1 Introduction

~~Photosynthesis has granted homotrophs access to vast amounts of carbohydrates which serve as abundant carbon sources for most heterotrophs.~~ (This paragraph can maybe be shortened a bit) Carbohydrate polymers (glycans) and glyco-conjugates ~~have thus become~~ 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 enzymatic breakdown.

The transfer of sugar moieties to acceptor molecules such as proteins, lipids or other sugars, is performed by enzymes called glycosyltransferases or GTs [4]. ~~GTs can be classified either by activity or by sequence similarity. The Enzyme Commission of the International Union of Biochemistry and Molecular Biology (IUBMB), has elaborated a classification system that integrates a description of the donor, acceptor and bond formed, summarized in the form of an EC number [5]. This activity-based classification, although enormously useful to avoid the proliferation of trivial names, has the limitation that it does not integrate the structural features of the enzymes nor can it easily accommodate enzymes that act on several substrates [5].~~

Campbell and colleagues (1997) proposed a sequence-based classification of GTs into 26 families, which was subsequently expanded to 65 families in 2003 [6]. 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. The constantly updated GT classification is presented in the carbohydrate-active enzymes database (CAZy; `www.cazy.org`) along with similar family classifications of other carbohydrate-active enzymes [7]. An additional advantage of the sequence-based classification is that it readily enables genome mining for the presence of family members. Today there are 116 GT families in the CAZy database and this number will continue growing as novel glycosyltransferases are progressively discovered or as known GTs are incorporated in the database. In contrast to the EC numbers [5], the sequence-based classification implicitly incorporates the structural features of GTs including the conservation of the catalytic residues. ~~Structurally, there are two major folds for the nucleotide-sugar dependent GTs, namely GT-A and GT-B, which both have Rossmann folds. By contrast, sugar-phospholipid-utilizing GTs are integral membrane proteins which have an overall GT-C fold with a number of transmembrane helices that varies from 8 to 13 [4].~~

(This paragraph can maybe be shortened a bit) 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 [8, 6]. 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]. 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 safely even if the precise transferred carbohydrate is not known.

The large majority of the 116 families of GTs listed in the CAZy database use donors activated by nucleotide diphosphates. Eleven families utilize nucleotide monophospho-sugars (sialyl and KDO transferases), while 12 families utilize lipid monophospho-sugars. Only one family in the CAZy database utilizes lipid diphospho-oligosaccharide donors: the oligosaccharyltransferases of family GT66, which transfer a pre-assembled oligosaccharide to asparagine residues in N-glycoproteins [4, 9]. Several lipid diphospho-oligosaccharide-utilizing GTs are currently missing in the CAZy database.

By contrast to the nucleotide-sugar dependent GTs, which are globular proteins with either a GT-A or GT-B fold, the sugar-phospholipid-utilizing GTs are integral membrane proteins which have an overall GT-C fold with a number of transmembrane helices that varies from 8 to 14 [4, 10]. Alexander and Locher recently suggested two subgroups of GT-C glycosyltransferases, GT-C_A and GT-C_B, based on the structural features of several of these families [10]. Several GTs from the GT-C_B subclass are missing in CAZy. They fall into four major functional classes, which are all involved in the synthesis of bacterial cell wall polysaccharides and, like CAZy family GT66, they catalyze the transfer of a glycan activated by the diphospholipid undecaprenyl diphosphate (Und-PP).

The first functional group is the peptidoglycan polymerases, SEDS (shape, elongation, division and sporulation) proteins. These proteins polymerize peptidoglycan in paris with class B penicillin-binding proteins, which perform peptidoglycan crosslinking [11]. The structure of a SEDS protein from *Thermus thermophilus* has been determined and consists of 10 transmembrane helices with several large extracellular loops containing functionally important residues [12]. A large hydrophobic groove containing highly conserved residues is thought to be the lipid binding site. An Asp residue has been shown to be essential for RodA function in both *T. thermophilus* and *B. subtilis* [12, 13].

