Supporting information:

Co-conserved sequence motifs are predictive of substrate specificity in a family of monotopic phosphoglycosyl transferases

Alyssa J. Anderson,[a,‡] Greg J. Dodge[a, ‡] , Karen N. Allen[b], Barbara Imperiali[a,*]

[a] Department of Biology and Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

[b] Department of Chemistry, Boston University, Boston, MA 02215, USA.

[‡]Equal contributions

*Corresponding author: imper@mit.edu

Materials and Methods

PglC Mutagenesis

PglC variants were generated using primers designed in the QuikChange primer tool (Agilent). Successful mutations were confirmed by Sanger sequencing. Forward and reverse primers are shown below, with altered codons underlined. Leucine was chosen as the aliphatic residue to introduce instead of alanine at position 150 because a small percentage of monoPGTs have a leucine at position 150.

Variant	Primers
W150F	5′-CGTATTCAAATTTTTTCTC**AAA**ACTTATGGCGTTTCTGC-3′ 5′-GCAGAAACGCCATAAGT**TTT**GAGAAAAAATTTGAATACG-3′
W150Y	5′-CGTATTCAAATTTTTTCTC**ATA**ACTTATGGCGTTTCTGC-3′ 5′-GCAGAAACGCCATAAGT**TAT**GAGAAAAAATTTGAATACG
W150L	5′-CGTATTCAAATTTTTTCTC**CAG**ACTTATGGCGTTTCTGC-3′ 5′-GCAGAAACGCCATAAGT**CTG**GAGAAAAAATTTGAATACG-3′
F197Y	5′-AGTTTTTGCCATT**ATA**TTTCTCCGTCGTCGCCTGC-3′ 5′-GCAGGCGACGACGGAGAAA**TAT**AATGGCAAAAACT-3′
F197W	5′-GAGTTAGTTTTTGCCATT**CCA**TTTCTCCGTCGTCGCCTGCC-3′ 5′-GGCAGGCGACGACGGAGAAA**TGG**AATGGCAAAAACTAACTC-3′
F197A	5′-GTTAGTTTTTGCCATT**CGC**TTTCTCCGTCGTCGCCTGCCCCTC-3′ 5′-GAGGGGCAGGCGACGACGGAGAAA**GCG**AATGGCAAAAACTAAC-3′

PglC expression and purification

BL21-CodonPlus (DE3)-RIL cells (Agilent) were transformed with a pET-His_6-SUMO-PglC construct for expression using the Studier auto-induction method. (*Protein Expr Purif* 2005 41 :207-34). A single colony was picked and starter cultures were grown overnight in 3 mL MDG media (0.5% (w/v) glucose, 0.25 (w/v) % aspartate, 2 mM $MgSO_4$, 25 mM Na_2HPO_4, 25 mM KH_2PO_4, 50 mM NH_4Cl, 5 mM Na_2SO_4 and 0.2× trace metal mix (from 1000× stock, Teknova, cat. # T1001) at 37 °C using kanamycin and chloramphenicol (30 µg/mL each). The overnight culture was transferred into 500 mL auto-induction media (1% (w/v) tryptone, 0.5% (w/v) yeast extract, 0.5% (v/v) glycerol, 0.05% (w/v) glucose, 0.2% (w/v) α-D-lactose, 2 mM $MgSO_4$, 25 mM Na_2HPO_4, 25 mM KH_2PO_4, 50 mM NH_4Cl, and 5 mM Na_2SO_4, 0.2× trace metal mix) containing kanamycin (90 µg/mL) and chloramphenicol (30 µg/mL). Cells were grown in a baffled Fernbach culture flask (2800 mL) at 200 rpm at 37 °C for 3h, after which time the temperature was reduced to 16 °C. The culture was allowed to grow for another 20 h and the cells were harvested at 3,700 × *g* for 30 min. The resulting cell pellet was washed with a buffer containing 50 mM HEPES, pH 7.5, 150 mM NaCl and used for protein purification.

