Diversity of sugar-diphospholipid-utilizing glycosyltransferase families

Ida K.S. Meitil[1], Garry P. Gippert[1], Kristian Barrett[1], Cameron J. Hunt[1], Bernard Henrissat[1,2,3*]

January 23, 2024

[1]Department of Biotechnology and Biomedicine, Technical University of Denmark, Kgs. Lyngby, Denmark
[2]Architecture et Fonction des Macromolécules Biologiques, CNRS, Aix-Marseille Université, Marseille, France
[3]Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

*Corresponding author. E-mail: bernard.henrissat@gmail.com

Abstract

Peptidoglycan polymerases, enterobacterial common antigen polymerases, O-antigen ligases, and other bacterial polysaccharide polymerases (BP-Pol) are glycosyltransferases (GT) that build bacterial surface polysaccharides. These integral membrane enzymes share the particularity of using diphospholipid-activated sugars and were previously missing in the carbohydrate-active enzymes database (CAZy; www.cazy.org). While the first three classes formed well-defined families of similar proteins, the sequences of BP-Pols were so diverse that a single family could not be built. To address this, we developed a new clustering method ~~where a~~ using a combination of a sequence similarity network ~~was used to define small groups of alignable sequences, hidden Markov models (HMMs) were built for each group, and the resulting HMMs were aligned to form new families~~and hidden Markov model comparisons. Overall, we have defined 17 new GT families including 14 of BP-Pols. We find that the reaction stereochemistry appears to be conserved in each of the defined BP-Pol families, and that the BP-Pols within the families transfer similar sugars even across Gram-negative and Gram-positive bacteria. Comparison of the new GT families reveals three clans of distantly related families, which also conserve the reaction stereochemistry.

1 Introduction

Carbohydrate polymers (glycans) and glyco-conjugates are the most abundant biomolecules on Earth and adopt a wide range of functions including energy storage, structure, signaling, and mediators of host-pathogen interactions[1]. Due to the stereochemical diversity of monosaccharides and the many possible linkages they can engage into, glycans display an enormous structural diversity[2;3]. Yet, our knowledge on their assembly is far from complete, especially in comparison to the enzymes catalyzing their breakdown.

The transfer of sugar moieties to acceptor molecules such as proteins, lipids or other sugars, is catalyzed by enzymes called glycosyltransferases or GTs[4]. Campbell and colleagues (1997) proposed a sequence-based classification of GTs into 26 families. The number of sequence-based families has since continued to grow based on the necessary presence of at least one experimentally characterized founding member to define a family, and is presented in the carbohydrate-active enzymes database (CAZy; www.cazy.org)[5]. An advantage of the sequence-based classification is that it readily enables genome mining for the presence of new family members. Today there are 116 GT families in the CAZy database and in contrast to the EC numbers[6], the sequence-based classification implicitly incorporates the structural features of GTs including the conservation of the catalytic residues.

It was recognized very early that sequence-based GT families group together enzymes that can utilize different sugar donors and/or acceptors, illustrating how GTs can evolve to adopt novel substrates and form novel products[7;8]. Mechanistically, glycosyltransferases can be either retaining or inverting, based on the relative stereochemistry of the anomeric carbon of the sugar donor and of the formed glycosidic bond[4]. With almost no exceptions, this feature is conserved in previously defined sequence-based families, providing predictive power to this classification, as the orientation of the glycosidic bond can be predicted even if the precise transferred carbohydrate is not known.

The large majority of the 116 GT CAZy families use donors activated by nucleotide diphosphates. Eleven families utilize nucleotide monophospho-sugars (sialyl and KDO transferases), while 12 families utilize lipid monophospho-sugars. Until now, only one family in the CAZy database utilizes sugar-diphospholipid donors: the oligosaccharyltransferases of family GT66, which transfer a pre-assembled oligosaccharide to Asp residues for protein N-glycosylation[4;9]. Several sugar-diphospholipid-utilizing GTs are currently missing in the CAZy database, and here we classify new sugar-diphospholipid-utilizing GTs from four major functional classes that are all involved in the synthesis of bacterial cell wall polysaccharides.

The first of these four functional classes corresponds to the peptidoglycan polymerases, SEDS (shape, elongation, division and sporulation) proteins. These proteins polymerize peptidoglycan in complex with class B penicillin-binding proteins[10]. Several 3-D structures of SEDS proteins have been determined, and they harbor 10 transmembrane helices and one long extracellular loop[11;12;13]. This loop contains an Asp residue, which has been shown to be essential for SEDS function[11;14].

The enzymes in the next two functional classes, bacterial polysaccharide polymerases (BP-Pol, also known as Wzy) and O-antigen ligase (O-Lig, also known as WaaL) are involved in the synthesis of lipopolysaccharides (LPS). LPS are polysaccharides on the membrane of Gram-negative bacteria, and consist of the highly diverse O-antigen attached to the Lipid A-core oligosaccharide located in the outer membrane[15]. The structure of the O-antigen determines the O-serotype of the bacteria. Most LPS structures are produced via the so-called Wzx/Wzy-dependent pathway[16;17], for which the genes are located in a specific gene cluster[16]. In this pathway, BP-Pol catalyzes the polymerization of pre-assembled oligosaccharides attached to undecaprenyl pyrophosphate (Und-PP). Little is known about the activity of BP-Pols. Firstly, because they are difficult to express heterologously, and to date, only one study has demonstrated the activity of O-Pol *in vitro*[18] and no experimentally determined 3-D structure is available. Secondly, because the sequences of BP-Pols are highly diverse with a sequence identity as low as 16% for different serotypes of the same species[16], it is difficult to identify conserved residues. However, several studies have identified BP-Pols in the gene clusters of various species, paving the way for analyzing BP-Pol sequences across a large range of taxonomic origin (see below). These include some Gram-negative bacteria which also employ the Wzx/Wzy-dependent pathway to produce capsular polysaccharides, including *Streptococcus pneumoniae*[19]. The third functional class, O-Lig catalyzes the final step in the synthesis of LPS; the ligation of the newly synthesized polymer (O-antigen) onto Lipid A-core oligosaccharide[20]. A structure of O-Lig in complex with Und-PP has been reported, which showed a fold with 12 transmembrane helices and a long periplasmic loop containing several conserved residues; two Args which bind to the phosphates of Und-PP and a His which is proposed to activate the acceptor[21].

The enzymes present in the fourth functional class, the enterobacterial common antigen polymerases (ECA-Pol, also known as WzyE) are involved in the synthesis of enterobacterial common antigen (ECA). In addition to the O-antigen, ECA is a specific polysaccharide that occurs on the cell surface in members of the Enterobacterales order. ECA consists of repeating units of N-acetylglucosamine, N-acetyl-D-mannosaminuronic acid and 4-acetamido-4,6-dideoxy-D-galactose[22]. ECA is also produced via the Wzy/Wzx-dependent pathway, where ECA-Pol performs the equivalent reaction to the BP-Pols [22].

Structurally, the sugar-diphospholipid-utilizing GTs have an overall GT-C fold common to other integral membrane GTs, which is different from the globular nucleotide-sugar-utilizing GTs; GT-A and GT-B[4]. GT-C enzymes have a number of transmembrane helices that varies from 8 to 14[4;23]. Alexander and Locher recently suggested two subgroups of GT-C glycosyltransferases, GT-C_A and GT-C_B[23], where O-Lig and SEDS make up GT-C_B[23]. As no structures have been published of ECA-Pol and BP-Pols, these have not been assigned to a structural subgroup.

We have identified 17 new GT families covering a large number of the sugar-diphospholipid-utilizing GTs, by detailed analysis of the primary sequence of SEDS proteins, ECA-Pols, BP-Pols and O-Ligs. In addition, we examined how sequence diversity correlates with the diversity of the transferred oligosaccharides and with the stereochemical outcome of the glycosyl transfer reaction. The analysis also revealed that the new GT families organize in three clans across the functional classes suggestive of common ancestry. Despite of poor sequence alignments we manage to identify conserved potentially critical amino acids common within the clans.

2 Results

2.1 Peptidoglycan Polymerases

For building the CAZy family of SEDS proteins, we used four characterized proteins as seed sequences: the proteins with PDB IDs 6BAR[11], 8TJ3[13] and 8BH1[12], and the protein with GenBank accession CAB15838.1[24]. Family ~~GTxx1~~ GT119 was created and initially populated by using BLAST against GenBank, and subsequently by searching against GenBank with an ~~HMM~~ hidden Markov model (HMM) built from the retrieved sequences. ~~GTxx1~~ GT119 is a very large family currently counting over 57,200 GenBank members in the CAZy database with a pairwise sequence identity of 19% over 221 residues for the most distant members.

The taxonomic distribution of family ~~GTxx1~~ GT119 follows what was reported in[14], namely that this protein family is present in all bacteria except for Mycoplasma. It is present in most but not all planctomycetes.

For SEDS proteins, the glycosyl donor for the polymerization reaction is Lipid II (Und-PP-muropeptide, an activated disaccharide carrying a pentapeptide), where the ~~undecaprenyl diphosphate~~ Und-PP is α-linked. The carbohydrate repeat unit of peptidoglycan being β-linked, the glycosyl transfer reaction thus inverts the stereochemistry of the anomeric carbon involved in the newly formed glycosidic bond.

2.2 Enterobacterial common antigen polymerases

The ECA-Pol which was studied in [25] was used as seed sequence for building the ECA-Pol family. Although the CAZy database only lists GenBank entries [26], we decided to build our multiple sequence alignments (MSAs) with sequences from the NCBI non-redundant database in order to capture more diversity. An ECA-Pol sequence library was thus constructed from the seed sequence using BLAST against the non-redundant database of the NCBI. The ECA-Pols were assigned to a single new CAZy family, ~~GTxx2~~GT120. To date this new family contains over 4800 GenBank members with high similarity (sequence identity greater than 38% over 414 residues), consistent with the conservation of acceptor, donor and product of the reaction.

As expected from their taxonomy-based designation, the ECA-Pol family (~~GTxx2~~GT120) essentially contains sequences from the Enterobacterales order but also a few members of the Pasteurellales, suggesting that ECA-Pols of the latter were acquired by horizontal gene transfer.

The ECA-Pol family uses a retaining mechanism, since the substrate repeat unit is axially linked to Und-PP and also axially linked in the final polymer.

2.3 O-antigen ligases

With the aim of including the O-Ligs in the CAZy database, we collected 37 O-Lig sequences (Supplementary Table 1) and constructed a sequence library from these seed sequences using BLAST against the NCBI non-redundant database. A phylogenetic tree of the sequence library revealed four distantly related clades (Supplementary ~~Fig.~~ Figure 1). The O-Ligs were included into one new CAZy family, ~~GTxx3~~ GT121 with more than 16,700 members distributed in the four subfamilies.

The greater diversity of the ~~GTxx3~~ GT121 O-Ligs compared to the ~~GTxx1~~ GT119 peptidoglycan polymerases and ~~GTxx2~~ GT120 ECA-Pol appears in the form of the four divergent clades in the O-Lig phylogenetic tree (Supplementary ~~Fig.~~ Figure 1). We hypothesize that this increased diversity originates from the extensive donor and moderate acceptor variability of O-Ligs [15]. Taxonomically, the ~~GTxx3~~ GT121 O-Lig family is present in most bacteria, including both ~~Gram-negatives~~ Gram-negative and Gram-positive bacteria. The reaction performed by O-Ligs involves an inversion of the stereochemistry of the anomeric carbon since the sugar donor is axially bound to Und-PP and the reaction product is equatorially bound to Lipid A [20].