The other three functional groups are involved in the synthesis of bacterial surface polysaccharides. Bacteria synthesize various surface polysaccharides which confer them antigenic properties. Lipopolysaccharide (LPS) is a polysaccharide specific of Gram-negative bacteria, and consists of the serotype-specific O-antigen attached to the Lipid A-core oligosaccharide which is located in the outer membrane [14]. On the other hand capsular

polysaccharides (CPS~~also known as K-antigens~~, containing the K-antigen) are produced by both Gram-negative and Gram-positive bacteria [15]. The covalent anchoring of CPS is still poorly understood, although it is found to be linked to peptidoglycan in some Gram-positives [15]. Bacteria from the Enterobacterales order produce yet another type of surface polysaccharides referred to as the enterobacterial common antigen (ECA), which consists of repeating units of N-acetylglucosamine, N-acetyl-D-mannosaminuronic acid and 4-acetamido-4,6-dideoxy-D-galactose [16]. Most of these surface polysaccharides are produced via the so-called Wzx/Wzy-dependent pathway, which takes place on the plasma membrane (inner membrane in Gram-negatives) [17]. In this pathway, sugar repeat units are assembled on an undecaprenyl-diphosphate (Und-PP) anchor on the cytoplasmic side of the membrane and then flipped to the outside of the membrane by the flippase Wzx. The repeat units are then polymerized by the bacterial polysaccharide polymerases (Wzy; BP-Pols), by transferring the growing polymer to the incoming new repeat units [17, 18]. In the case of LPS, the polymer (O-antigen) is then ligated onto Lipid A-core oligosaccharide by the O-antigen ligase (WaaL; O-Lig) [19]. ECA is produced via the same pathway, but with another set of enzymes including the polymerase (WzyE). In order to distinguish these polymerases from the serotype-specific polymerases, they are here referred to as ~~ECA~~ enterobacterial common antigen polymerases (ECA-Pols).

~~Several of the GTs from these pathways are missing from the CAZy database including ECA-Pols, BP-Pols, and O-Ligs, as well as some peptidoglycan polymerases. These enzymes share with CAZy family GT66 the particularity of catalyzing the transfer of oligosaccharides and, like GT66, their donor is also activated by a diphospholipid (Und-PP).~~ In an attempt to complete the sequence-based classification of GTs, we have performed a detailed analysis of the primary sequence of ~~peptidoglycan polymerases, polysaccharide polymerases and O-antigen ligases~~ SEDS proteins, ECA-Pols, BP-Pols and O-Ligs to assign their sequences to CAZy families and examined how sequence diversity correlates with the diversity of the transferred oligosaccharides and with the stereochemical outcome of the glycosyl transfer reaction.

2 Results

2.1 Peptidoglycan Polymerases

~~The synthesis of peptidoglycan is primarily performed by class A penicillin binding proteins(PBPs), which harbor a GT51 domain and a transpeptidase domain [20, 13]. However, it has been shown that peptidoglycan polymerization is also performed by the proteins RodA [11] and FtsW [21], often called shape, elongation, division and sporulation (SEDS) proteins. FtsW operates in complex with a transpeptidase that performs the peptide cross linking [12]). For RodA and FtsW~~For building the CAZy family of SEDS proteins, we used the sequence from the published structure, 6BAR [12], as a starting point. The family GTxx1 was created and populated by using BLAST against Genbank, and subsequently using an HMM search against Genbank. GTxx1 is a very large family currently counting over 57,200 Genbank members in the CAZy database with a sequence similarity greater than 19% sequence identity over 221 residues.

The taxonomic distribution of family GTxx1 follows what was reported in [13], 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 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.