Protein purification was carried out at 4 °C. A 20 g cell pellet was re-suspended in 100 mL buffer A (50 mM HEPES, pH 7.5, 150 mM NaCl) containing 50 mg lysozyme (RPI, cat. # L38100), 100 µL

EDTA-free protease inhibitor cocktail (EMD cat. # 539134) and 50 µL DNase I (NEB, cat. # M0303S). Cells were placed on a rotating mixer to tumble for 15 min at 4 °C followed by sonication (Sonics Vibra-Cell; 50% amplitude, 1 sec ON – 2 sec OFF, 2 × 1.5 min) for effective cell lysis. Cells were kept on ice during sonication and rested for 5 min in between the two sonication cycles. The resultant suspension was tumbled for 15 min at 4 °C followed by centrifugation at 9,000 × *g* for 45 min at 4 °C using a Ti45 rotor. The resulting supernatant was further centrifuged at 140,000 × *g* for 65 min at 4 °C. The membrane pellet, also known as cell envelope fraction (CEF), was resuspended in 2 mL of buffer A. The total volume of the solution was ~5 mL. To this resuspended CEF, 23 mL of buffer containing 50 mM HEPES, pH 7.5, 100 mM NaCl, 1% DDM (Anatrace, cat. # D310A) and 28 µL protease inhibitor cocktail solution was added. The suspension was tumbled overnight at 4 °C. The solution was centrifuged at 150,000 × *g* for 65 min at 4 °C using a Ti70 rotor. The supernatant was incubated with 1 mL fresh Ni-NTA resin pre-equilibrated with an equilibration buffer containing 50 mM HEPES, pH 7.5, 100 mM NaCl, 20 mM imidazole and 5% glycerol. After tumbling the protein solution with the resin for 1 h, the flow-through was separated. The column was washed with 20 mL of wash-1 buffer (equilibration buffer + 0.03% DDM), followed by 20 mL of wash-2 buffer (equilibration buffer containing 45 mM imidazole + 0.03% DDM). The protein was eluted from the column using elution buffer (equilibration buffer + 500 mM imidazole + 0.03% DDM). Elution fractions (2 × 1 mL) were combined and immediately desalted using a 5 ml HiTrap desalting column (GE Healthcare, cat. # 17-1408-01) that was pre-equilibrated with a desalting buffer containing 50 mM HEPES, pH 7.5, 100 mM NaCl, 0.03% DDM and 5% glycerol.

Purified His_6-SUMO-PglC variants were incubated with 0.14 equivalents of SUMO protease (*S. cerevisiae*) at 16 °C with gentle shaking at 80 rpm for 6 h. The SUMO protease was expressed and purified following a previously published protocol. The resulting solution was incubated with 250 µL fresh Ni-NTA resin that was pre-equilibrated with a buffer containing 50 mM HEPES, pH 7.5, 100 mM NaCl, 20 mM imidazole, 0.03% DDM and 5% glycerol. After 45 min incubation, the flow-through was collected, the column was washed with two column volumes of the desalting buffer and the wash fractions were combined with the flow-through.

Stability Measurements

Approximately 10 µL of 1 mg/mL *Cc* PglC variants were loaded in triplicate into Prometheus NT.48 high sensitivity capillaries. Sample turbidity was monitored over a temperature range of 20-80 °C, with a rate of increase of 1 degree per minute using the Nanotemper Prometheus Panta. Triplicate data sets were merged, and standard deviations and turbidity onset temperatures were determined using the Nanotemper PR.Stability Analysis software. Buffer signal was background subtracted in Excel, and data was plotted in GraphPad Prism.

Supplementary Tables and Figures

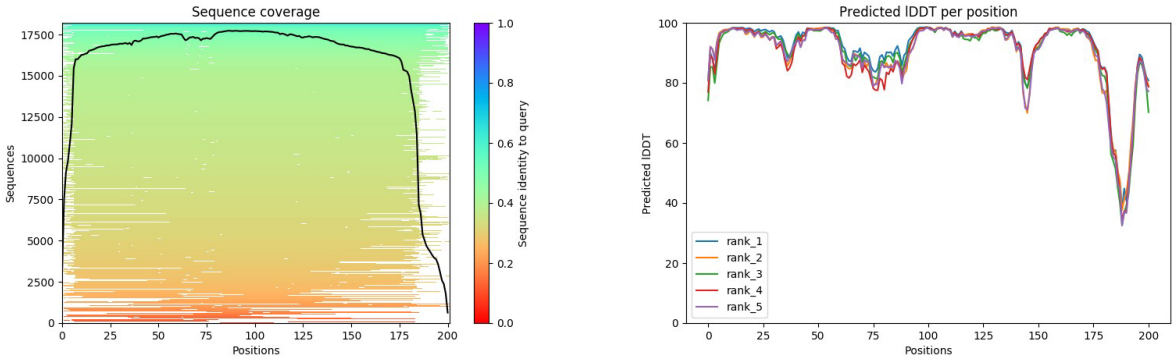


Figure S1: Sequence coverage and pLDDT score for full-length AlphaFold model of *C. concisus* PglC. Only a small proportion of the sm-PGTs sampled have an extended C-terminus, similar to that observed in diNAcBac-specific PGTs, such as those from *Campylobacter*.

