have lysine residues +2 and/or +4 relative to the serine glycosylation site, a drastic departure from the much smaller residues seen in the other orthologs including $ComP_{110264}$. Our above truncation series of the $ComP_{110264}$ cysteine region showed that the 11-mer iGTΔ6-6 was the smallest in the tested series that retained efficient levels of glycosylation. The iGTΔ6-6 sequence is framed by an N-terminal isoleucine and C-terminal asparagine. Interestingly, all of the ComP orthologues have a corresponding conserved asparagine residue +5 relative to the serine glycosylation site, and most have an isoleucine residue a few residues upstream of the predicted glycosylated serine. The conserved identity or chemical character of these residues, paired with our truncation series experiments, suggests that these residues may be important for PglS-mediated glycosylation in all ComP proteins.

or serines ("iGTss"). (**C**) Western blots on periplasmic extracts of *E. coli* SDB1 expressing the CPS8 glycan, EPA_{iGTcc} or EPA_{iGTss}, with (+) or without (−) PglS. Anti-EPA channel. (**D**) Anti-His channel. (**E**) Merge of **C** and **D**. Equivalent amounts of periplasmic extracts based on OD_{600} were loaded per lane. Protein mass markers (in kDa) are indicated to the left of panels. The image in panel A was generated using Pymol (https://pymol.org/2/).

Table I. ComP orthologs identified in *Acinetobacter* and *Psychrobacter* strains

Organism	NCBI Accession	Protein Size (AA)	Disulfide Size (AA)
Acinetobacter radioresistens 50v1	WP_101235146.1	142	21
Acinetobacter sp. SFC	WP_067726803	150	21
Acinetobacter soli GFJ2	WP_076033120	146	21
Psychrobacter sp. Marseille-P5312	WP_105244723	140	23
Psychrobacter sp. ANT_H59	WP_149408907	142	21
Acinetobacter parvus DSM 16617	ENU35291	150	19
Acinetobacter baylyi ADP1	WP_004923779	147	21
Acinetobacter soli CIP 110264	WP_004934263	145	23

Discussion

The exploitation of OTases to generate bioconjugate vaccines offers several advantages over conventional methods for producing conjugate vaccines. Namely, bioconjugation is considered a great simplification of the production process, requiring less complex and time-consuming manufacturing and purification processes (Frasch 2009; Rappuoli et al. 2019). However, until now, the PglS bioconjugation system was only compatible with its native substrate ComP or with an engineered EPA fusion containing a 117-amino acid ComP fragment. As such, the size of the ComP fragment limited its application in designing and developing next-generation bioconjugate vaccines with the PglS system. A smaller ComP sequon affords numerous advantages. First, it can be engineered into the coding sequence of other carrier proteins, something that is not easily possible with the larger 117-amino acid ComP fragment. While we limited our current internal sequon engineering to EPA, a similar strategy of using solvent-exposed loops for ComP sequon integration could be used for other common vaccine carriers such as CRM_{197}, the tetanus toxin C fragment or the cholera toxin B subunit. A PglS-derived bioconjugate vaccine would ideally not rely on using standalone ComP or ComP as a separate domain in a fusion, as this protein has not previously been used as a carrier in conjugate vaccines developed for human use. In contrast, EPA and CRM_{197} are established carriers with a track record of clinical use and safety in humans (Berti and Adamo 2018). Using a short, internal ComP sequon (ideally the 11-amino acid sequence IASGASAATTN) results in an engineered protein that closely resembles the native EPA and is thus more favorable from a vaccine development perspective. Second, shortening the ComP fragment significantly reduces the mass of the carrier protein, which increases the polysaccharide-to-protein ratio in the resulting bioconjugate. Here, we reduced the size of the ComP fragment from the previously used 117 amino acids to 11 amino acids, a difference of 10.9 kDa, ultimately reducing carrier mass by 14%. We have also shown that at least the iGTΔ3-4 sequon can be integrated at multiple sites on EPA. Thus, the polysaccharide-to-protein ratio could be further increased by engineering multiple ComP sequons into a single protein, an approach that has been successfully used with EPA and PglB (Ihssen and others 2010). Finally, a smaller sequon facilitates screening for optimized glycosylation sequences for PglS, as was done for PglL (Pan and others 2016). We identified several other candidate ComP proteins in this study and these proteins represent additional sequence space that can be utilized in order to test different ComP sequons or combinations thereof, with PglS to produce a range of different bioconjugates.