~~The three-dimensional structure of RodA from *Thermus thermophilus* has been determined and consists of 10 transmembrane helices with several large extracellular loops containing functionally important residues [12]. A large hydrophobic groove containing highly conserved residues is thought to be the lipid binding site. An Asp residue has been shown to be essential for RodA function in both *T. thermophilus* and *B. subtilis* [12, 13].~~

~~Sequence-wise we found excellent sequence similarity between RodA and FtsW proteins from various sources and they were easily grouped together in a single, very large family (GTxx1) currently counting over 57,200 members in the CAZy database and showing no significant sequence similarity to other GT families.~~

~~The taxonomic distribution of family GTxx1 follows what was reported in [13], namely that this protein family is present in all bacteria except for Mycoplasma. It is present in most but not all planctomycetes.~~

2.2 Enterobacterial common antigen polymerases

The ECA-Pol which was studied in [22] was used as seed sequence for the ECA-Pol family. Although the CAZy database only lists Genbank entries [23], we decided to build our multiple sequence alignments (MSAs) with 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. The ECA-Pols ~~display a high sequence conservation, consistent with the conservation of acceptor, donor and product of the reaction. ECA-Pols were therefore~~ were assigned to a single new ~~and homogeneous CAZy family~~CAZy family, GTxx2.

To date this new family contains over 4800 members ~~. The repeat unit being axially bound to Und-PP and axially linked in the final polymer, this reaction is retaining the configuration of the anomeric carbon undergoing catalysis~~with 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) 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 (~~vide infra).~~see below).why see below here?

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 was constructed ~~with~~ of the sequence library using our in-house Aclust tool which revealed four distantly related clades (Supplementary Fig. 1). The O-Ligs were included into one new CAZy family, GTxx3 with >16,700 members distributed in four subfamilies.

The greater diversity of the GTxx3 ~~O-antigen ligases~~ O-Ligs compared to the GTxx1 peptidoglycan polymerases and GTxx2 ECA-Pol appears in the form of the four divergent clades in the O-Lig phylogenetic tree (Supplementary Fig. 1). We hypothesize that this increased diversity originates from the extensive donor and moderate acceptor variability of O-Ligs [14]. Taxonomically, the GTxx3 O-Lig family is present in most bacteria, including both Gram-negatives and Gram-positives. 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 [19].

A recently discovered ~~O-antigen ligase~~O-Lig, WadA, is bimodular with a GTxx3 domain appended to a globular glycosyltransferase domain of family GT25, which adds the last sugar to the oligosaccharide core [24]. We have constructed a tree with representative WadA homologs from the GTxx3 family (Supplementary Fig. 2) and observe that most of the sequences appended to a GT25 domain ~~cluster together in one area, which~~ form one clade in the tree, except for a few outliers. This suggests a coupled action of the GT25 and of the GTxx3 at least for the bimodular ~~O-antigen ligases and maybe~~ O-ligs and possibly for the entire family. The bimodular WadA ~~ligase~~ O-Lig is observed in five genera including Mesorhizobium and Brucella.

2.4 Other bacterial polysaccharide polymerases

~~We collected 365 bacterial~~ The fourth functional subgroup of GT-C_B are the BP-Pols. There is to our knowledge only one experimentally characterized BP-Pol [18]. However, several studies have identified BP-Pols from the polysaccharide gene clusters, and we decided to build our families based on these. We thus collected 363 predicted BP-Pol sequences (also referred to as Wzy) from seven ~~different studies that have reported BP-Pol sequences from~~ review papers for various species, both Gram-negatives and Gram-positives: *Escherichia coli* [25], *Shigella boydii*, *Shigella dysenteriae*, *Shigella flexneri* [26], *Salmonella enterica* [27], *Yersinia pseudotuberculosis*, *Yersinia similis* [28], *Pseudomonas aeruginosa* [17], *Acinetobacter baumanii*, *Acinetobacter nosocomialis* [29] and *Streptococcus pneumoniae* [30] (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 [17]. 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 hydrophobic 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 ~~expanded the~~ first built a sequence library by running BLAST against the NCBI non-redundant database for each of the 365 BP-Pol seeds. ~~However, clustering~~ 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. 1a). 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 [31]. We then combined the SSN clusters into ~~"superclusters"~~ "superclusters" in a network analysis based on the HHblits scores (Fig. 1b). For this, we used a score cut-off of 160 in order to get a meaningful sequence and organismal diversity, resulting in 28 ~~"superclusters "~~ superclusters of varying sizes and 86 singleton clusters. Interestingly, the BP-Pols cluster across