A recently discovered O-Lig, WadA, is bimodular with a ~~GTxx3~~ GT121 domain appended to a globular glycosyltransferase domain of family GT25, which adds the last sugar to the oligosaccharide core [27]. We have constructed a tree with representative WadA homologs from the ~~GTxx3~~ GT121 family (Supplementary ~~Fig.~~ Figure 2) and observe that most of the sequences appended to a GT25 domain form one clade in the tree, except for a few outliers. This suggests a coupled action of the GT25 and of the ~~GTxx3~~ GT121 at least for the bimodular O-ligs and possibly for the entire family. The bimodular WadA O-Lig is observed in five genera including Mesorhizobium and Brucella.

2.4 Other bacterial polysaccharide polymerases

The fourth functional ~~subgroup of GT-C$_B$~~ class of Und-PP-sugar-utilizing GTs are the BP-Pols. As previously mentioned, there is only one experimentally characterized BP-Pol [18], but several studies have identified BP-Pols from the polysaccharide gene clusters, and we decided to build our families based on these published reports. We thus collected 363 predicted BP-Pol sequences from seven studies for various species, both ~~Gram-negatives and Gram-positives~~Gram-negative and Gram-positive bacteria: *Escherichia coli* [28], *Shigella boydii*, *Shigella dysenteriae*, *Shigella flexneri* [29], *Salmonella enterica* [30], *Yersinia pseudotuberculosis*, *Yersinia similis* [31], *Pseudomonas aeruginosa* [16], *Acinetobacter baumanii*, *Acinetobacter nosocomialis* [32] and *Streptococcus pneumoniae* [19] (Supplementary Table 2).

In contrast to ECA-Pols, the donors as well as the acceptors of BP-Pols are highly variable. Others have reported an exceptional sequence diversity of BP-Pols even within the same species [16]. We also found that the sequences of BP-Pols are extremely diverse, and global alignments failed to reveal any conserved residue due to both sequence diversity and to the difficulty in aligning proteins with multiple and variable numbers of transmembrane helices. It was therefore not possible to build a single family that could capture the diversity of BP-Pols.

In order to group BP-Pols into similarity clusters that we could include as families in the CAZy database, we first built a sequence library by running BLAST against the NCBI non-redundant database for each of the ~~365~~ 363 BP-Pol seeds. Clustering of the BP-Pols proved challenging. A phylogenetic analysis was not possible because of their great diversity, and a sequence similarity network (SSN) analysis alone would either result in very small clusters (using a strict threshold) or larger clusters that were linked because of insignificant relatedness (using a loose threshold).

Instead, we used a combination of SSN and HMM comparisons: First, we used an SSN with a strict threshold which would allow us to build good MSAs for the resulting clusters. This resulted in 204 clusters (~~Fig.~~ Figure 1a).

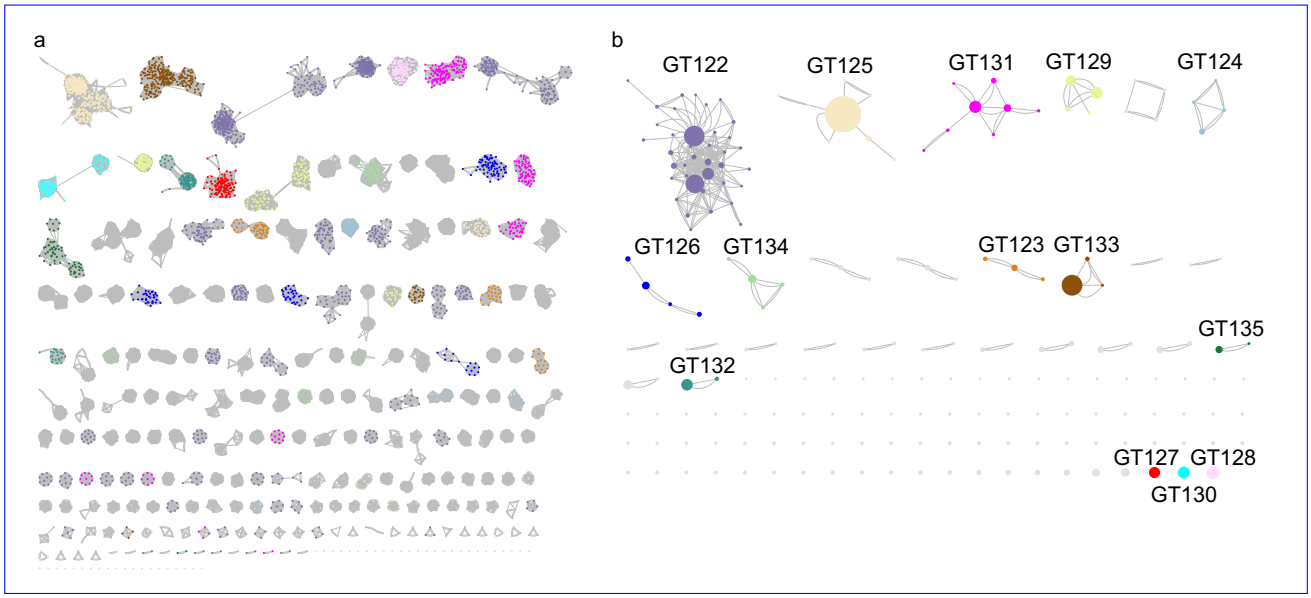

Figure 1: Clustering of BP-Pol sequences. a) The first step of the clustering: SSN network with nodes representing individual proteins and edges representing pairwise alignment bit scores. Proteins are linked by edges if they have a pairwise score above 110. The resulting clusters are sorted according to number of protein members, with the largest cluster in the upper left corner. b) The second step of the clustering: HMM models were built for each SSN cluster and the HMMs were compared using HHblits. A network was built with nodes ~~presenting~~ representing SSN clusters and edges representing HHblits scores. SSN clusters are linked by edges if they have an HHblits score higher than 160. The resulting clusters are referred to as superclusters and are sorted according to number of SSN clusters. There are two edges between nodes, when the HHblits score is above ~~the threshold~~ 160 in both directions. The size of the nodes represents the number of members in the SSN cluster. The 14 largest superclusters (>150 GenBank members) ~~defined~~ define CAZy families ~~GTxx4~~ GT122 - ~~GTx17~~GT135. ~~In~~ Nodes are colored consistently according to their respective CAZy family in both panel a and b~~, the SSN clusters are coloured according to which supercluster they belong to~~.

Next, we created an HMM profile of each SSN cluster and compared the HMMs by all-vs-all pairwise HHblits, a program that aligns two HMMs and calculates a similarity score[33]. We then combined the SSN clusters into superclusters in a network analysis based on the HHblits scores (~~Fig.~~ Figure 1b), resulting in ~~28~~ 27 superclusters of varying sizes and 86 singleton clusters. Interestingly, the BP-Pols clustered across taxonomy, and even BP-Pols from Gram-positive and Gram-negative bacteria clustered together. The 14 largest superclusters define new GT families in the CAZy database (~~GTxx4-GTx17~~GT122-GT135) with a number of members ranging from 159 to 5,979 at the time of submission. Only 150 of the 363 original seeds are included in the new families. We thus expect that many more BP-Pol families will be created in the future, as the amount and diversity of data increase.

All of the BP-Pol families are present in a wide taxonomic range, and outside of the taxonomic orders of the original seeds. Several of the families contain members from both Gram-positive and Gram-negative bacteria, for example ~~GTxx4, GTx12, and GTx16~~GT122, GT130, and GT134.

As a way of evaluating our families, we performed structural superimpositions of AlphaFold models of distantly related members of each family. As an example, superimpositions of five distantly related members of ~~GTxx4~~ GT122 are shown in Supplementary ~~Fig. 4.~~ Figure 3. The sequence identity between these members is relatively low (between 21.4 and 24.3%). Yet, they still produce a meaningful superimposition, and notably, the conserved residues are oriented very similarly.

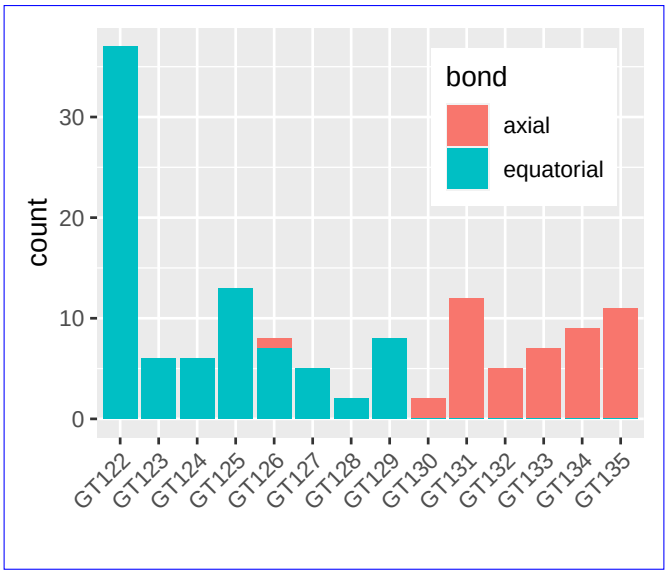

Figure 2: ~~Conservation~~ Level of conservation of stereochemical outcome of the reaction catalyzed in the various BP-Pol families. ~~Equatorial bonds~~ The bars represent the number of enzymes that are ~~shown in orange, implying~~ known to employ either retaining (making an axial bond) or inverting ~~mechanism. Axial bonds are shown~~ (making an equatorial bond) mechanisms in ~~blue, implying a retaining mechanism~~each of the new BP-Pol families.

2.5 Analyzing the sugars transferred by bacterial polysaccharide polymerases

Next, we investigated how the BP-Pol families relate to the structures of the transferred oligosaccharide repeat units. We retrieved the serotype-specific sugar structures, which were reported in the review papers[34;29;30;31;16;32;19]. Additionally, nine sugar structures were included, which were published after the review papers[35;36;37;38;39]. Out of the 150 BP-Pol seed sequences that were included in the new CAZy families, we matched 131 with a sugar structure. The repeat units are oligosaccharides with 3-7 monomers within the backbone, often with branches. In most of the cases, the bond which is formed by the polymerase has been identified in the review papers based on the other GTs in the gene cluster which assemble the repeat units.

Having retrieved the sugar structures, we first analyzed the stereochemistry of the bond catalyzed by the polymerase. As mentioned above, the stereochemical mechanism (inverting or retaining) is usually well conserved in the CAZy GT families. The repeat unit structures are always axially linked (α for D-sugars and β for L-sugars) to the Und-PP moiety before polymerization. There are two possible mechanisms for the BP-Pol-catalyzed polymerization reaction, either retaining or inverting the axial configuration. Thus, if the bond formed by the polymerase is axial, the mechanism is retaining and if the bond formed by the polymerase is equatorial, the mechanism is inverting.

We found that the stereochemical outcome of BP-Pols appears well conserved within the new BP-Pol CAZy families and varies from one family to another (~~Fig.~~ Figure 2). There is only one exception; in family

GTxx8GT126, the polymerase linkages are all equatorial except for the O-antigen in *Pseudomonas aeruginosa* O4, where it is axial. It is possible that there could This could be due to an error in the chemical structure or that the serotype designation was incorrect or that the *P. aeruginosa* O4 polymerase constitutes an exception.