While $ComP_{110264}$ could be truncated to an 11-amino acid sequon, some of the longer fragments including iGTcc, iGTΔ1-2, iGTΔ2-3 and iGTΔ3-4 may be glycosylated slightly better than iGTΔ6-6 (Figure 4C). However, it is not clear if this modest increase in glycosylation translates into meaningful increases in bioconjugate production. We were surprised to find that some of the shortest $ComP_{110264}$ tags, iGTΔ9-8 and iGTΔ9-9, appeared to be better glycosylated than the larger sequons iGTΔ7-7 to iGTΔ8-9 (Figure 4C and Supplementary Figure S9). Further experiments will be necessary to ascertain whether this is a general phenomenon with the ComP sequon or if it is specific to the site in EPA where the iGTs were integrated. Perhaps after truncation to five residues, the remaining sequence combined with the surrounding EPA residues was sufficient to generate a pseudo-sequon that was recognized more efficiently by PglS. The additional residues retained in the larger iGTΔ8-8, iGTΔ7-8 and iGTΔ8-7 sequons could have interfered with this low-level recognition and thus glycosylation by PglS. The iGT truncations also led to the identification of two residues in $ComP_{110264}$ that may be important for glycosylation: Ile77 and Asn87. Deletion of either residue resulted in a dramatic reduction in glycosylation (Figure 4C). However, we cannot say with certainty whether the loss in glycosylation is because these residues play a key role in the PglS recognition process or if it is due to a shortening of the sequon below a critical threshold when either is removed. Intriguingly, the alignment in Figure 5 shows that these two residues, or residues with similar chemical character, are conserved within putative ComP orthologs. The bacterial *N*-linking sequon includes threonine and aspartate residues that are proposed to form Van der Waals interactions and hydrogen bonds with residues in the active site of PglB (Gerber et al. 2013). We speculate that the ComP isoleucine and asparagine residues perform similar roles in the active site of PglS, properly orientating the peptide substrate in the OTase active site for glycan transfer. It will be interesting to determine whether other factors modulate the binding affinity of peptide substrate and PglS and whether amino acid substitutions or synthetic/hybrid sequons result in improved glycosylation above wild-type ComP sequences.

Our findings indicate that $ComP_{110264}$, and possibly other ComP proteins, are stabilized by a disulfide bond between Cys71 and Cys93. However, in our experiments, this putative disulfide bond was only required for glycosylation when the ComP fragment was fused to the N- or C-terminus of a carrier protein (Figure 1C–E). Many types of pilins are stabilized by disulfide bridges, supporting the supposition that these two cysteines do in fact form a crosslink in ComP proteins as well as in the fragments (Piepenbrink et al. 2016). When the ComP disulfide region was placed in a different structural context, i.e. on a surface loop of EPA rather than used as an individual domain, we observed glycosylation even without the two cysteine residues. As such, when ComP is used as a separate domain in a fusion, more residues are required for glycosylation than when the sequon is placed