taxonomy, and even BP-Pols from Gram-positive and Gram-negative bacteria cluster together. The 14 largest ~~"superclusters "~~ superclusters have been included as new GT families in the CAZy database (GTxx4-GTx17) with a number of members ranging from 159 to 5,979 at the time of submission. Only ~~152 of the 365~~ 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 range of taxonomy, 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.

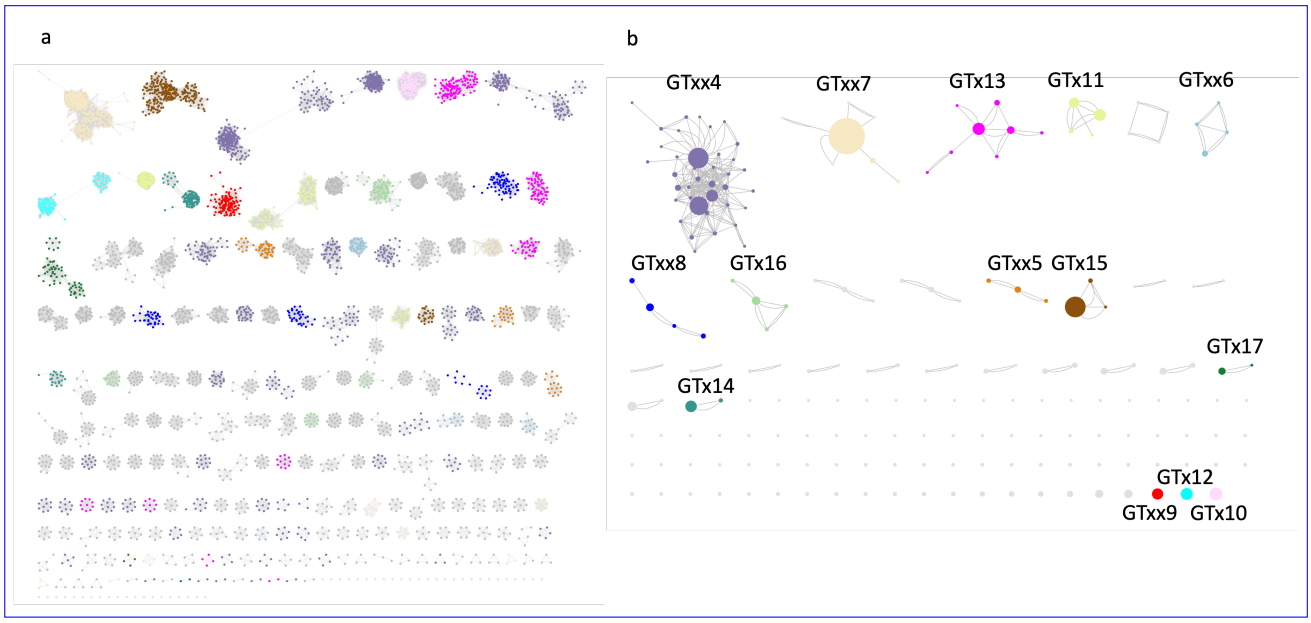


Figure 1: Clustering of BP-Pol sequences. a) SSN network with nodes representing proteins and edges representing pairwise alignment bit scores. b) HHblits network with nodes presenting SSN clusters and edges representing HHblits scores. The resulting clusters are referred to as "superclusters". There are two edges between nodes, when the HHblits score is above the threshold 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 CAZy families GTxx4 - GTx17. In both a and b, the SSN clusters are coloured according to which supercluster they belong to.