Next, we investigated whether there was a correlation between the structures of the transferred sugars and the sequence similarity of the BP-Pols. We created phylogenetic trees of the BP-Pols in each family and visualized them with the corresponding transferred repeat units. We observe that the sugars within each family show similarity and this similarity appears to correlate with the structure of the tree(Fig. , implying that polymerases with similar sequence utilize similar substrates (Figure 3, Supplementary Fig. Figure 4). The ends of the repeat units, ie. the subsite moieties immediately upstream (+1) and downstream (-1) of the newly created bond (Fig. Figure 4) seem to be most conserved whereas more variability occurs in the middle part. We hypothesize that the +1 and -1 subsites are the moieties most important for recognition by the active site of the BP-Pol.

We observe examples of BP-Pols from distant taxonomic origin that cluster in the same CAZy family and have highly similar sugars. For example, *Escherichia coli* O178 and *Streptococcus pneumoniae* 47A in GTxx7 GT125 transfer sugars with almost identical backbones, suggestive of horizontal gene transfer (Figure 3). There is only a slight variance in the middle of the repeat unit. This suggests that there is less constraints on the central part of the repeat unit than on the extremeties extremities that define the donor and the acceptor.

We next attempted to quantify the correlation between BP-Pol sequence and carbohydrate structure. For this we developed an original pairwise oligosaccharide similarity score. In our scoring scheme, the similarity of two glycans is estimated by examining the -1 and +1 subsites, as we expect that these are the moities moieties most fitting the active site of the BP-Pol (Fig. Figure 4). The minimum match between two oligosaccharides corresponds to identical moieties at both subsites -1 and +1, which yields a score of 2. Thereafter, the score increases by one unit for each additional match at contiguous subsites, -2, -3, etc., and +2, +3, etc., up to a maximum value of 7 subsites found for the glycans encountered in this study (for details see Methods).

Using our glycan similarity scoring system, we found a correlation between sugar similarity and polymerase sequence similarity (Fig. Figure 5), supported by a preponderance of similarity scores appearing close to the score matrix diagonal and within each individual family.

2.6 Comparison of families

Others have previously reported sequence and structural similarity between SEDS, O-Lig and some BP-Pols [13;23;21;14] 13;14;21;23. In order to investigate the relatedness of the new CAZy families, we compared the family HMMs by all-vs-all HHblits analyses[33] (Fig. Figure 6). Strikingly, we observe that the retaining BP-Pol families cluster together on the heatmap along with the retaining ECA-Pols, while the inverting BP-Pols form two distinct groups, one of them containing the inverting SEDS (GT119) and the inverting O-Ligs (GT121). The background noise between some inverting and retaining enzymes is likely due to the general conservation of the successive transmembrane helices, which is altered in the GTxx4-GTxx5-GTxx6 GT122-GT123-GT124 subgroup due to their different architecture (see below). Peptidoglycan polymerases, GTxx1, segregate away from the other families.

In the CAZy database, clans have been defined for the glycoside hydrolases (GHs), which group together CAZy families with distant sequence similarity, similar fold, similar catalytic machinery and stereochemical outcome[40]. In extension of the report of the GT-C_B class by Alexander and Locher[23], and based on the above-mentioned similarities between the new CAZy families, we can now define three clans within GT-C_B sequence-based clans: GT-C_{B1} consisting of inverting BP-Pol families, SEDS and O-Lig, GT-C_{B2} consisting of retaining BP-Pol families and ECA-Pol, and GT-C_{B3} consisting of inverting BP-Pol families (Table 1, Figure 6). The families within each clan share residual, local, sequence similarity, insufficient to produce a multiple sequence alignment, but suggestive of common ancestry.

In the absence of a three-dimensional structure, and based solely on the number of transmembrane helices, we assigned on the sequence similarity to SEDS and O-Ligs, we have assigned the BP-Pol families of clan GT-C_{B2} and GT-C_{B3} to the structural subclass GT-C_B of Alexander and Locher[23]. In addition, we also present in Table 1 the families of GT-C glycosyltransferases that have not yet been assigned to a structural class.

We then examined residue conservation and the general architecture of the enzymes in the clans. Based on the above mentioned pairwise HHblits analyses and structural superimpositions (Supplementary Figure 5-7), we tried to evaluate which architectural features and conserved residues are common within the clans. Indeed, there are some common features across most families. In all the families, all the conserved residues are located on the outer face of the membrane (Figure 7). Enzymes of clans GT-C_{B1} and GT-C_{B2} have a long extracellular loop close to the C-terminus (Fig. containing conserved residues (Figure 7). In stark contrast, families GTxx4, GTxx5 and GTxx6 GT122, GT123 and GT124 of clan GT-C_{B3} have an architecture completely different from that of the two other clans (Fig. Figure 7), with the a long loop located close to the N-terminus, and a conservation of one Asp, one His and two Arg residues.

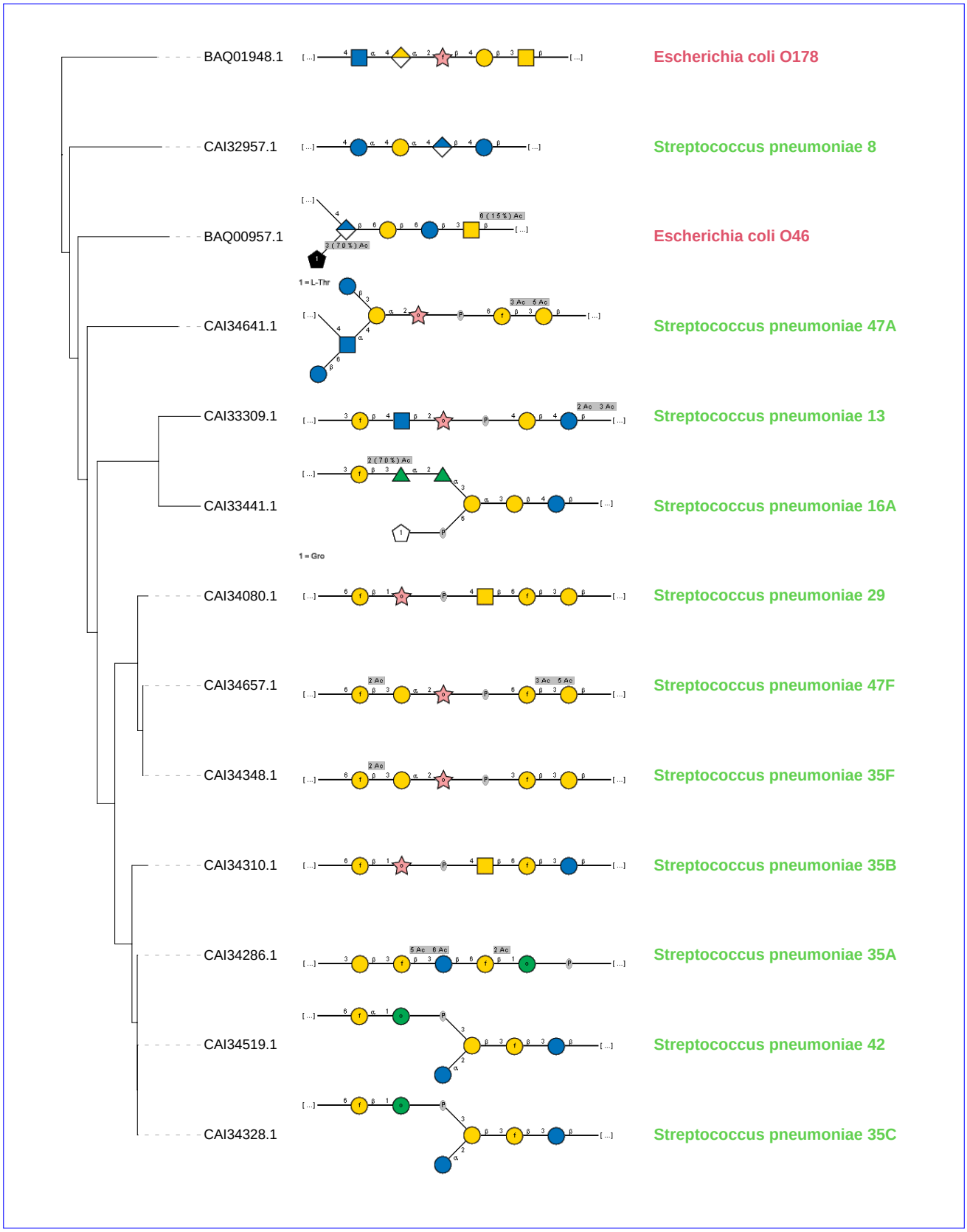

Figure 3: ~~Similarity~~ Comparison of ~~transferred~~ repeat unit sugars transferred by BP-Pols in ~~GTxx7~~GT125. The transferred repeat unit structures (in SNFG representation) are shown on a phylogenetic tree of BP-Pols in family ~~GTxx7~~GT125. There is an overall similarity between all the transferred sugars in the family and the similarity appears to correlate with the tree structure, ie. BP-Pol sequence similarity. In particular, the ends of the repeat units (+1 and -1 subsites) appear to be often conserved, whereas there is more variety in the central region where the enzyme does not interact with the sugar. Note that the +1 site corresponds to the non-reducing end of the depicted sugar structures and the -1 site corresponds to the reducing end. Notably, the family contains BP-Pols from distant ~~taxonomy which~~ taxonomic origin and that yet transfer similar ~~sugars~~repeat units.

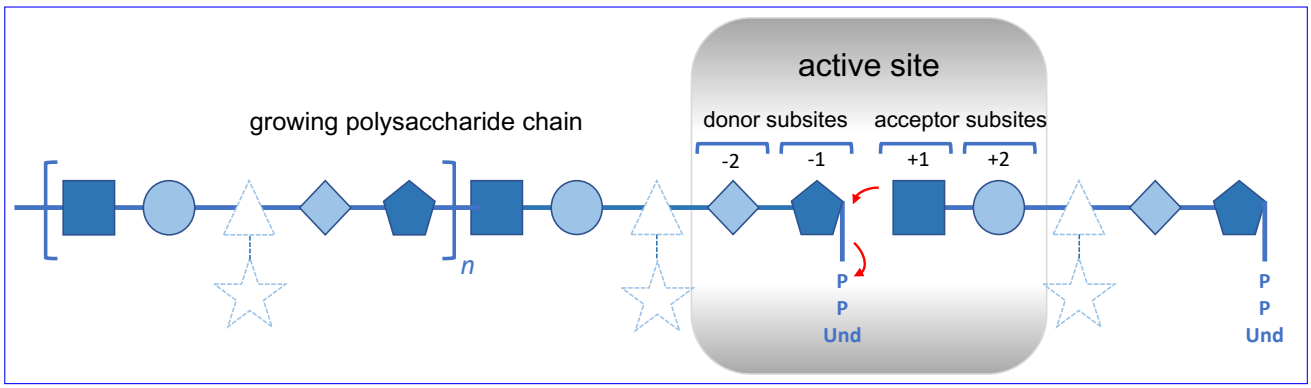

Figure 4: An idealized representation of a BP-Pol. The donor is the growing glycan chain activated by ~~undecaprenyl pyrophosphate~~ Und-PP while the acceptor is a single repeat unit ~~monomer~~linked to Und-PP. The reaction is hypothesized to chiefly involve the sugar residues of the donor (subsites -2 and -1) and of the acceptor (subsites +1 and +2) that are proximal to the reaction center rather than residues and branches that are more distal (depicted in dashed lines). The reaction is represented by red arrows.