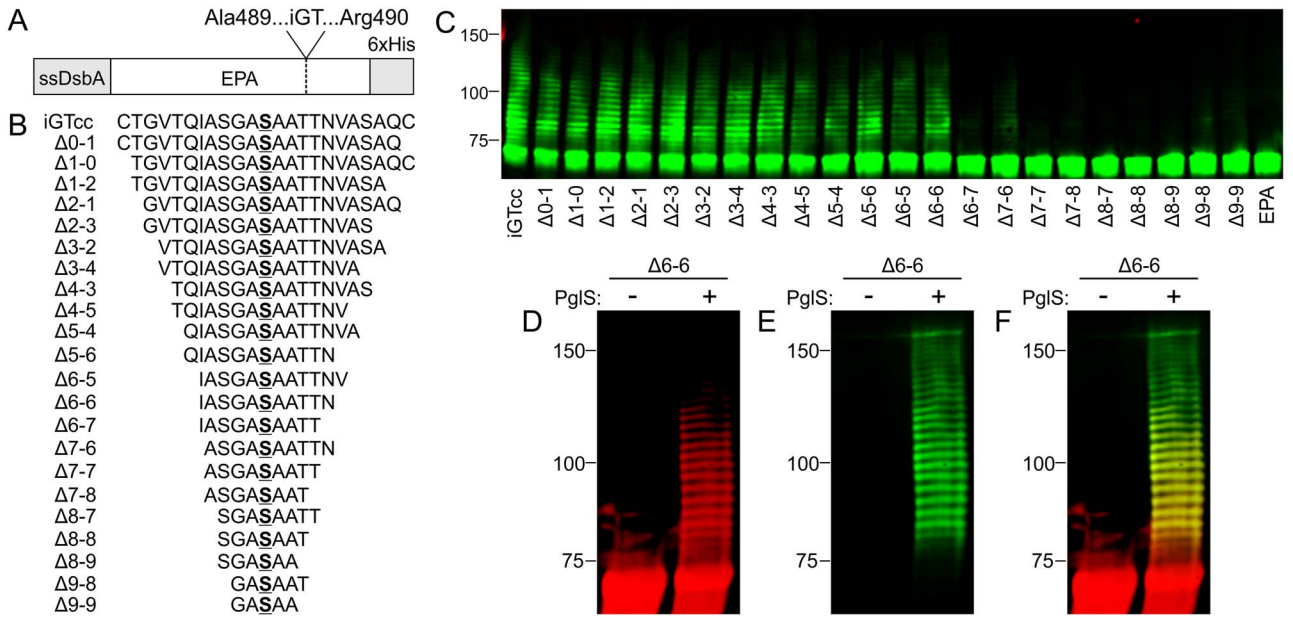

Fig. 4. Truncation of iGTcc and identification of a minimal ComP sequon required for PglS glycosylation. (**A**) Diagram of the internally-tagged EPA construct used for these experiments. Truncated variants of iGTcc between 22 and 5 amino acids in size were inserted into the EPA coding sequence between Ala489 and Arg489. (**B**) Amino acid sequences of the truncated iGT variants. The underlined, bolded serine is the glycosylation site. (**C**) Western blot analysis on periplasmic extracts of *E. coli* SDB1 expressing PglS, CPS8 and an EPA_{iGT} truncation carrier protein probing with an anti-EPA antibody. EPA_{iGTcc} is shown for comparison. The "EPA" lane corresponds to EPA lacking any ComP-derived sequences and serves as a negative control. Equivalent amounts of periplasmic extract based on OD_{600} were loaded per lane. (**D**) Western blot analysis of Ni-affinity chromatography purified $EPA_{iGT\Delta 6\text{-}6}$ from SDB1 cells expressing the CPS glycan in the presence (+) or absence (−) of PglS. Anti-His channel. (**E**) Anti-CPS8 channel. (**F**) Merge of **D** and **E**. Protein mass markers (in kDa) are indicated to the left of panels **C**–**F**.

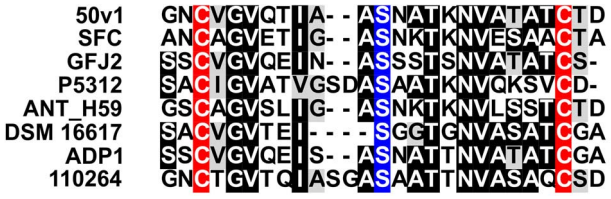

Fig. 5. Multiple sequence alignment of eight putative ComP orthologs. A conserved pair of cysteine residues (red) is found flanking the serine glycosylation site (blue) in all ComP orthologs. The sequence alignment was generated using T-Coffee (http://tcoffee.crg.cat/).

within a heterologous protein. It is possible that the disulfide bridge in ComP has additional roles besides protein stability, such as providing a conformational epitope that is preferentially recognized by PglS. Both of the internal sites that we chose on the surface of EPA for the integration of the iGTs are in hairpin loop structures (Figure 3A and Supplementary Figure S10). The formation of a covalent cysteine crosslink in native ComP could result in the formation of a hairpin-like structure similar to that where we integrated the iGTs in EPA. Although it is currently unknown whether PglS glycosylates folded or unfolded proteins, in the former case, it is conceivable that PglS recognition and transfer of a lipid-linked glycan requires a secondary structure, such as a hairpin loop, containing the sequon, rather than only a linear sequence of conserved amino acids. If this is the case, integration of the iGTs in the hairpin on the surface of EPA might mimic a native, disulfide-stabilized conformation in ComP that is necessary for PglS recognition. Experiments integrating the ComP sequon into non-hairpin structures on the surface of EPA or other carriers are necessary to confirm this hypothesis.

Using the $ComP_{C1}$ domain, we establish that the PglS bioconjugation system can be extended to other types of immunogenic carriers besides EPA, namely CRM_{197}. As shown above, we generated a bioconjugate containing the CPS8 glycan using the CRM_{197}–$ComP_{C1}$ carrier. CRM_{197} is the most common carrier protein in the global conjugate vaccine market. It is used as the sole carrier protein for the 13-valent blockbuster pneumococcal Prevnar13 conjugate vaccine that is licensed in more than 100 countries, as well in other vaccines against *N. meningitidis* and *H. influenza* type B (Berti and Adamo 2018). CRM_{197} will also be used for Pfizer's upcoming 20-valent pneumococcal conjugate vaccine (PCV), which includes a pneumococcal serotype 8 conjugate. However, in Pfizer's recently announced Phase 3 study on their 20-valent PCV in vaccine-naïve adults, immunogenicity responses to serotype 8 did not pass noninferiority criteria comparing individuals vaccinated with the polysaccharide-only vaccine Pneumovax 23 (Pfizer 2020). It is unknown whether the serotype 8 noninferiority results are due to alterations of the serotype 8 conjugate during manufacture that decrease immunogenicity, a result of carrier protein suppression that can decrease the immune response to multivalent conjugate vaccines using the same carrier protein (Pobre et al. 2014) or simply coincidental. Nevertheless, testing serotype 8 bioconjugates as well as other pneumococcal bioconjugates in a head-to-head study with PCV 20 will be required to determine if bioconjugates can offer improved immunogenicity. In our studies above, we fused the C-terminal $ComP_{C1}$ fragment to the N- or C-terminus of CRM_{197}; however, one or multiple smaller ComP iGT sequons integrated within the CRM_{197} sequence are also expected to be glycosylated

by PglS given that the iGTΔ3-4 sequon was efficiently glycosylated at two separate internal sites of EPA. The crystal structures of both the diphtheria toxin and CRM_{197} are known (Malito et al. 2012), aiding future ComP iGT sequon engineering. Moreover, the CD4+ T-cell epitopes of CRM_{197} are also known, allowing ComP iGT sequon integration without compromising the natural immunological properties of CRM_{197} to elicit strong T-cell-dependent responses (Raju et al. 1995).