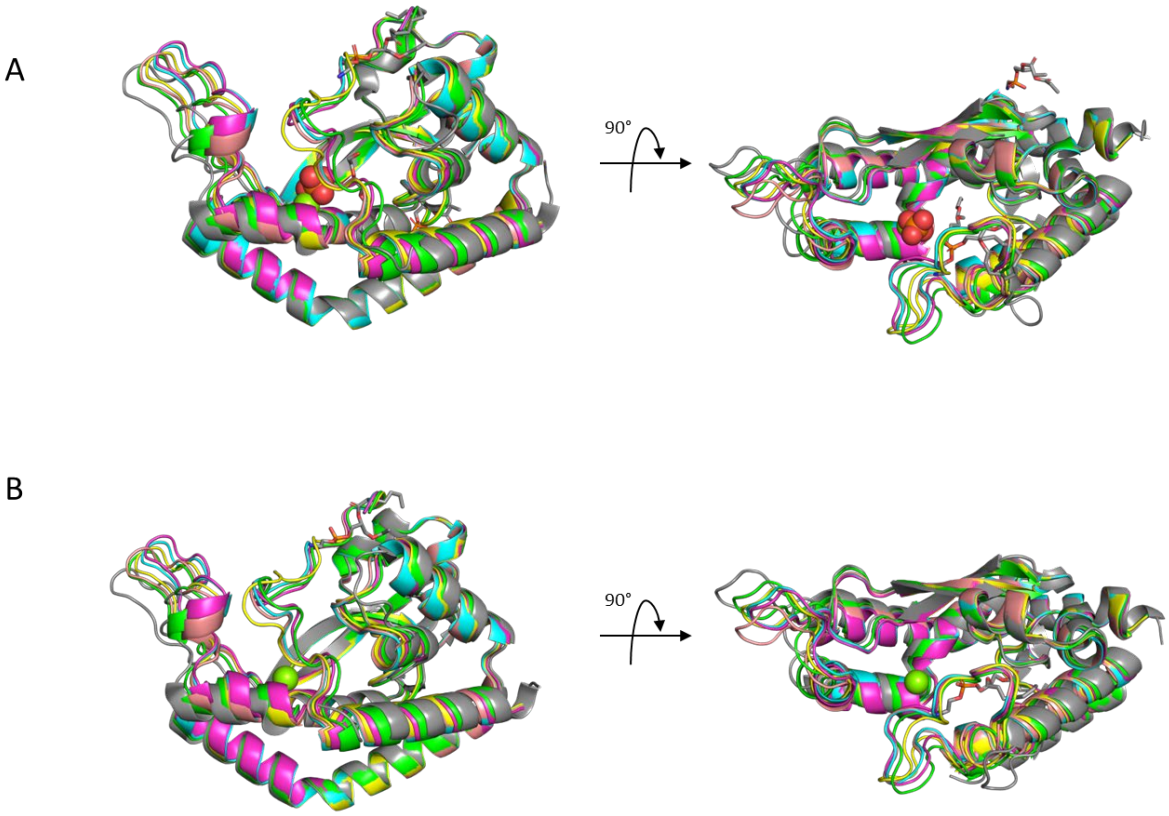
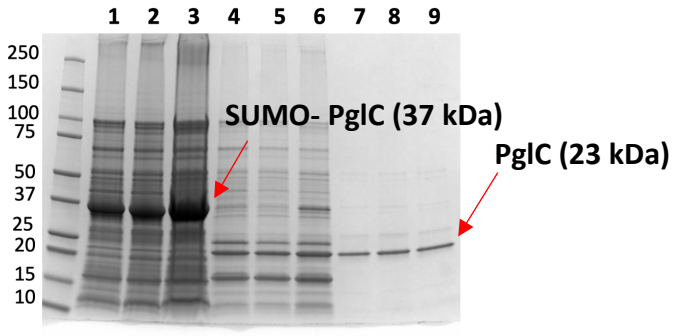


Figure S2: Superposition of *C. concisus* PglC experimental structure (gray; phosphatidylethanolamine head-group moieties shown in stick, bound Mg^{2+} shown as sphere) chain A (Top) and chain B (Bottom) with AlphaFold models Rank 1-5 ((cyan, magenta, yellow, tan, green).