~~Structural subclassAlexander & Locher~~ Structural subclass Alexander & Locher	~~CAZy clan~~ CAZy clan
GT-C_A	-
	-
	-
	-
	-
GT-C_B	~~- GTxx1 Inverting Lipid-PP-oligosaccharide~~ GT-C_{B1}
height-	GT-C_{B2}
height-	GT-C_{B3}
-	-
	-
	-
	-

Table 1: Structural subclasses, clans and families of GT-C fold glycosyltransferases and relationships to mechanism and glycosyl donor.

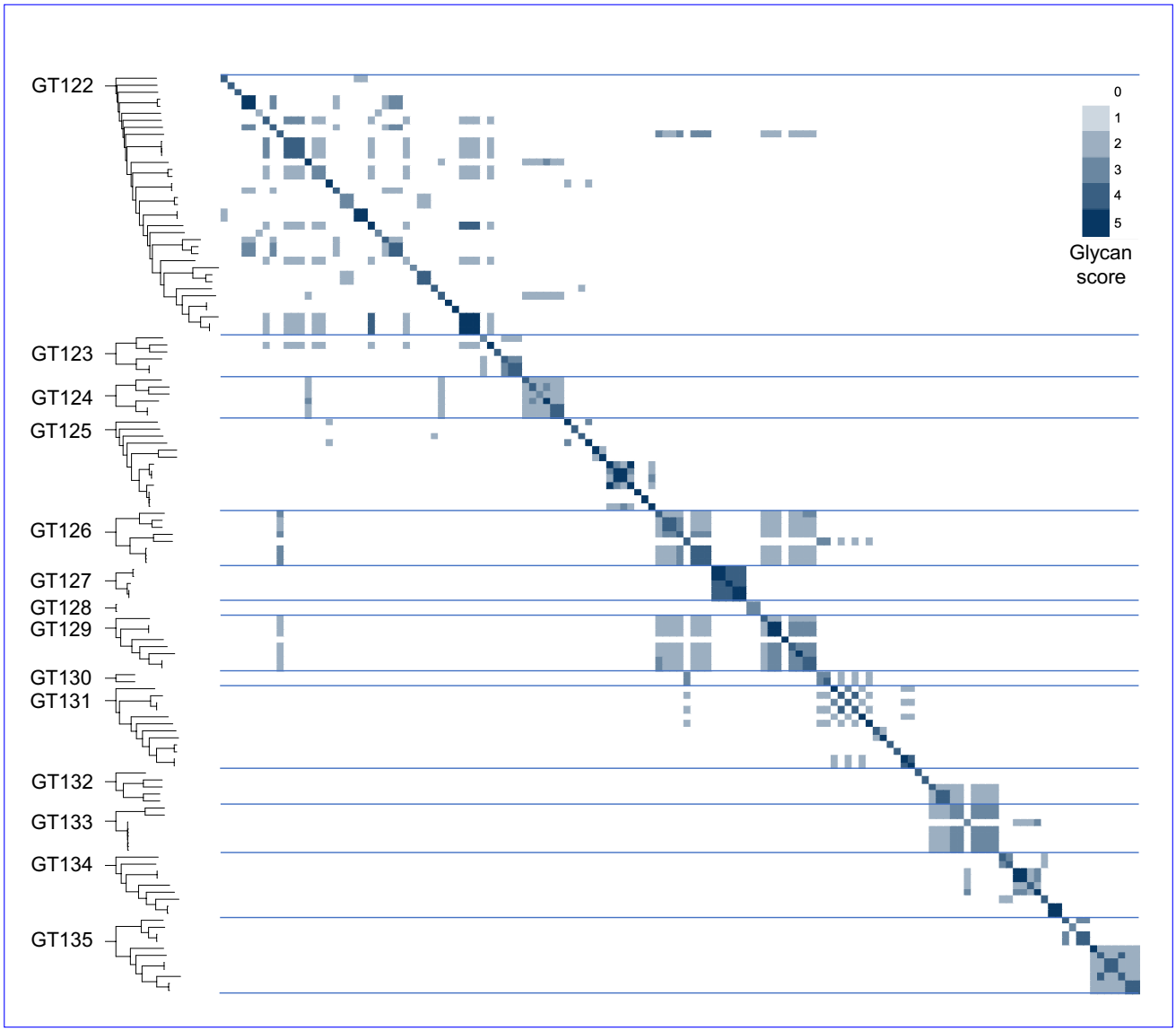

Figure 5: Glycan similarity of sugar repeat units polymerized by BP-Pols. All "seed" BP-Pols where the corresponding transferred oligosaccharide was known were included in the heatmap. A phylogenetic tree is shown for the polymerases in each CAZy family on the left. The glycan similarity scores are shown in a color scale of light blue (score value of 2 corresponding to identical matches at both –1 and +1 sites) to dark blue (score value of 5 corresponding to identical matches for at least three additional sequential positions). ~~Blue~~ Horizontal lines separate the families. The darker colors close to the diagonal and within the families indicate specific substrate similarities in each family.

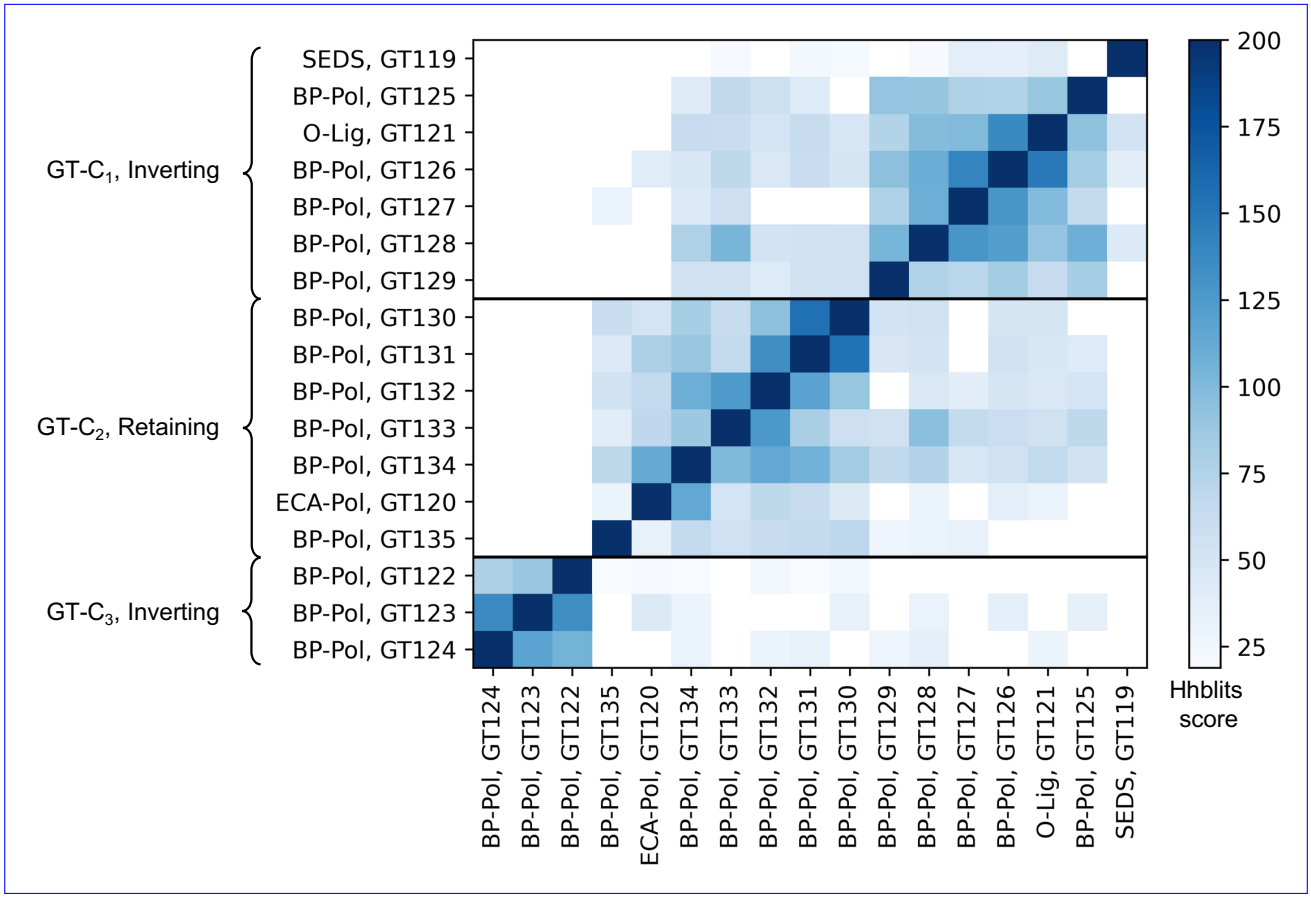

Figure 6: ~~Heatmap~~ Relatedness of ~~inter-family~~ the new CAZy families and definition of clans. Inter-family HHblits bit scores ~~. The HHblits scores~~ are shown in a heatmap on a color scale from white (low similarity score) to dark blue (high similarity score). The HHblits scores depend on the direction of the alignment, and therefore the heatmap is not symmetrical. The inverting BP-Pols form two clans, GT-C_1 which also contains the inverting SEDS (GT119) and the inverting O-Ligs (GT121) and GT-C_3 containing only BP-Pols. The retaining BP-Pol families form one clan, GT-C_2, which also contains the retaining ECA-Pol family (GT120).