2.5 Analyzing the sugars transferred by bacterial polysaccharide polymerases

~~There are two possible outcomes for the BP-Pol-catalyzed polymerization reaction, either retaining or inverting the α-configuration of the anomeric carbon of the carbohydrate carrying the Und-PP moiety. Examination of the structure of the polysaccharides produced thus often reveals the stereochemistry of the bond formed by the polymerases. In order to assess if the stereochemical mechanism is conserved in the families, we retrieved the structures of the transferred sugar repeat units from the Carbohydrate Structure Database (CSDB)[32]. As mentioned above, 152 of the original 365 original~~ 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 ([33, 26, 27, 28, 17, 29, 30]). Additionally, eight sugar structures were included, which were published after the review papers [34, 35, 36, 37, 38]. Out of the 150 BP-Pol seed sequences that were included in the new ~~families. Out of these 152 BP-Pols, 132 were matched~~ CAZy families, we matched 131 with a sugar structure. ~~In these structures, the~~ The repeat units are oligosaccharides with 3-7 monomers within the backbone, often with branches. In ~~several of the studies from which the BP-Pol sequences were retrieved~~most of the cases, the bond which is formed by the polymerase has been identified ~~[29, 28, 30, 27]. In cases where the polymerase linkage was not clear from the literature, we identified it by comparing with similar sugar structures from similar polymerases.~~ in the review papers. explain a bit more (details in section 4.)

Having retrieved the sugar structures transferred by the BP-Pols, we first analyzed the stereochemistry of the bond catalyzed by the polymerase. As mentioned above, the stereochemical mechanism (inverting or retaining) is usually 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 BP-Pol CAZy families and varies from one family to another (Fig. 2). There ~~are two apparent exceptions, however, where one of the polymerase linkages has a different stereochemistry than the rest of the family . This is attributable to when a wrong polysaccharide was assigned to the polysaccharide gene cluster comprising the polymerase (for example if the bacteria produces several surface polysaccharides), or when there was~~ is only one exception: in family GTxx8 the polymerase linkages are all equatorial except for the O-antigen in *Pseudomonas aeruginosa* O4, where it is axial. It seems likely that there is either an error in the chemical structure ~~reported for the polysaccharide, or when the linkage made by the polymerase was wrongly predicted. For example in family GTxx4, one of the polymerase linkages is axial while the other 37 are equatorial~~ or that the serotype designation was incorrect.