Cc PglC W150 variants

1. SUMO F197A PglC
2. SUMO F197W PglC
3. SUMO F197Y PglC
4. F197A SUMO cleavage
5. F197W SUMO cleavage
6. F197Y SUMO cleavage
7. F197A PglC
8. F197W PglC
9. F197Y PglC



Cc PglC F197 variants

1. SUMO F197A PglC
2. SUMO F197W PglC
3. SUMO F197Y PglC
4. F197A SUMO cleavage
5. F197W SUMO cleavage
6. F197Y SUMO cleavage
7. F197A PglC
8. F197W PglC
9. F197Y PglC

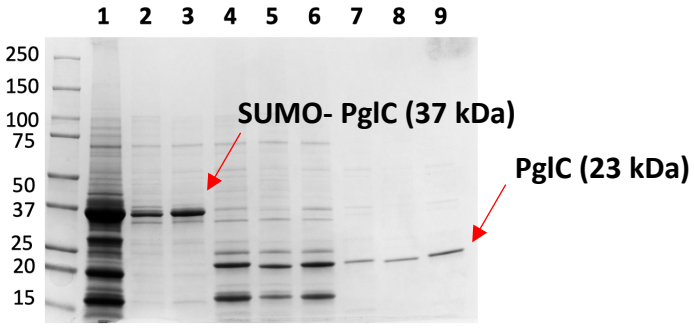


Figure S3: SDS-PAGE of samples from purification of *C. concisus* PglC aromatic box variants. SUMO-PglC elutions from Ni-NTA resin (Lanes 1-3), SUMO cleavage reactions with SUMO protease (Lanes 4-6), and the reverse Ni-NTA purification of SUMO cleaved PglC (Lanes 7-9) were characterized by SDS-PAGE.

.

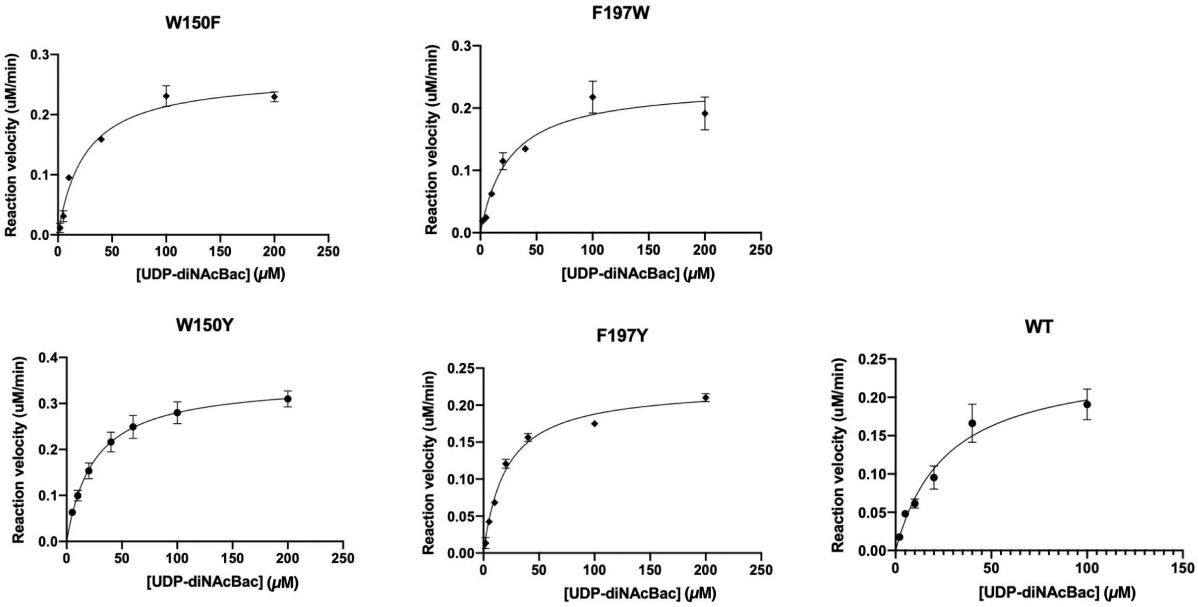
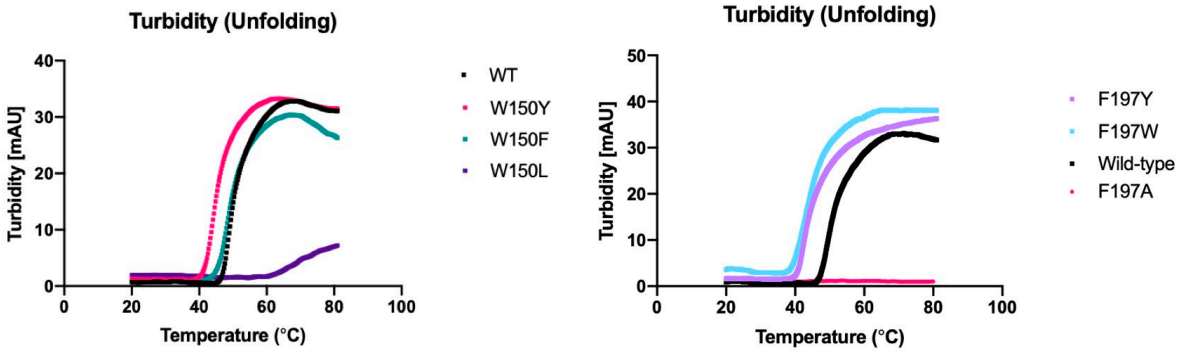