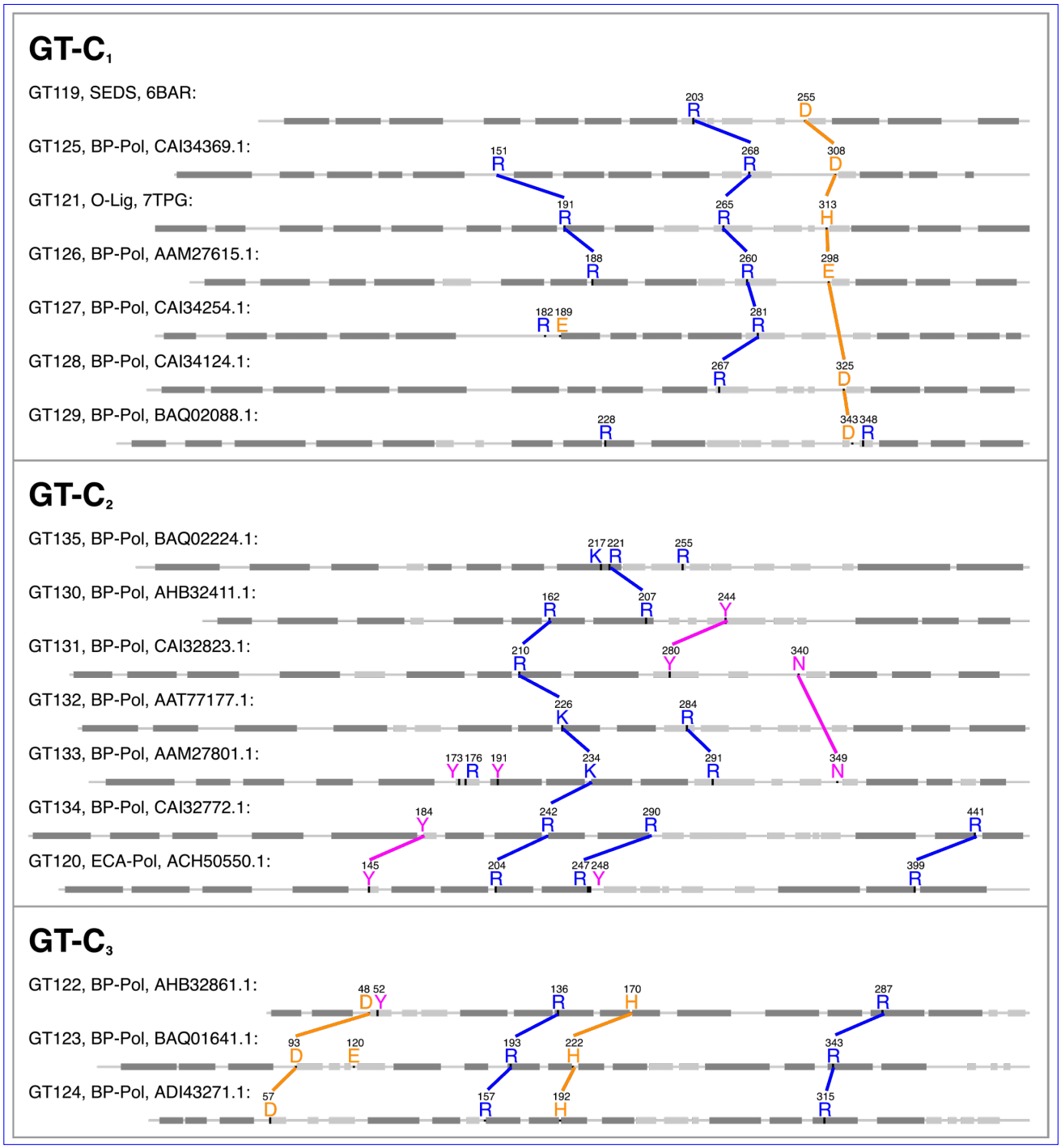

Figure 7: Equivalent conserved residues in the clans. ~~The conserved~~ Conserved residues of each of the new CAZy families are shown on sequences of representative family members. ~~Lines~~ Colored lines are shown between conserved residues ~~that~~ from different families, which align in HHblits alignments and ~~that~~ co-localize in structural superimpositions (Supplementary ~~Fig. 5~~Figure 5-7). Transmembrane helices are shown in dark gray boxes, ~~non-transmembrane~~ extracellular helices are shown in light gray boxes. The secondary structures were taken from the crystal structures for family ~~GTxx1~~ GT119 and ~~GTxx3~~ GT121 (6BAR and 7TPG respectively) and from AlphaFold models for all other families. The R210 in ~~GTx13~~ GT131 is either K or R in the family.~~Conserved aliphatic residues are not shown.~~

The families in GT-C_1 show a distinct pattern of residue conservation. As mentioned above, the structure of O-Lig in complex with Und-PP revealed several important residues; Arg-191 and Arg-265 which bind to the phosphate groups of Und-PP, and His-313 which is proposed to activate the acceptor[21]. The ~~other families in GT-C_{B1} appear to have a similar pattern. All~~ SEDS family (GT119) also has a conserved Arg which aligns with the second conserved Arg in O-Lig and a conserved essential Asp which aligns with the conserved His in O-Lig (Figure 7). Likewise, all the BP-Pols in the clan have 1-2 conserved Args, ~~most of which are conserved~~ some of which align to the O-Lig Args in the HHblits ~~alignment~~alignments, and we hypothesize that they also play the role of binding to the diphosphate. Similarly, all the families in the clan except for ~~GTxx9~~ GT127 have either a conserved Asp or Glu, which align with the ~~His-313 in~~ conserved His of O-Lig and the conserved Asp of SEDS (Figure 7). We hypothesize that ~~the~~ these Glu and Asp residues ~~in the BP-Pols play the same role as the His-313 in O-Lig~~also play the role of activating the acceptor. As an example, the superimposition of the published O-Lig structure (7TPG)[21] and an AlphaFold model from one representative of the inverting BP-Pol family ~~GTxx8~~ GT126 is shown in ~~Fig.~~ Figure 9a. The superimposition produced an overall RMSD of 5.3Å over 192 residues. Even with such a high RMSD, the two conserved Args are oriented very similarly, and the conserved His and Glu are in the same position. As mentioned above, GT127 does not have a conserved Asp or Glu in ~~O-Lig is placed in~~ the same position as the rest of the families. However, it has a conserved Glu in ~~the BP-Pol~~a loop between transmembrane helices 5 and 6, which likely plays the same role.

In the retaining clan GT-C_{B2}, the pattern of conservation is different. Here, most of the families have 2-3 conserved Arg/Lys and 1-2 conserved Tyr (Figure 7). Interestingly, we observe that the ECA-Pol family ~~GTxx2~~ GT120 shows high similarity with one of the BP-Pol families, ~~GTx16~~GT134. A superimposition of AlphaFold models from each family shows that the conserved residues are oriented very similarly, despite the low overall similarity (RMSD 5.4Å over 360 residues) (~~Fig.~~ Figure 8b).

~~Although the peptidoglycan polymerase family, GTxx1 does not cluster in any of the three clans, it does display topographical similarity to~~ The families in the inverting clan GT-C_{B3} ~~. In terms of architecture it also contains a long extracellular loop with a conserved Argand the conserved and essential Aspresidue~~[11]~~. The Asp residue is in a similar position as the Asp/Glu/Hisin the other families in clan GT-C_{B1}. We therefore hypothesize that this conserved Asp may play the role of activating the acceptor in clan GT-C_{B1} glycosyltransferases as the His in O-Lig~~[21]all have two conserved Arg, a conserved Asp, and a conserved His, all of which align between the families in the HHblits alignments (Figure 7).

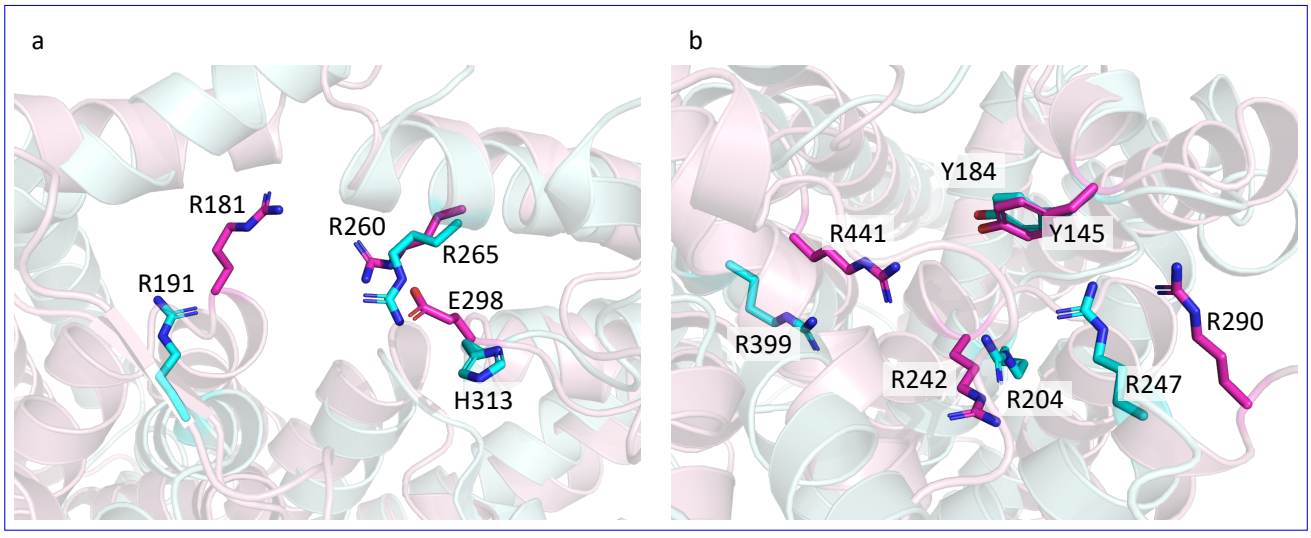

Figure 8: Structural ~~superimposition~~ superimpositions of members of different ~~families with conserved residues~~functional classes belonging to the same clans. a) Superimposition of O-Lig ~~from GTxx3~~ in cyan (GT121, PDB: 7TPG) and AlphaFold model of BP-Pol ~~from GTxx8~~ in pink (GT126, Genbank accession: AAM27615.1) (RMSD 5.3Å over 192 residues, sequence identity 20.8% over 485 residues) ~~. The~~ showing that the conserved Glu in ~~GTxx8 is aligning~~ the BP-Pol aligns with the conserved His in ~~GTxx3~~the O-Lig, which ~~is~~ has been proposed to activate the acceptor[21]. b) Structural superimpositions of AlphaFold models of ECA-Pol ~~from GTxx2~~ in cyan (GT120, Genbank accession: ACH50550.1) and BP-Pol ~~from GTx16~~ in pink (GT134, Genbank accession: CAI32772.1) illustrating structural similarity and co-localization of the conserved residues (RMSD 5.4Å over 360 residues, sequence identity 17.1% over 543 residue). ~~The conserved residues occupy similar positions.~~

3 Discussion

Here we have added 17 glycosyltransferase families (~~GTxx1 to GTx17~~GT119 to GT135) to the CAZy database bringing the total of covered families from ~~116 to 133.~~ 118 to 135. In the CAZy database, families are built by aggregating similar sequences around a biochemically characterized member. The known difficulties in the direct experimental characterization of integral membrane GTs render this constraint impractical. To circumvent this problem, but to remain connected to actual biochemistry, we decided to build our families around seed sequences for which knowledge of the glycosidic bond formed could be deduced from examination of the polysaccharide product from the literature.

To our knowledge, this is the first time that BP-Pols from different species have been successfully clustered. Indeed, forming groups of BP-Pols has been very difficult previously because of their extreme diversity even within strains of a single species[28], and, as a consequence, the knowledge on conserved and functional residues has been very limited. By combining BP-Pols from a wide range of taxonomical origins and expanding with the current sequence diversity, we were able to form larger families of similar polymerases from widely different taxonomies, thereby revealing conserved residues that are most likely functionally important.

~~We observed that the O-Lig family (GTxx3) was present in many Gram-positive bacteria such as *Streptococcus pneumoniae*. The covalent anchoring of CPS in Gram-negative bacteria is still poorly understood, although it is found to be linked to peptidoglycan in some Gram-positive bacteria [17;41]. Thus a hypothesis could be that the GTxx3 members in *S. pneumoniae* are responsible for the ligation of CPS to the peptidoglycan layer in these bacteria.~~

Because families are more robust when built with enough sequence diversity, many clusters of O-antigen polymerases were judged too small to build meaningful CAZy families. Additional polymerase families are thus expected in the future with the accumulation of sequence data. For instance the small cluster that contains 47% identical BP-Pols from *E. coli* Q108 (GenBank BAQ01516.1) and *A. baumanii* Q24 (GenBank AHB32586.1) only contains eight sequences and will remain unclassified until enough sequence diversity has accumulated. This arbitrary decision comes from the need to devise a classification that can withstand a massive increase in the number of sequences without the need to constantly revise the content of the families.

Moreover, we observe that the sequence diversity within the families we have built is minimal for peptidoglycan polymerases (~~GTxx1)~~ GT119) and ECA-Pols (GT120), and then increases gradually ~~from ECA-Pols (GTxx2) to~~ for O-Ligs (~~GTxx3~~GT121) and is maximal for BP-Pols (~~GTxx4-GTx17~~GT122-GT135). We hypothesize that sequence diversity reflects the donor and acceptor diversity in each family since the latter increases accordingly; the enzymes in the SEDS and ECA-Pol families act with the same donor and same acceptor, the enzymes in the O-Lig family act with different donors but same acceptor, and for the enzymes in the BP-Pol families act on different donors and different acceptors.