In conclusion, these results represent a significant improvement in our understanding of bacterial *O*-linking sequons and greatly expand the glycoengineering toolbox. ComP has now been utilized for bioconjugate production with PglS in several different incarnations: as a full-length native protein, as a truncated 117 amino acid domain fusion, as a C- or N-terminal 25–45 amino acid disulfide-stabilized peptide fusion and as a minimal 11–23-amino acid sequon placed inside a heterologous carrier protein. This collectively demonstrates the flexible, modular possibilities in carrier protein design with the PglS bioconjugation system. The ability to use small, 11–23-amino acid ComP sequons offers numerous advantages over larger fusions, including the ability to generate bioconjugates that closely resemble commonly used, FDA-approved carrier proteins and variants thereof that contain multiple ComP sequons. Future studies will determine whether synthetic or hybrid sequons show improved glycosylation with PglS or if the wild-type ComP sequences are preferred. Discovery and experimental validation of additional ComP orthologs will also increase the sequence space available for the development of improved next-generation bioconjugate carrier proteins for PglS.

Materials and methods

Primers, media and antibiotics

All primers used in this study are presented in Supplementary Table SI. All antibiotics were purchased from Gold Biotechnology and used at the following concentrations: ampicillin, 100 μg/mL, spectinomycin, 50 μg/mL, tetracycline, 10 μg/mL, kanamycin, 20 μg/mL, chloramphenicol, 20 μg/mL. *Escherichia coli* strains were maintained on LB agar plates with appropriate antibiotics. All *E. coli* strains were cultured in LB media.

Cloning of PglS into pEXT21

The pEXT21 vector used for *pglS* cloning was previously characterized (Dykxhoorn et al. 1996). The *pglS* gene was PCR amplified using *A. baylyi* ADP1 genomic DNA as template. Forward and reverse primers used to amplify *pglS* contained 20 bp of homology to the pEXT21 multiple cloning site. The PCR product was purified and subsequently gel extracted using GeneJet kits (ThermoFisher, USA). Gel-purified *pglS* was then cloned into PCR-linearized pEXT21 using the NEBuilder HiFi DNA Assembly kit (New England Biolabs, Massachusetts, USA). The reaction was transformed into Stellar *E. coli* cells and transformants were selected on L-agar supplemented with spectinomycin. A correct clone containing the pEXT21–$PglS_{ADP1}$ plasmid was sequence verified and used subsequently for bioconjugation experiments.

High-throughput cloning of $ComP_{110264}$ fragments onto EPA

The pEXT20 vector used for carrier protein cloning was previously characterized (Dykxhoorn et al. 1996). The *comP*$_{110264}$ gene containing 15 base pairs homology on the 5′ and 3′ ends to the multiple cloning site of pEXT20 was synthetically built as a gBlock (Integrated DNA Technologies, IDT, Iowa, USA). The reconstituted *comP*$_{110264}$ gBlock was cloned into pEXT20 linearized via PCR using the In-Fusion HD EcoDry Kit (Takara Bio, USA). The In-Fusion reaction was transformed into *E. coli* Stellar cells and transformants were selected on L-agar supplemented with ampicillin. A correct clone containing the pEXT20-*comP*$_{110264}$ plasmid was sequence verified and used subsequently as template for all $ComP_{110264}$ fragment cloning. Primers to amplify the 80+ ComP fragments (listed in Supplementary Table SI) were used in a 96-well plate PCR format with the pEXT20–$ComP_{110264}$ plasmid as template. Each forward primer contained 15-bp homology to the GGGS linker immediately downstream of EPA in the pCH4 plasmid and each reverse primer contained 15-bp homology to the hexahistidine tag in the pCH4 plasmid. PCR products were purified using a Vac-Man 96 Vacuum Manifold and the Wizard SV 96 PCR Clean-up System (Promega, USA). Purified PCR products were then cloned onto the C-terminus of EPA in place of the 117-amino acid fragment of $ComP_{110264}$ using linearized pCH4 as template in an In-Fusion HD EcoDry reaction. The In-Fusion reactions were transformed into Stellar cells according to the manufacturers recommendations and plated on L-agar supplemented with ampicillin. Clones were individually picked in 96-well plate format, cultured and mini-prepped using a Wizard SV 96 Plasmid DNA Purification System. Correct clones containing the each $ComP_{110264}$ fragment were sequence verified.