Figure S4: Kinetic analysis of *C. concisus* PglC variants. Data was plotted and fit to the Michaelis-Menten equation in GraphPad Prism Error bars represent the standard deviation from triplicate measurements. Assays were carried out using various concentrations of UDP-diNAcBac (2.5–200μM) with 0.3 nM PglC for WT and 3 nM for PglC variants. Reactions were quenched at several time points (3 min, 6 min, 9 min, 12 min, and 15 min) and reaction rates were calculated in Excel. Rates in the linear range with less than 10% substrate turnover were used for steady-state kinetic analysis. The data was fit to the Michaleis-Menten equation in GraphPad Prism.

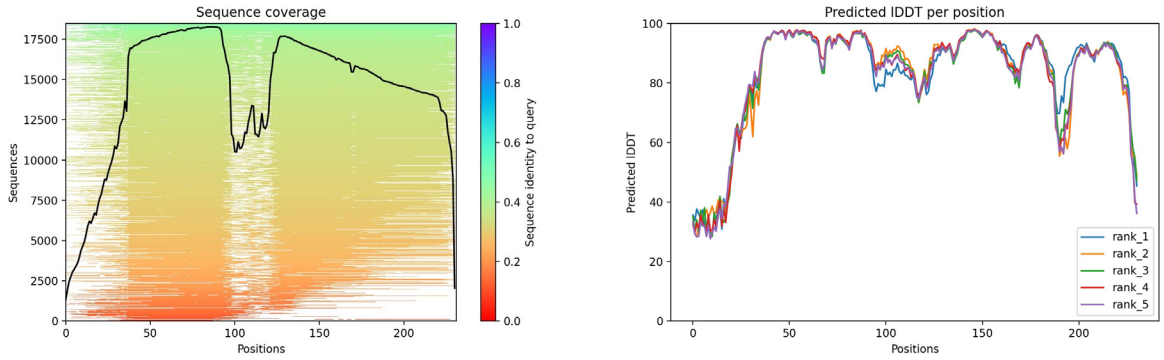


Variant	Turbidity Onset (°C)
WT	45.3 +/- 0.35
W150Y	39.5 +/- 0.38
W150F	42.1 +/- 0.35
W150L	n.d.
F197Y	38.31 +/- 0.42
F197W	36.08 +/- 0.97
F197A	n.d.

Figure S5: Thermal stability of *C. concisus* PglC aromatic box variants.
Turbidity of *C. concisus* PglC variants was measured with the Nanotemper Prometheus Panta over a temperature range of 20 – 80 °C. The temperature at which the turbidity of the sample began to increase is reported in the table as turbidity onset. Error represents the standard deviation from triplicate measurements. Stability measurements could not be made for the W150L and F197A variants, as they were unstable at room temperature.