It has been observed that for classical GT-A and GT-B fold glycosyltransferases, the catalytic mechanism is conserved within a family, but families with the same fold can have different mechanisms, possibly because the stereochemical outcome of the glycosyl transfer reaction is essentially dictated by the precise positioning and activation of the acceptor above (S_N2) or below (S_Ni) the sugar ring of the donor[4]. Very occasionally, retaining glycosyltransferases have been shown to operate via a double displacement mechanism that involves Asp/Glu residues to form a glycosyl enzyme intermediate and to activate the acceptor that attacks this intermediate[42]. The families defined here display globally similar GT-C folds, and they also show conservation of the catalytic mechanism with about half of the families retaining and the other half inverting the anomeric configuration of the donor, suggesting that the outcome of the reaction catalyzed by GT-C glycosyltransferases is also dictated by the positioning of the acceptor with respect to the sugar plane of the acceptor. In turn this also suggests that retaining BP-Pols also operate by an S_Ni mechanism rather than by the formation of a glycosyl enzyme intermediate. This hypothesis is supported by the lack of invariant Asp or Glu residues which could be involved in the formation and subsequent breakdown of a glycosyl enzyme intermediate in the retaining families ~~GTxx2 and GTx12-GTx17~~GT120 and GT130-GT135. Additionally, the S_Ni mechanism may provide protection against the interception of a glycosyl enzyme intermediate by a water molecule resulting in an undesirable hydrolysis reaction and termination of the polysaccharide elongation.

The wealth of structural data of GT-C glycosyltransferases now permits a deeper evaluation of the intrinsic properties of this large class of enzymes. Alexander and Locher have recently evaluated the structural similarities between GT-C fold glycosyltransferases and have divided them in two fold subclasses[23]. The GT families that we describe here significantly expand the GT-C class in the CAZy database (www.cazy.org) and allow to combine the structural classes with mechanistic information. Lairson *et al.* have proposed the subdivision of GT-A and GT-B fold glycosyltransferases in clans that integrate the stereochemical outcome of the reaction[4]. Here we also note the conservation of the stereochemistry in the families of BP-Pols and we thus propose to group them into three clans which share the same fold, residual sequence conservation and the same catalytic mechanism (Table 1). As more families of BP-Pols emerge, these three clans will likely grow. Table 1 shows the three clans we defined here and how they relate to the structural classes defined by Alexander and Locher. Of note are families ~~GTxx4, GTxx5, and GTxx6~~ GT122, GT123, and GT124 which do not bear any similarity, even distant, with

the GT families of the other two clans. These three families also stand out by the location in the sequence of the long loop that harbors the catalytic site in the other GT-C families. In absence of relics of sequence relatedness to the other families, ~~GTxx4, GTxx5 and GTxx6~~ GT122, GT123 and GT124 were assigned to clan GT-C~~$_{B3}$~~3. ~~With 10 transmembrane helices, it is tempting to suggest that this clan may belong to the fold subclass GT-C$_B$ of Alexander and Locher.~~

The analysis presented here shows that not only the stereochemistry of the glycosyl transfer is conserved in the BP-Pol families, but our development of an original method to estimate glycan similarity also reveals a certain degree of structural similarity of the oligosaccharide repeat units, suggesting that the latter constitutes a significant evolutionary constraint applying to the sequence and structure of BP-Pols. A closer inspection of the oligosaccharide repeat units within the families further reveals that the carbohydrates that appear the most constrained are the carbohydrates located (i) at the non-reducing end of the acceptor and (ii) close to the ~~undecaprenyl pyrophosphate~~ Und-PP of the donor, i.e. the residues closest to the reaction center (~~Fig.~~ Figure 4). By contrast, residues away from the two extremities engaged in the polymerization reaction appear more variable, and can tolerate insertions/deletions or the presence of flexible residues such as linear glycerol or ribitol, with or without ~~or~~ the presence of a phosphodiester bond.

The version of the glycan similarity score presented here was inspired in part by observed structural similarities in different O-antigen repeat units assembled by very similar BP-Pols[16]. The ~~repeat-unit~~ repeat unit comparison involves a translation of glycan IUPAC nomenclature to a reduced alphabet of terms representing only backbone configuration, i.e., ignoring chemical modifications and sidechains. Furthermore, a positive similarity score requires an entire identical match of all backbone elements at both donor and acceptor positions (–1 and +1 sites in ~~Fig.~~ Figure 4, respectively). Despite these simplifications, the similarity score reveals, with exceptions, an overall greater intra- rather than inter-family oligosaccharide similarity (Fig 5). These limitations will be addressed at a later stage (G.P. Gippert, in preparation).

We have next looked at the distribution of the new GT families in genomes, and particularly the families of BP-Pols. This uncovers broadly different schemes, with some bacteria having only one polymerase (and therefore only able to produce a single polysaccharide) while others having several, and sometimes more than 5, an observation in agreement with the report that *Bacteroides fragilis* produces no less than 8 different polysaccharides from distinct genomic loci[43]. The multiplicity of polysaccharide biosynthesis loci in some genomes makes it sometimes difficult to assign a particular polysaccharide structure to a particular biosynthesis operon.

We observed that the O-Lig family (GT121) was present in many Gram-positive bacteria such as *Streptococcus pneumoniae*. The covalent anchoring of CPS in Gram-negative bacteria is still poorly understood, although it is found to be linked to peptidoglycan in some Gram-positive bacteria [17;41]. Thus a hypothesis could be that the GT121 members in *S. pneumoniae* are responsible for the ligation of CPS to the peptidoglycan layer in these bacteria.

As already shown in other occasions, the sequence-based classification of carbohydrate-active enzymes of the CAZy database has predictive power. The case of the GT families described here supports this view as the invariant residues in the families not only co-localize in the same area of the three-dimensional structures (whether actual or AlphaFold-predicted), but also correspond to the residues found essential for function in the families where this has been studied experimentally. The families described herein also show mechanistic conservation and thus the stereochemistry of glycosyl transfer can be predicted. Finally, the observed similarity in oligosaccharide repeat units that accompanies sequence similarity has also predictive power and paves the way to the future possibility of *in silico* serotyping based on DNA sequence.

4 Methods

4.1 Alignment-based Clustering (Aclust)

Phylogenetic trees were generated using an in-house tool called Aclust (G.P.Gippert, manuscript in preparation. ~~Source code may be obtained via GitHub at https://github.com/GarryGippert/Aclust). Aclust~~ Aclust employs a hierarchical clustering algorithm comprising the following steps. (1) A distance matrix is computed from all-vs-all pairwise local ~~pairwise~~ sequence alignments[44], or from a multiple sequence alignment provided by MAFFT[45]. The distance calculation is based on a variation of Scoredist ~~?~~ [46] where distance values are normalized to the shorter pairwise sequence length rather than to pairwise alignment length. (2) The distance matrix is embedded into orthogonal coordinates using metric matrix distance geometry[47], and (3) a bifurcating tree is computed using nearest-neighbor joining and centroid averaging in the orthogonal coordinate space. The last centroid created in this process is defined as the root node. (4) Beginning with the root node of the initial tree, each left and right subtree constitutes disjoint subsets of the original sequence pool, which are reembedded and rejoined separately (i.e., steps 2 and 3 repeated for each subset), and the process repeated recursively — having the effect of gradually reducing deleterious effects on tree topology arising from long distances between unrelated

proteins.

4.2 Building the peptidoglycan polymerase family (~~GTxx1~~GT119)

The peptidoglycan polymerase family, ~~GTxx1~~GT119, was built by using Blastp from BLAST+ 2.12.0+[48] with the sequences of the characterized SEDS proteins (PDB 6BAR, 8TJ3~~and~~, 8BH1 and GenBank accession CAB15838.1) against GenBank with a threshold of approximately 30% to retrieve the family members. Next, an MSA was generated with MAFFT v7.508 using the L-INS-i strategy[45], and an HMM model was built with hmmbuild of HMMER 3.3.2[49]. The family was further populated using hmmsearch from HMMER 3.2.2 against GenBank.

4.3 Building the Enterobacterial common antigen polymerases family (~~GTxx2~~GT120)

A sequence library of ECA-Pols was constructed by using Blastp with the seed sequence (GenBank accession AAC76800.1) against the NCBI non-redundant database version 61 with an E-value threshold of 1e-60. The hits were redundancy reduced using CD-HIT 4.8.1[50] with a threshold of 99%. The redundancy-reduced pool of ECA-Pol sequences was clustered using our in-house tool Aclust (see above), and the tree showed one large clade and a few outliers. All the sequences in the large clade were used to build an MSA using MAFFT v7.508 with the L-INS-i strategy[45]. An HMM was built based on this MSA using hmmbuild of HMMER 3.3.2[49]. The family ~~GTxx3~~ GT121 was built in CAZy and populated using Blastp against GenBank with an approximate threshold of 30% and hmmsearch against GenBank.

4.4 Building the O-antigen ligase family (~~GTxx3~~GT121)

37 O-Lig sequences were selected from literature (Supplementary Table 1) and expanded using Blastp against the NCBI non-redundant database with an E-value cut-off of 1e-60. Redundancy reduction was performed on the resulting sequence pool using CD-HIT with a threshold of 99%, resulting in a pool of 1,402 sequences. A phylogenetic tree of the pool of O-Lig sequences was generated using Aclust (see above), which showed deep clefts between main branches, and branches with sufficient internal diversity (Supplementary Figure 2). Based on these results, four subfamilies were determined. An MSA was built for the family as well as for the subfamilies with MAFFT v7.508 using the L-INS-i strategy. HMMs were built based on the MSAs using the hmmbuild of HMMER 3.3.2[49]. The family was populated using Blastp against GenBank ~~using~~ with an approximate threshold of 30% identity with the seed sequences and using hmmsearch with the family and subfamily HMMs.

4.5 Building the Bacterial polysaccharide polymerase families (~~GTxx4-GTx17~~GT122-GT135

363 BP-Pol sequences were retrieved from review papers on biosynthesis of O-antigens and capsular polysaccharides in different species: *Escherichia coli*[28], *Shigella boydii*, *Shigella dysenteriae*, *Shigella flexneri*[29], *Salmonella enterica*[30], *Yersinia pseudotuberculosis*, *Yersinia similis*[31], *Pseudomonas aeruginosa*[16], *Acinetobacter baumanii*, *Acinetobacter nosocomialis*[32] and *Streptococcus pneumoniae*[19] (complete list in Supplementary Table 2). The BP-Pols for *A. baumannii* O7 and O16 were omitted, because of uncertainty of their serotypes[32]. The BP-Pol from *P. aeruginosa* O15 was also omitted, because it has been shown that this BP-Pol is inactivated and that the O-antigen is synthesized via the ABC-dependent pathway rather than the Wzx/Wzy-dependent pathway[51].

The sequence library was expanded using Blastp for each seed sequence against the NCBI non-redundant database with an E-value threshold of 1e-15. Redundancy reduction was performed using CD-HIT with a threshold of 95% identity.

To find clusters of BP-Pol sequences that were large enough to create a CAZy family, we developed a clustering method consisting of two steps. First, in order to make a sequence similarity network (SSN), all-vs-all pairwise local ~~aignments~~ alignments of the BP-Pol sequence pool were performed using Blastp from BLAST+ 2.12.0+. A series of networks were built using different bit score thresholds. The members of the resulting SSN clusters were identified using NetworkX[52] and MSAs of the members were built with MAFFT v7.508 using the L-INS-i strategy. The MSAs were inspected using Jalview[53], and a bit score threshold of 110 was selected, as it was the lowest score for which the SSN clusters had adequate sequence conservation (approximately 15 conserved residues).