CRM197–$ComP_{C1}$ and $ComP_{C1}$–CRM197 cloning strategy

A synthetic gene consisting of the FlgI signal sequence fused to the CRM_{197} coding sequence (excluding its start codon) in-turn fused to a hexahistidine tag and stop codon along with 15-bp homology on the 5′ and 3′ ends to the multiple cloning site of pEXT20 was synthetically built as a gBlock (IDT). The reconstituted FlgI–CRM_{197}–hexahistidine gBlock was subsequently cloned into pEXT20 linearized via PCR using an NEBuilder HiFi DNA Assembly kit. The reaction was transformed into Stellar *E. coli* cells and transformants were selected on L-agar supplemented with ampicillin. A correct clone containing the pEXT20–FlgI–CRM_{197}–hexahistidine plasmid was sequence verified and used subsequently as template for cloning the $ComP_{C1}$ fragments N- or C-terminal of the CRM_{197} coding sequence. Briefly, to generate the CRM_{197}–$ComP_{C1}$ construct, the $ComP_{C1}$ fragment was amplified with primers containing 5′ homology to the end of the CRM_{197} sequence and 3′ homology to the hexahistidine tag immediately downstream of the CRM197 coding sequence. The $ComP_{C1}$ PCR product was purified and subsequently gel extracted using a GeneJet kit. Gel-purified $ComP_{C1}$ was then cloned between CRM_{197} and the hexahistidine tag linearized via PCR using the NEBuilder HiFi DNA Assembly kit. The reaction was transformed into Stellar *E. coli* cells and transformants selected on L-agar supplemented with ampicillin. A correct clone containing the pEXT20–CRM_{197}–$ComP_{C1}$ plasmid was sequence verified. A similar approach was used to generate the pEXT20–$ComP_{C1}$–CRM_{197} plasmid except the 5′ and 3′ 20-bp homology arms on the primers used to amplify the $ComP_{C1}$ fragment were homologous to the FlgI signal sequence and start of the CRM_{197} coding sequence, respectively.

Internal EPA_{iGT} cloning strategy

The iGT $ComP_{110264}$ fragments were PCR amplified using primers listed in Supplementary Table SI. For iGT fragments cloned between Ala489 and Arg490, forward and reverse primers contained 20-bp homology to allow for site-directed recombination between Ala489

and Arg490 residues using the NEBuilder HiFi DNA Assembly kit. Briefly, iGT fragments were PCR amplified using pEXT20–$ComP_{110264}$ as template, purified and gel extracted. An iGT fragment was then cloned into pEXT20–DsbA–EPA linearized via PCR using the NEBuilder HiFi DNA Assembly kit. The reaction was transformed into Stellar *E. coli* cells and transformants were selected on L-agar supplemented with ampicillin. Clones containing pEXT20–EPA_{iGT} plasmids were sequence verified. A similar approach was used to generate the iGTΔ3-4 sequence between residues Glu548 and Gly549 of EPA.

Bioconjugation experiments

Escherichia coli strain SDB1 was used for all bioconjugation experiments. SDB1 was made electrocompetent by centrifuging cells during mid-log phase and washing in sterile 10% glycerol twice prior to freezing at −80°C. Plasmids were electroporated into SDB1 and transformants selected on LB agar plates containing appropriate antibiotics. The following day, 8–10 single colonies were picked and inoculated into LB broth supplemented with antibiotics for overnight outgrowth at 30°C. Bioconjugation screening experiments with the 87 EPA–ComP fragment fusion constructs were performed in deep-well plates containing 1 mL of LB broth. Deep-well plates were grown at 225 rpm at 30°C for 4 h, induced with 1 mM IPTG, and grown overnight. Whole-cell lysates corresponding to 0.9 OD_{600} units were removed, centrifuged and the pellets resuspended in 1× Laemmli buffer. For bioconjugation experiments in shake flasks, desired starter strains were cultured overnight and the next morning diluted to a starting OD_{600} of 0.05 in 70 mL LB media in 125 mL flasks. Cultures were grown at 30 °C and induced with 1 mM IPTG after 3–4 h when the OD_{600} had reach 0.4–0.6. The temperature was then maintained at 30 °C or reduced to 20 °C depending on the experiment. Typically, strains were cultured for 20–24 h before harvest. The bioconjugation experiments were performed in triplicate with similar results in each trial.

Protein extraction and western blotting

For periplasmic extraction of proteins, we used the osmotic shock method described by (Caillava et al. 2019). Briefly, *E. coli* cells were harvested by centrifugation at 3000 RCF and the supernatant removed. The pellet was resuspended in hyperosmotic buffer (20 mM Tris–HCl pH 8.0, 20% w/v sucrose, 1 mM EDTA) to an OD_{600} of 5.0 and incubated on ice while shaking for 30 minutes. The suspension was centrifuged, supernatant removed, and the pellet suspended in hypoosmotic buffer (20 mM Tris–HCl pH 8.0, 1% w/v sodium deoxycholate) to an OD_{600} of 5.0 and incubated on ice while shaking for 1 h. The samples were centrifuged and supernatant containing the periplasmic fraction was removed and diluted in Laemmli buffer. All samples were normalized to the same OD_{600} per lane. The periplasmic extracts were separated on SDS-PAGE gels (BioRad, California, USA) and proteins transferred to a nitrocellulose membrane using a semi-dry transfer system. Membranes were incubated with blocking buffer (Li-Cor, Nebraska, USA) for 1 h before incubating with primary antibodies for another hour in 1:1 blocking buffer and TBST (10 mM Tris–HCl, 150 mM NaCl, 0.05% v/v Tween-20). We used commercial rabbit anti-EPA and mouse anti-6×His antibodies from Millipore-Sigma (USA), mouse anti-CRM_{197} antibodies from AIC Biotech and pneumococcal rabbit antisera for serotype 8 from SSI Diagnostica (via Cedarlane Labs, North Carolina, USA). Following incubation with primary antibodies, membranes were washed with TBST for 15 min. The membranes were then incubated with IRDye 680RD goat anti-mouse and/or IRDye 800CW goat anti-rabbit secondary antibodies (Li-Cor) 1:1 blocking buffer and TBST for 30 min followed by another 15-min TBST wash. The membranes were imaged using a Li-Cor Odyssey CLx.