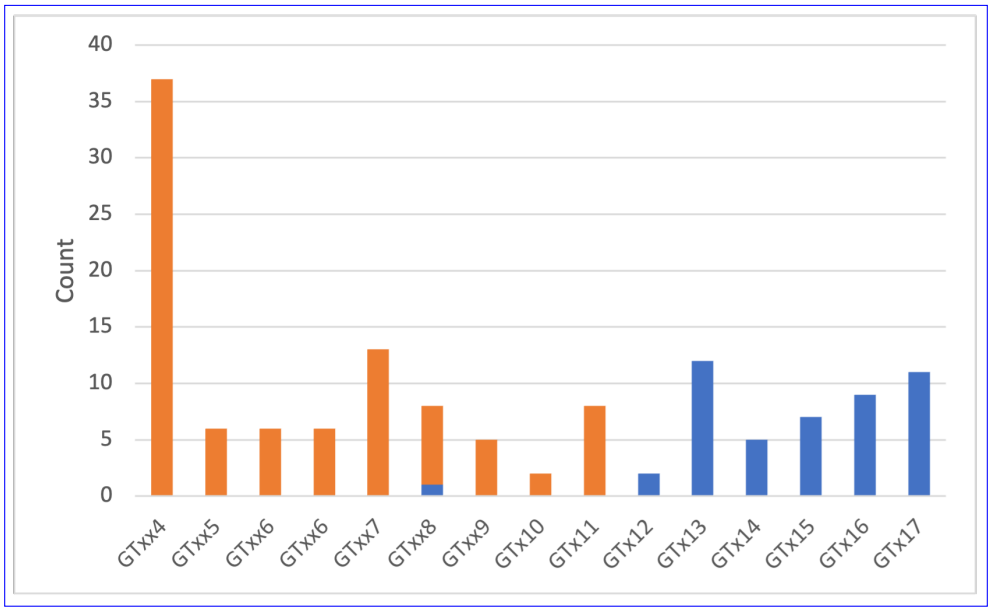


Figure 2: Conservation of stereochemical outcome of the reaction catalyzed in the various BP-Pol families. Equatorial bonds are shown in orange and axial bonds are shown in blue.

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 see that the sugars within each family are similar, and correlates with the structure of the tree reformulate (Fig. ~~2~~3, Supplementary Fig. 3). ~~It seems likely that there is either an error in the chemical structure or that the bacteria contains two different polymerases, one that catalyzes the formation of an equatorial bond and one that catalyzes the formation of an axial bond.~~ Notably, we observe examples of BP-Pols from distant taxonomic serotypes that cluster in the same CAZy family and have highly similar sugars. For example, (*Escherichia coli* O178 and *Streptococcus pneumoniae*) 47A in GTxx7 transfer sugars with almost identical backbones. There is only a slight variance in the middle of the repeat unit. This suggest that horizontal gene transfer has occured. reformulate

To quantify the correlation between BP-Pol sequence and sugar structure, we ~~analyzed the oligosaccharide repeat units associated with each of the CAZy families. For this purpose, we~~ developed an original pairwise oligosaccharide similarity score.

In our scoring scheme, the similarity of two glycans is estimated by examining subsite moieties immediately upstream and downstream of the newly created interosidic bond, as we hypothesize that these are the moieties most fitting the active site of the polymerase (Fig. ~~3~~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).

~~An idealized representation of a BP-Pol. The donor is the growing glycan chain activated by undecaprenyl pyrophosphate while the acceptor is a repeat unit monomer. 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.~~

Using our glycan similarity scoring system, we found a correlation between sugar similarity and polymerase