Litoreibacter albidus PGT AlphaFold Model
Sequence Coverage and pLDDT

A



Oceanicola sp. PGT AlphaFold Model
Sequence Coverage and pLDDT

B

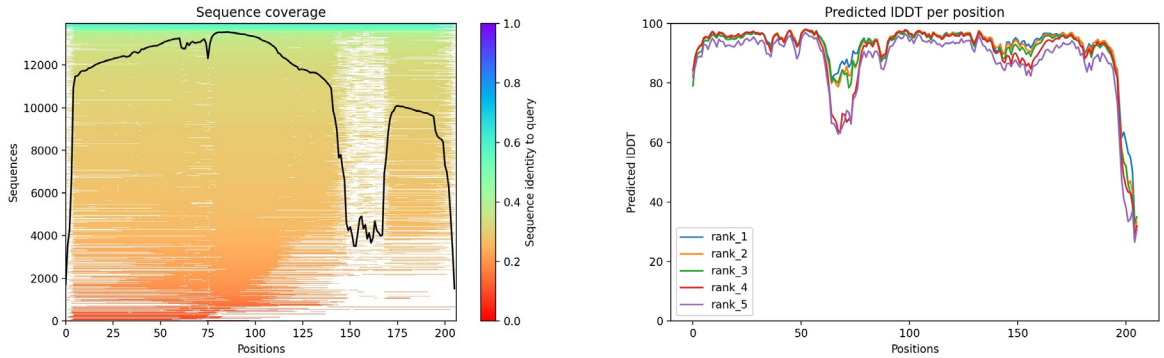


Figure S6: Sequence coverage and pLDDT scores from AlphaFold model of sm-PGTs from *Litoreibacter albidus* (Uniprot: A0A1H2W6Q3) and *Oceanicola sp*. (Uniprot: A0A254R773) Areas with low sequence coverage correspond to structural deviations from the core PGT fold.

AF model Rank	AF model #	RMSD Chain A	RMSD Chain B
1	4	1.48	1.64
2	2	1.09	1.14
3	5	1.27	1.39
4	3	1.51	1.67
5	1	0.59	0.76

Table S1: Summary of superposition of *C. concisus* PglC models with PDB structure.

	C. concisus PglC with C-terminus	*C. concisus* PglC Original model
Data Collection		
Resolution range	61.32 - 2.74 (2.838 - 2.74)	
Space group	P 3_2 2 1	
Cell Dimensions		
a, b, c *(Å)*	70.802 70.802 188.442	
α, β, γ (°)	90, 90, 120	
Unique reflections	27735 (1941)	
Multiplicity	18.0 (11.8)	
Completeness (%)	0.99 (1.0)	
Mean I/σ(I)	21.4 (2.5)	
Wilson B-Factor	68.21	
R_{merge}	0.0987 (1.1)	
R_{meas}	0.101 (1.16)	
$CC_{1/2}$	1.0 (0.71)	
CC^{*}	1.0 (0.89)	
Refinement		
R-work	0.258	0.259
R-free	0.295	0.281
Number of Atoms		
macromolecules	3145	3043
ligands	58	82
solvent	20	20
Protein residues	386	366
R.m.s deviations		
Bond lengths (Å)	0.003	0.003
Angles (°)	0.63	0.66
Ramachandran favored (%)	95.5	96.17
Ramachandran allowed (%)	3.97	3.28
Ramachandran outliers (%)	0.53	0.55
B-Factor		
Protein	72.4	79.5
Ligand/ion	84.9	93.4
Water	60.4	64.3

Table S2: Crystallography Data Collection and Refinement statistics. Left: New PglC model with additional C-terminal residues. Right: Original PglC model.

Aromatic Box Motif				
***C. concisus* PglC residue #**				
118	122	150	154	197
diNAcBac Cluster Consensus				
Y	Y	W	F	F
diNAcBac Cluster Conservation (%)				
89.5	87.9	91.5	74.1	67.2
non-diNAcBac Cluster Consensus				
E	E	L	V	n/a
non-diNAcBac Cluster Conservation (%)				
16.2	11.1	5.4	35.2	n/a

GLLLP Motif				
***C. concisus* PglC residue #**				
71	72	73	74	75
diNAcBac Cluster Consensus				
G	L	L	L	P
diNAcBac Cluster Conservation (%)				
73.3	59.7	2.3	74.1	54.9
non-diNAcBac Cluster Consensus				
R	N	G	A	E
non-diNAcBac Cluster Conservation (%)				
0.8	21.1	29.5	11.5	<0.1

Table S3*:* Conservation of Motifs in sm-PGTs. Residue numbers indicate position in *C. concisus* PglC. Cluster consensus displays the most common residues at these positions within the two clusters extracted from the alignment tree. Conservation displays the percentage of sequences containing the consensus residue. n/a: Position is after the C-terminus of the non-diNAcBac PGTs.