Protein purification

Bioconjugate producing SDB1 cells were harvested by centrifugation and pellets suspended in 20 mM Tris, 10 mM imidazole and 500 mM NaCl pH 8.0. Bacteria were lysed by sonication using a Sonica Q700 sonicator with the following settings: 75% amplitude, 5 s on, 5 s off for 20 min. Lysates were centrifuged at 25,000 rcf for 30 min and the supernatants collected and filtered with a 0.2-µm polyethersulfone (PES) filter. Filtered lysates were passed over Ni-NTA resin, washed with 20 column volumes of 20 mM Tris, 10 mM imidazole and 500 mM NaCl pH 8.0 and eluted with five column volumes of 20 mM Tris, 300 mM imidazole and 500 mM NaCl pH 8.0. The Ni-NTA eluate was diafiltrated using Amicon concentrators into 20 mM Tris pH 8.0, filtered with a 0.2-µm PES filter and loaded onto a Mono Q 5/50 GL column. Proteins were eluted using a linear gradient (0–25%) over 50 column volumes with 20 mM Tris 1 M pH 8.0. Fractions containing bioconjugates were identified via SDS-PAGE analysis, pooled, concentrated using Amicon concentrators and subsequently loaded onto a Superdex 200 Increase 10/300 GL column. Bioconjugates were eluted using an isocratic flow of phosphate-buffered saline. Bioconjugates were identified via SDS-PAGE and western blot analysis.

Supplementary data

Supplementary data are available at *Glycobiology* online.

Funding

National Institute of Allergy and Infectious Diseases (project number R44AI131742) awarded to VaxNewMo, LLC.

Conflict of interest statement

C.J.K, L.S.R and C.M.H. have a financial stake in VaxNewMo, a for-profit entity developing bioconjugate vaccines using patented technology derived from the data presented in this and other published manuscripts.

Abbreviations

CPS, capsular polysaccharide; CRM_{197}, cross-reactive material 197, genetically inactivated diphtheria toxin; EPA, *Pseudomonas aeruginosa* exotoxin A, also known as exoprotein A; iGT, internal glycotag, derived from $ComP_{110264}$; OTase, oligosaccharyltransferase; PCV, pneumococcal conjugate vaccine; SDS-PAGE, sodium dodecyl sulfate-polyacrylamide gel electrophoresis.

References

Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M, Donohoe K, Harris D, Murphy L, Quail MA *et al.* 2006. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet*. 2:e31.

Berti F, Adamo R. 2018. Antimicrobial glycoconjugate vaccines: an overview of classic and modern approaches for protein modification. *Chem Soc Rev*. 47:9015–9025.

Berti F, Campisi E, Toniolo C, Morelli L, Crotti S, Rosini R, Romano MR, Pinto V, Brogioni B, Torricelli G *et al.* 2014. Structure of the type IX group B *Streptococcus* capsular polysaccharide and its evolutionary relationship with types V and VII. *J Biol Chem*. 289:23437–23448.

Caillava AJ, Ortiz GE, Melli LJ, Ugalde JE, Ciocchini AE, Comerci DJ. 2019. Improving bioreactor production of a recombinant glycoprotein in *Escherichia coli*: Effect of specific growth rate on protein glycosylation and specific productivity. *Biotechnol Bioeng*. 116:1427–1438.

Dow JM, Mauri M, Scott TA, Wren BW. 2020. Improving protein glycan coupling technology (PGCT) for glycoconjugate vaccine production. *Expert Rev Vaccines*. 19:507–527.

Dykxhoorn DM, St Pierre R, Linn T. 1996. A set of compatible tac promoter expression vectors. *Gene*. 177:133–136.

Faridmoayer A, Fentabil MA, Haurat MF, Yi W, Woodward R, Wang PG, Feldman MF. 2008. Extreme substrate promiscuity of the *Neisseria* oligosaccharyl transferase involved in protein *O*-glycosylation. *J Biol Chem*. 283:34596–34604.

Faridmoayer A, Fentabil MA, Mills DC, Klassen JS, Feldman MF. 2007. Functional characterization of bacterial oligosaccharyltransferases involved in *O*-linked protein glycosylation. *J Bacteriol*. 189:8088–8098.

Feldman MF, Mayer Bridwell AE, Scott NE, Vinogradov E, McKee SR, Chavez SM, Twentyman J, Stallings CL, Rosen DA, Harding CM. 2019. A promising bioconjugate vaccine against hypervirulent *Klebsiella pneumoniae*. *Proc Natl Acad Sci U S A*. 116:18655–18663.