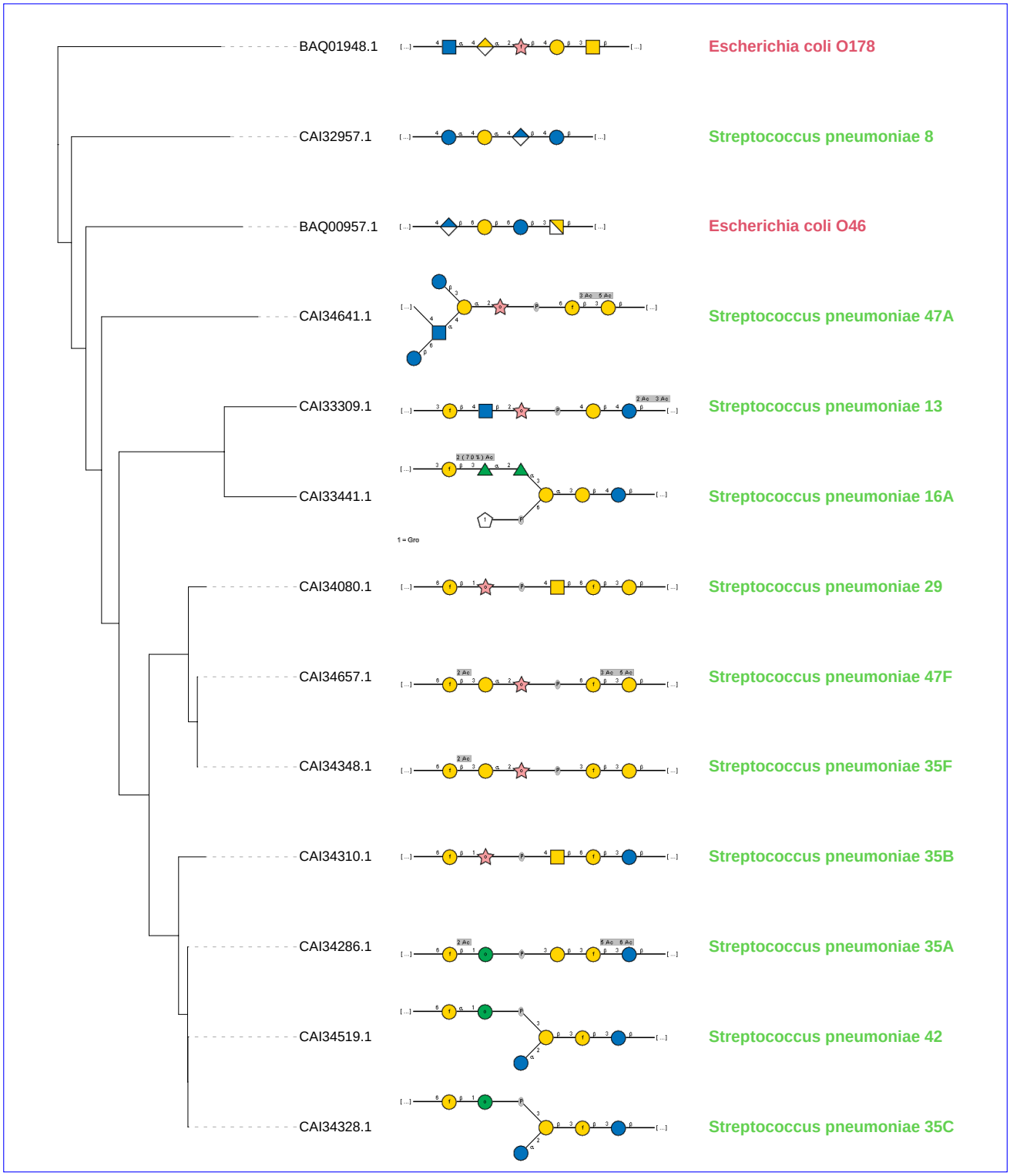


Figure 3: ~~Conservation~~ Phylogenetic tree of ~~stereochemical outcome~~ BP-Pols in family GTxx7 with structures of the ~~reaction catalyzed~~ corresponding sugar repeat units in ~~the various BP-Pol families~~SNFG format. The ~~bond formed by the polymerase was retrieved~~ family contains BP-Pols from ~~literature or deduced by comparison to similar sugars~~ bacteria from distant taxonomies which transfer similar ~~polymerases~~sugars. ~~Equatorial bonds~~ The trees for all the families are shown in ~~orange and axial bonds are shown in blue.~~Supplementary Figure 3.

sequence similarity (Fig. ~~4)(Detailed: Supplementary Fig. 3)~~5), supported by a preponderance of similarity scores appearing close to the score matrix diagonal and within each individual family.