HMMs were then built for each SSN cluster using hmmbuild of HMMER 3.3.2, and the HMMs were compared using HHblits 3.3.0[54]. A series of HHblits networks were built using different HHblits score thresholds. Again, the members of the resulting superclusters were identified using NetworkX and MSAs of the ~~members~~ superclusters were built with MAFFT v7.508 using the L-INS-i strategy. A bit score threshold of 160 was selected as it resulted in superclusters with adequate diversity for building CAZy families (approximately 5 conserved residues). CAZy families were created for the 14 largest superclusters and populated with sequences

present in GenBank by a combination of Blastp with the seed sequences and hmmsearch. The networks were visualized with Cytoscape[55].

4.6 Analysis of sugar repeat unit structures

In order to analyze the relation between BP-Pol sequence and structure of the transferred repeat unit, we retrieved the repeat unit structures for the serotypes for the BP-Pols that were included in the new CAZy families. The repeat unit structures were retrieved from the same review papers from which we ~~got~~ retrieved the BP-Pol sequences[32;19;31;30;29;16], except for the sugars for *E. coli*, where the sugar structures have been reported elsewhere[34]. Nine additional repeat unit structures were included for *S. pneumoniae*, which were published after the review paper; serotypes 16A[35], 33A[36], 33C and 33D[37], 35C and 35F[38], 42 and 47F[56] and 47A[57]. For *Y. pseudotuberculosis* O3 and *S. pneumoniae* 33B, we used the revised ~~structure from [39] and [37] respectively~~structures [39;37]. *Pseudomonas aeruginosa* O2 and O16 contain two BP-Pol genes; one BP-Pol localized in the O-antigen biosynthesis cluster, which polymerizes the sugar repeat units with an α bond and one BP-Pol localized outside the biosynthesis cluster which polymerizes the repeat units with a β bond[58]. Since the ~~BP-Pol~~ BP-Pols reported in[16] are ~~the BP-Pols~~ from the O-antigen cluster, we report the sugar structure with the α bond.

The linkages formed by the polymerase ~~has~~ have been determined in all of these papers, except for a few cases. This determination is based on the ~~initial GT transfering specific monosaccharides, and sometimes also based on~~ other GTs in the gene cluster, in particular the initial GT which transfers the first monosaccharides to the Und-PP anchor. The cases where the polymerase linkage has not been ~~determined unambigously~~ unambiguously determined in the review papers are *E. coli* O166, O78, O152, O81, O83, O11, O112ab, O167, O187, O142, O117, O107, O185, O42, O28ac, O28ab, for which there ~~were~~ are two or more possible polymerase linkages. For the structures that were published after the review papers, the polymerase bond had not been determined in *S. pneumoniae* 33A and 47A. For *S. pneumoniae* 33A, we determined the ~~polymerase~~ linkage based on the ~~gene cluster having~~ presence of the initial trasferase ~~WchA~~ *wchA* in the gene cluster, which transfers a ~~glucose~~ [32]. glucose-1-phosphate to Und-PP [19]. In *S. pneumoniae* 47A ~~has WcjG as~~ the initial transferase is WcjG, which transfers Gal*p* or Gal*f* [32][19]. Since the repeat unit contains both Gal and Galp, we could not determine the polymerase linkage unambiguously. However, the repeat unit is very similar to other repeat units in the family (most similar to that of *S. pneumoniae* 13), and we proposed the equivalent polymerase linkage.

The CSDB database (http://csdb.glycoscience.ru)[59] was used to retrieve literature, SNFG image representations and linear sugar strings of the repeat unit structures. Phylogenetic trees for BP-Pol families with sugar structures were generated using MAFFT v7.508[45] with the L-INS-i strategy to supply an initial multiple sequence alignment, followed by Aclust (section 4.1) for distance matrix embedding and clustering. The trees were visualized in iTOL[60]. The barplot was generated using R [61], Rstudio [62], and the ggplot2 package [63].

4.7 Oligosaccharide backbone similarity score

A similarity score function was developed that quantifies the number of identical subunits at both donor and acceptor ends of oligosaccharides, specifically positions [..., –2, –1, +1, +2, ...] with respect to the bond formation site (Figure 4). The minimum non-zero similarity score between a pair of oligosaccharides is 2, requiring identity at both positions –1 and +1. Thereafter the comparison extends by one position in each positive (+2, +3, ...) and negative (–2, –3, ...) chain ~~directions~~direction, adding one to the score for each additional identical match, but terminating at the first non-identity or possible re-use of a backbone position.

To facilitate comparison, ~~oligosaccharides~~ oligosaccharide sequences are translated from IUPAC nomenclature into symbols that represent elements of backbone geometry, only considering monomer dimension and stereochemistry of acceptor and anomeric donor carbon atoms, and ignoring sidechains and chemical modification (~~Fig.~~ Figure 9). Briefly, the monomer dimension is represented by a single letter P, F or L depending on whether the monomer sugar is a pyranose, furanose or is linear, respectively. Stereochemistry of the acceptor and donor carbon atoms is represented by the index number of the carbon position within the ring/monomer, followed by a single letter U, D or N depending on whether the linked oxygen atom is U (up=above the monomer ring), D (down=below the monomer ring), or N (neither above or below the ring). The N symbol is assigned in cases of conformational flexibility such as with alditols or C6 linkages. At present, in scoring the similarity of two thus translated residues, the entirety of the translation strings must be identical to achieve a score of +1. Further details and limitations will be presented elsewhere (G.P. Gippert, manuscript in preparation).

4.8 Comparison of the families

Pairwise HHblits analyses[33] were performed for each of the new CAZy families. The HHblits scores were visualized in a heatmap using Python Matplotlib[64].

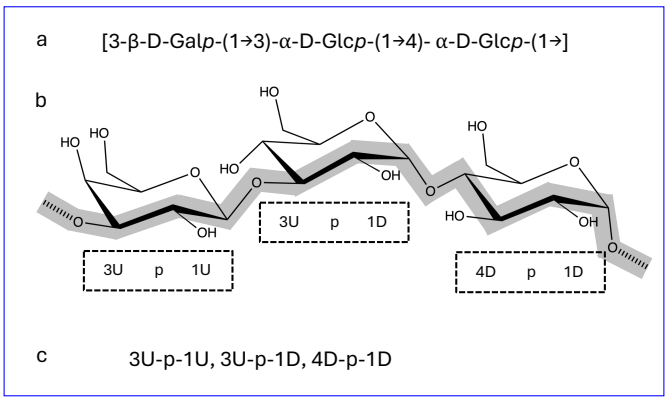

Figure 9: Oligosaccharide translation from IUPAC nomenclature to backbone (geometric) subunits for a trisaccharide consisting of one D-galactopyranose and two D-glucopyranose residues joined by intramolecular β1→3 and α1→4 bonds, respectively, and an intermolecular α1→3 bond formed in the polymerase reaction. (a) IUPAC nomenclature (b) Stereochemical projection highlighting backbone (thick grey line) and transfer bond (hatched line segments), and translated geometric subunits below ~~(see text)~~. (c) Completed translation.

AlphaFold2[14] structures were generated of representative proteins from the families using the ColabFold implementation[65] on our internal GPU cluster processed with the recommended settings. The best ranked relaxed model was used. The protein structures were visualized in PyMOL [?][66] and pairwise structural superimpositions were performed using the CEalign algorithm[67].

5 Data availability

Accessions to the seed sequences utilized in this work are given in Supplementary Table 1-2; the constantly updated content of families ~~GTxx1~~ GT119 - ~~GTx17~~ GT135 is given in the online CAZy database at www.cazy.org.

6 ~~Acknowledgements~~Code availability

Source code for Aclust may be obtained via GitHub at https://github.com/GarryGippert/Aclust.

7 Acknowledgments

This work was supported by the Novo Nordisk Foundation [grant number NNF20SA0067193]. Drs. Vincent Lombard and Nicolas Terrapon are gratefully acknowledged for their assistance in incorporating our data into the CAZy database. We also thank Dr. Philip Toukach for kindly providing a copy of the CSDB.

8 Author contributions

I.M. performed data acquisition, sequence analysis and interpretation; G.P.G. developed methodologies, supervised, analyzed and interpreted data; K.B. supervised, analyzed and interpreted data; C.J.H. performed custom structure predictions; B.H. conceived the study, supervised and interpreted results. The manuscript was written by I.M. and B.H. with help from all co-authors.

9 Competing interests

~~None~~ The authors declare no competing interests.

References

[1] Varki, A. *et al.* (eds.) *Essentials of Glycobiology* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2022), 4th edn. URL http://www.ncbi.nlm.nih.gov/books/NBK579918/.

[2] Laine, R. A. A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 x 10(12) structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology* **4**, 759–767 (1994).

[3] Lapébie, P., Lombard, V., Drula, E., Terrapon, N. & Henrissat, B. Bacteroidetes use thousands of enzyme combinations to break down glycans. *Nature Communications* **10**, 2043 (2019). URL https://www.nature.com/articles/s41467-019-10068-5.

[4] Lairson, L., Henrissat, B., Davies, G. & Withers, S. Glycosyltransferases: Structures, Functions, and Mechanisms. *Annual Review of Biochemistry* **77**, 521–555 (2008). URL https://www.annualreviews.org/doi/10.1146/annurev.biochem.76.061005.092322.

[5] Drula, E. *et al.* The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Research* **50**, D571–D577 (2022). URL https://academic.oup.com/nar/article/50/D1/D571/6445960.

[6] McDonald, A. G. & Tipton, K. F. Fifty-five years of enzyme classification: advances and difficulties. *The FEBS journal* **281**, 583–592 (2014).

[7] Campbell, J. A., Davies, G. J., Bulone, V. & Henrissat, B. A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *The Biochemical Journal* **326**, 929–939 (1997).

[8] Coutinho, P. M., Deleury, E., Davies, G. J. & Henrissat, B. An Evolving Hierarchical Family Classification for Glycosyltransferases. *Journal of Molecular Biology* **328**, 307–317 (2003). URL https://linkinghub.elsevier.com/retrieve/pii/S0022283603003073.

[9] Knauer, R. & Lehle, L. The oligosaccharyltransferase complex from yeast. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1426**, 259–273 (1999). URL https://linkinghub.elsevier.com/retrieve/pii/S0304416598001287.

[10] Cho, H. Assembly of Bacterial Surface Glycopolymers as an Antibiotic Target. *Journal of Microbiology* **60**, 359–367 (2023). URL https://link.springer.com/10.1007/s12275-023-00032-w.

[11] Sjodt, M. *et al.* Structure of the peptidoglycan polymerase RodA resolved by evolutionary coupling analysis. *Nature* **556**, 118–121 (2018). URL http://www.nature.com/articles/nature25985.

[12] Käshammer, L. *et al.* Cryo-EM structure of the bacterial divisome core complex and antibiotic target FtsWIQBL. *Nature Microbiology* **8**, 1149–1159 (2023). URL https://www.nature.com/articles/s41564-023-01368-0.

[13] Nygaard, R. *et al.* Structural basis of peptidoglycan synthesis by E. coli RodA-PBP2 complex. *Nature Communications* **14**, 5151 (2023). URL https://www.nature.com/articles/s41467-023-40483-8.

[14] Meeske, A. J. *et al.* SEDS proteins are a widespread family of bacterial cell wall polymerases. *Nature* **537**, 634–638 (2016). URL http://www.nature.com/articles/nature19331.