Feldman MF, Wacker M, Hernandez M, Hitchen PG, Marolda CL, Kowarik M, Morris HR, Dell A, Valvano MA, Aebi M. 2005. Engineering N-linked protein glycosylation with diverse O antigen lipopolysaccharide structures in *Escherichia coli*. *Proc Natl Acad Sci U S A*. 102:3016–3021.

Frasch CE. 2009. Preparation of bacterial polysaccharide-protein conjugates: analytical and manufacturing challenges. *Vaccine*. 27:6468–6470.

Garcia-Quintanilla F, Iwashkiw JA, Price NL, Stratilo C, Feldman MF. 2014. Production of a recombinant vaccine candidate against Burkholderia pseudomallei exploiting the bacterial N-glycosylation machinery. *Front Microbiol*. 5:381.

Geno KA, Gilbert GL, Song JY, Skovsted IC, Klugman KP, Jones C, Konradsen HB, Nahm MH. 2015. Pneumococcal capsules and their types: past, present, and future. *Clin Microbiol Rev*. 28:871–899.

Gerber S, Lizak C, Michaud G, Bucher M, Darbre T, Aebi M, Reymond JL, Locher KP. 2013. Mechanism of bacterial oligosaccharyltransferase: in vitro quantification of sequon binding and catalysis. *J Biol Chem*. 288:8849–8861.

Goffin P, Dewerchin M, De Rop P, Blais N, Dehottay P. 2017. High-yield production of recombinant CRM197, a non-toxic mutant of diphtheria toxin, in the periplasm of *Escherichia coli*. *Biotechnol J*. 12:1700168.

Harding CM, Feldman MF. 2019. Glycoengineering bioconjugate vaccines, therapeutics, and diagnostics in *E. coli*. *Glycobiology*. 29:519–529.

Harding CM, Nasr MA, Kinsella RL, Scott NE, Foster LJ, Weber BS, Fiester SE, Actis LA, Tracy EN, Munson RS Jr *et al.* 2015. *Acinetobacter* strains carry two functional oligosaccharyltransferases, one devoted exclusively to type IV pilin, and the other one dedicated to *O*-glycosylation of multiple proteins. *Mol Microbiol*. 96:1023–1041.

Harding CM, Nasr MA, Scott NE, Goyette-Desjardins G, Nothaft H, Mayer AE, Chavez SM, Huynh JP, Kinsella RL, Szymanski CM *et al.* 2019. A platform for glycoengineering a polyvalent pneumococcal bioconjugate vaccine using *E. coli* as a host. *Nat Commun*. 10:891.

Hatz CF, Bally B, Rohrer S, Steffen R, Kramme S, Siegrist CA, Wacker M, Alaimo C, Fonck VG. 2015. Safety and immunogenicity of a candidate bioconjugate vaccine against *Shigella dysenteriae* type 1 administered to healthy adults: A single blind, partially randomized phase I study. *Vaccine*. 33:4594–4601.

Hug I, Feldman MF. 2011. Analogies and homologies in lipopolysaccharide and glycoprotein biosynthesis in bacteria. *Glycobiology*. 21:138–151.

Huttner A, Hatz C, van den Dobbelsteen G, Abbanat D, Hornacek A, Frolich R, Dreyer AM, Martin P, Davies T, Fae K *et al.* 2017. Safety, immunogenicity, and preliminary clinical efficacy of a vaccine against extraintestinal pathogenic *Escherichia coli* in women with a history of recurrent urinary tract infection: a randomised, single-blind, placebo-controlled phase 1b trial. *Lancet Infect Dis*. 17:528–537.

Ihssen J, Kowarik M, Dilettoso S, Tanner C, Wacker M, Thony-Meyer L. 2010. Production of glycoprotein vaccines in *Escherichia coli*. *Microb Cell Fact*. 9:61.

Kay EJ, Yates LE, Terra VS, Cuccui J, Wren BW. 2016. Recombinant expression of *Streptococcus pneumoniae* capsular polysaccharides in *Escherichia coli*. *Open Biol*. 6:150243.

Kowarik M, Young NM, Numao S, Schulz BL, Hug I, Callewaert N, Mills DC, Watson DC, Hernandez M, Kelly JF *et al.* 2006. Definition of the bacterial N-glycosylation site consensus sequence. *EMBO J*. 25:1957–1966.

Malito E, Bursulaya B, Chen C, Lo Surdo P, Picchianti M, Balducci E, Biancucci M, Brock A, Berti F, Bottomley MJ *et al.* 2012. Structural basis for lack of toxicity of the diphtheria toxin mutant CRM197. *Proc Natl Acad Sci U S A*. 109:5229–5234.

Nothaft H, Szymanski CM. 2010. Protein glycosylation in bacteria: sweeter than ever. *Nat Rev Microbiol*. 8:765–778.

Pan C, Sun P, Liu B, Liang H, Peng Z, Dong Y, Wang D, Liu X, Wang B, Zeng M *et al.* 2016. Biosynthesis of conjugate vaccines using an O-linked glycosylation system. *MBio*. 7:e00443–e00416.

Pan YJ, Lin TL, Chen CT, Chen YY, Hsieh PF, Hsu CR, Wu MC, Wang JT. 2015. Genetic analysis of capsular polysaccharide synthesis gene clusters in 79 capsular types of *Klebsiella* spp. *Sci Rep*. 5:15573.

Pfizer. 2020. 20-valent pneumococcal conjugate vaccine demonstrated comparable safety and immunogenicity profile to licensed pneumococcal vaccines. https://investors.pfizer.com/investor-news/press-release-details/2020/Pfizer-Announces-Top-Line-Results-from-Phase-3-Study-of-20-Valent-Pneumococcal-Conjugate-Vaccine-in-Pneumococcal-Vaccine-Nave-Adults-Aged-18-Years-or-Older/default.aspx.

Piepenbrink KH, Lillehoj E, Harding CM, Labonte JW, Zuo X, Rapp CA, Munson RS Jr, Goldblum SE, Feldman MF, Gray JJ *et al.* 2016. Structural diversity in the type iv pili of multidrug-resistant *Acinetobacter*. *J Biol Chem*. 291:22924–22935.

Pobre K, Tashani M, Ridda I, Rashid H, Wong M, Booy R. 2014. Carrier priming or suppression: understanding carrier priming enhancement of anti-polysaccharide antibody response to conjugate vaccines. *Vaccine*. 32:1423–1430.

Raetz CR, Whitfield C. 2002. Lipopolysaccharide endotoxins. *Annu Rev Biochem*. 71:635–700.

Raju R, Navaneetham D, Okita D, Diethelm-Okita B, McCormick D, Conti-Fine BM. 1995. Epitopes for human CD4+ cells on diphtheria toxin: structural features of sequence segments forming epitopes recognized by most subjects. *Eur J Immunol*. 25:3207–3214.

Rappuoli R, De Gregorio E, Costantino P. 2019. On the mechanisms of conjugate vaccines. *Proc Natl Acad Sci U S A*. 116:14–16.

Reglinski M, Ercoli G, Plumptre C, Kay E, Petersen FC, Paton JC, Wren BW, Brown JS. 2018. A recombinant conjugated pneumococcal vaccine that protects against murine infections with a similar efficacy to Prevnar-13. *NPJ Vaccines*. 3:53.

Schaffer C, Messner P. 2017. Emerging facets of prokaryotic glycosylation. *FEMS Microbiol Rev*. 41:49–91.

Sun P, Pan C, Zeng M, Liu B, Liang H, Wang D, Liu X, Wang B, Lyu Y, Wu J *et al.* 2018. Design and production of conjugate vaccines against *S. Paratyphi* A using an O-linked glycosylation system in vivo. *NPJ Vaccines*. 3:4.

Szymanski CM, Yao R, Ewing CP, Trust TJ, Guerry P. 1999. Evidence for a system of general protein glycosylation in *Campylobacter jejuni*. *Mol Microbiol*. 32:1022–1030.

Valvano MA. 2003. Export of O-specific lipopolysaccharide. *Front Biosci*. 8:s452–s471.

van den Dobbelsteen G, Fae KC, Serroyen J, van den Nieuwenhof IM, Braun M, Haeuptle MA, Sirena D, Schneider J, Alaimo C, Lipowsky G *et al.* 2016. Immunogenicity and safety of a tetravalent *E. coli* O-antigen bioconjugate vaccine in animal models. *Vaccine*. 34:4152–4160.

Vik A, Aas FE, Anonsen JH, Bilsborough S, Schneider A, Egge-Jacobsen W, Koomey M. 2009. Broad spectrum O-linked protein glycosylation in the human pathogen *Neisseria gonorrhoeae*. *Proc Natl Acad Sci U S A*. 106:4447–4452.

Wacker M, Feldman MF, Callewaert N, Kowarik M, Clarke BR, Pohl NL, Hernandez M, Vines ED, Valvano MA, Whitfield C *et al.* 2006. Substrate specificity of bacterial oligosaccharyltransferase suggests a common transfer mechanism for the bacterial and eukaryotic systems. *Proc Natl Acad Sci U S A*. 103:7088–7093.

Wacker M, Linton D, Hitchen PG, Nita-Lazar M, Haslam SM, North SJ, Panico M, Morris HR, Dell A, Wren BW *et al.* 2002. N-linked glycosylation in *Campylobacter jejuni* and its functional transfer into *E. coli*. *Science*. 298:1790–1793.

Wacker M, Wang L, Kowarik M, Dowd M, Lipowsky G, Faridmoayer A, Shields K, Park S, Alaimo C, Kelley KA *et al.* 2014. Prevention of *Staphylococcus aureus* infections by glycoprotein vaccines synthesized in *Escherichia coli*. *J Infect Dis*. 209:1551–1561.

Wedekind JE, Trame CB, Dorywalska M, Koehl P, Raschke TM, McKee M, FitzGerald D, Collier RJ, McKay DB. 2001. Refined crystallographic structure of *Pseudomonas aeruginosa* exotoxin A and its implications for the molecular mechanism of toxicity. *J Mol Biol*. 314: 823–837.