[15] Di Lorenzo, F. *et al.* A Journey from Structure to Function of Bacterial Lipopolysaccharides. *Chemical Reviews* **122**, 15767–15821 (2022). URL https://pubs.acs.org/doi/10.1021/acs.chemrev.0c01321.

[16] Islam, S. T. & Lam, J. S. Synthesis of bacterial polysaccharides via the Wzx/Wzy-dependent pathway. *Canadian Journal of Microbiology* **60**, 697–716 (2014). URL http://www.nrcresearchpress.com/doi/10.1139/cjm-2014-0595.

[17] Whitfield, C., Wear, S. S. & Sande, C. Assembly of Bacterial Capsular Polysaccharides and Exopolysaccharides. *Annual Review of Microbiology* **74**, 521–543 (2020). URL https://www.annualreviews.org/doi/10.1146/annurev-micro-011420-075607.

[18] Woodward, R. *et al.* In vitro bacterial polysaccharide biosynthesis: defining the functions of Wzy and Wzz. *Nature Chemical Biology* **6**, 418–423 (2010). URL http://www.nature.com/articles/nchembio.351.

[19] Bentley, S. D. *et al.* Genetic Analysis of the Capsular Biosynthetic Locus from All 90 Pneumococcal Serotypes. *PLoS Genetics* **2**, e31 (2006). URL https://dx.plos.org/10.1371/journal.pgen.0020031.

[20] Ruan, X., Loyola, D. E., Marolda, C. L., Perez-Donoso, J. M. & Valvano, M. A. The WaaL O-antigen lipopolysaccharide ligase has features in common with metal ion-independent inverting glycosyltransferases*. *Glycobiology* **22**, 288–299 (2012). URL https://academic.oup.com/glycob/article-lookup/doi/10.1093/glycob/cwr150.

[21] Ashraf, K. U. *et al.* Structural basis of lipopolysaccharide maturation by the O-antigen ligase. *Nature* **604**, 371–376 (2022). URL https://www.nature.com/articles/s41586-022-04555-x.

[22] Rai, A. K. & Mitchell, A. M. Enterobacterial Common Antigen: Synthesis and Function of an Enigmatic Molecule. *mBio* **11**, 1–19 (2020). URL https://journals.asm.org/doi/10.1128/mBio.01914-20.

[23] Alexander, J. A. N. & Locher, K. P. Emerging structural insights into C-type glycosyltransferases. *Current Opinion in Structural Biology* **79**, 102547 (2023). URL https://linkinghub.elsevier.com/retrieve/pii/S0959440X23000210.

[24] Emami, K. *et al.* RodA as the missing glycosyltransferase in Bacillus subtilis and antibiotic discovery for the peptidoglycan polymerase pathway. *Nature Microbiology* **2**, 16253 (2017). URL http://www.nature.com/articles/nmicrobiol2016253.

[25] Maczuga, N., Tran, E. N. H., Qin, J. & Morona, R. Interdependence of Shigella flexneri O Antigen and Enterobacterial Common Antigen Biosynthetic Pathways. *Journal of Bacteriology* **204**, e00546–21 (2022). URL https://journals.asm.org/doi/10.1128/jb.00546-21.

[26] Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research* **42**, D490–D495 (2014). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1178.

[27] Servais, C. *et al.* Lipopolysaccharide biosynthesis and traffic in the envelope of the pathogen Brucella abortus. *Nature Communications* **14**, 911 (2023). URL https://www.nature.com/articles/s41467-023-36442-y.

[28] Iguchi, A. *et al.* A complete view of the genetic diversity of the Escherichia coli O-antigen biosynthesis gene cluster. *DNA Research* **22**, 101–107 (2015). URL https://academic.oup.com/dnaresearch/article-lookup/doi/10.1093/dnares/dsu043.

[29] Liu, B. *et al.* Structure and genetics of *Shigella* O antigens. *FEMS Microbiology Reviews* **32**, 627–653 (2008). URL https://academic.oup.com/femsre/article-lookup/doi/10.1111/j.1574-6976.2008.00114.x.

[30] Liu, B. *et al.* Structural diversity in *Salmonella* O antigens and its genetic basis. *FEMS Microbiology Reviews* **38**, 56–89 (2014). URL https://academic.oup.com/femsre/article-lookup/doi/10.1111/1574-6976.12034.

[31] Kenyon, J. J., Cunneen, M. M. & Reeves, P. R. Genetics and evolution of Yersinia pseudotuberculosis O-specific polysaccharides: a novel pattern of O-antigen diversity. *FEMS Microbiology Reviews* **41**, 200–217 (2017). URL https://academic.oup.com/femsre/article/41/2/200/2996588.

[32] Hu, D., Liu, B., Dijkshoorn, L., Wang, L. & Reeves, P. R. Diversity in the Major Polysaccharide Antigen of Acinetobacter Baumannii Assessed by DNA Sequencing, and Development of a Molecular Serotyping Scheme. *PLoS ONE* **8**, e70329 (2013). URL https://dx.plos.org/10.1371/journal.pone.0070329.

[33] Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9**, 173–175 (2012). URL http://www.nature.com/articles/nmeth.1818.

[34] Liu, B. *et al.* Structure and genetics of *Escherichia coli* O antigens. *FEMS Microbiology Reviews* **44**, 655–683 (2020). URL https://academic.oup.com/femsre/article/44/6/655/5645236.

[35] Li, C. *et al.* Structural, Biosynthetic, and Serological Cross-Reactive Elucidation of Capsular Polysaccharides from Streptococcus pneumoniae Serogroup 16. *Journal of Bacteriology* **201**, 13 (2019).

[36] Lin, F. L. *et al.* Identification of the common antigenic determinant shared by Streptococcus pneumoniae serotypes 33A, 35A, and 20 capsular polysaccharides. *Carbohydrate Research* **380**, 101–107 (2013). URL https://linkinghub.elsevier.com/retrieve/pii/S000862151300284X.

[37] Lin, F. L. *et al.* Structure elucidation of capsular polysaccharides from Streptococcus pneumoniae serotype 33C, 33D, and revised structure of serotype 33B. *Carbohydrate Research* **383**, 97–104 (2014). URL https://linkinghub.elsevier.com/retrieve/pii/S0008621513003947.

[38] Bush, C. A., Cisar, J. O. & Yang, J. Structures of Capsular Polysaccharide Serotypes 35F and 35C of Streptococcus pneumoniae Determined by Nuclear Magnetic Resonance and Their Relation to Other Cross-Reactive Serotypes. *Journal of Bacteriology* **197**, 2762–2769 (2015). URL https://journals.asm.org/doi/10.1128/JB.00207-15.

[39] Kondakova, A. N. *et al.* Reinvestigation of the O-antigens of Yersinia pseudotuberculosis: revision of the O2c and confirmation of the O3 antigen structures. *Carbohydrate Research* **343**, 2486–2488 (2008). URL https://linkinghub.elsevier.com/retrieve/pii/S0008621508003443.

[40] Henrissat, B. & Bairoch, A. Updating the sequence-based classification of glycosyl hydrolases. *The Biochemical Journal* **316**, 695–696 (1996).

[41] Paton, J. C. & Trappetti, C. *Streptococcus pneumoniae* Capsular Polysaccharide. *Microbiology Spectrum* **7**, 7.2.33 (2019). URL https://journals.asm.org/doi/10.1128/microbiolspec.GPP3-0019-2018.

[42] Doyle, L. *et al.* Mechanism and linkage specificities of the dual retaining β-Kdo glycosyltransferase modules of KpsC from bacterial capsule biosynthesis. *Journal of Biological Chemistry* **299**, 104609 (2023). URL https://linkinghub.elsevier.com/retrieve/pii/S002192582300251X.

[43] Krinos, C. M. *et al.* Extensive surface diversity of a commensal microorganism by multiple DNA inversions. *Nature* **414**, 555–558 (2001). URL https://www.nature.com/articles/35107092.

[44] Smith, T. & Waterman, M. Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197 (1981). URL https://linkinghub.elsevier.com/retrieve/pii/0022283681900875.

[45] Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–780 (2013). URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mst010.

[46] Sonnhammer, E. L. & Hollich, V. Scoredist: A simple and robust protein sequence distance estimator. *BMC Bioinformatics* **6**, 108 (2005). URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-108.

[47] Crippen, G. & Havel, T. *Distance Geometry and Molecular Conformation*. Chemometrics research studies series (Research Studies Press, 1988). URL https://books.google.dk/books?id=XjRCAQAAIAAJ.

[48] Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-421.

[49] Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**, W29–W37 (2011). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr367.

[50] Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006). URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl158.

[51] Huszczynski, S. M., Hao, Y., Lam, J. S. & Khursigara, C. M. Identification of the Pseudomonas aeruginosa O17 and O15 O-Specific Antigen Biosynthesis Loci Reveals an ABC Transporter-Dependent Synthesis Pathway and Mechanisms of Genetic Diversity. *Journal of Bacteriology* **202** (2020). URL https://journals.asm.org/doi/10.1128/JB.00347-20.

[52] Hagberg, A., Swart, P. & S Chult, D. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Annual Python in Science Conference, Pasadena, CA, August 19 –24, 2008*, 11–16 (2008). URL https://www.osti.gov/biblio/960616.

[53] Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009). URL https://academic.oup.com/bioinformatics/article/25/9/1189/203460.

[54] Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019). URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3019-7.

[55] Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* **13**, 2498–2504 (2003). URL http://genome.cshlp.org/lookup/doi/10.1101/gr.1239303.

[56] Petersen, B. O., Meier, S., Paulsen, B. S., Redondo, A. R. & Skovsted, I. C. Determination of native capsular polysaccharide structures of Streptococcus pneumoniae serotypes 39, 42, and 47F and comparison to genetically or serologically related strains. *Carbohydrate Research* **395**, 38–46 (2014). URL https://linkinghub.elsevier.com/retrieve/pii/S0008621514002560.

[57] Petersen, B. O., Hindsgaul, O., Paulsen, B. S., Redondo, A. R. & Skovsted, I. C. Structural elucidation of the capsular polysaccharide from Streptococcus pneumoniae serotype 47A by NMR spectroscopy. *Carbohydrate Research* **386**, 62–67 (2014). URL https://linkinghub.elsevier.com/retrieve/pii/S0008621513004084.

[58] Lam, J. S., Taylor, V. L., Islam, S. T., Hao, Y. & Kocíncová, D. Genetic and Functional Diversity of Pseudomonas aeruginosa Lipopolysaccharide. *Frontiers in Microbiology* **2** (2011). URL http://journal.frontiersin.org/article/10.3389/fmicb.2011.00118/abstract.

[59] Toukach, P. V. & Egorova, K. S. Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts. *Nucleic Acids Research* **44**, D1229–D1236 (2016). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv840.

[60] Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49**, W293–W296 (2021). URL https://academic.oup.com/nar/article/49/W1/W293/6246398.

[61] R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2023). URL https://www.R-project.org/.

[62] Posit team. *RStudio: Integrated Development Environment for R* (Posit Software, PBC, Boston, MA, 2023). URL http://www.posit.co/.

[63] Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016). URL https://ggplot2.tidyverse.org.

[64] Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **9**, 90–95 (2007).

[65] Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nature Methods* **19**, 679–682 (2022). URL https://www.nature.com/articles/s41592-022-01488-1.

[66] Schrödinger, L. The PyMOL Molecular Graphics System, Version 2.5 (2020).

[67] Shindyalov, I. N. & Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering* **11**, 739–747 (1998).