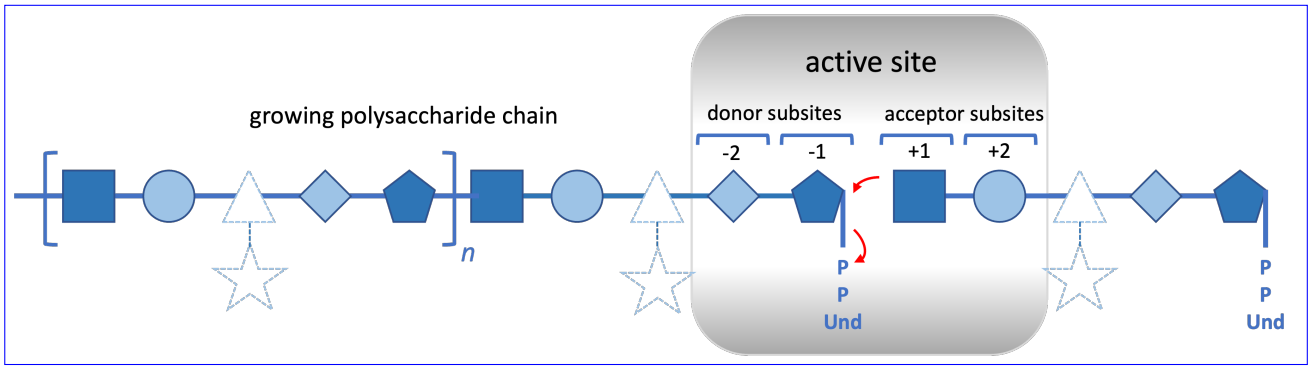


Figure 4: An idealized representation of a BP-Pol. The donor is the growing glycan chain activated by undecaprenyl pyrophosphate while the acceptor is a repeat unit monomer. The reaction is hypothesized to chiefly involve the sugar residues of the donor (subsites -2 and -1) and of the acceptor (subsites +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.

~~Notably, we observe examples of BP-Pols from distant taxonomic serotypes that cluster in the same CAZy family and have highly similar sugars. In Fig. 5, we show two examples of that. In GTxx7, two BP-Pols from very distant taxonomical origin (*Escherichia coli* and *Streptococcus pneumoniae*) transfer sugars with almost identical backbones. There is only a slight variance in the middle of the repeat unit. In GTx15, three BP-Pols from three different genera transfer glycans that, although consisting of different monosaccharides, have identical composition when translated into the backbone geometry description (Fig 5). These could be examples of horizontal gene transfer.~~

~~Examples of potential horizontal gene transfer. Top: Two BP-Pols from family GTxx7 from distant taxonomies transfer similar sugars. Bottom: Three different BP-Pols from different genera transfer similar sugars. The glycans are shown in SNFG representation and backbone geometry descriptors.~~

2.6 Comparison of families

Others have previously reported sequence and structural similarity between RodA, O-Lig and some BP-Pols [39, 10, 40, 13]. In order to investigate the relatedness of the new CAZy families, we compared the family HMMs by all-vs-all HHblits ~~analyzes~~ analyses [31] (Fig. 6). Strikingly, we observe that the retaining BP-Pol families cluster together on the heatmap together with the retaining ECA-Pols, while the inverting BP-Pols form two distinct groups, one of them containing the inverting O-Ligs. The background noise between some inverting and the retaining enzymes likely due to the general conservation of the successive transmembrane helices, which is altered in the GTxx4-GTxx5-GTxx6 subgroup due to their different architecture (~~vide infra~~see below); on the other hand, peptidoglycan polymerases segregate away from the other families.

~~Alexander and Locher recently suggested two subgroups of GT-C glycosyltransferases, GT-C$_A$ and GT-C$_B$, based on the structural features of several of these families [10].~~ 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 [41]. In extension of the report of the GT-C$_B$ class by Alexander and Locher ~~([10])~~[10], and based on the above-mentioned similarities between the new CAZy families, we can now define three clans within GT-C$_B$: GT-C$_{B1}$ consisting of inverting BP-Pol families 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). 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 clan GT-C$_{B3}$ to the structural subclass GT-C$_B$ of Alexander and Locher [10]. 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, 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 on the ~~outside~~ outer face of the membrane. Enzymes of clans ~~GT-CB1 and GT-CB2~~ GT-C$_{B1}$ and GT-C$_{B2}$ have a long extracellular loop close to the C-terminus (Fig. 7). In stark contrast, families GTxx4, GTxx5 and GTxx6 of clan ~~GT-CB3~~ GT-C$_{B3}$

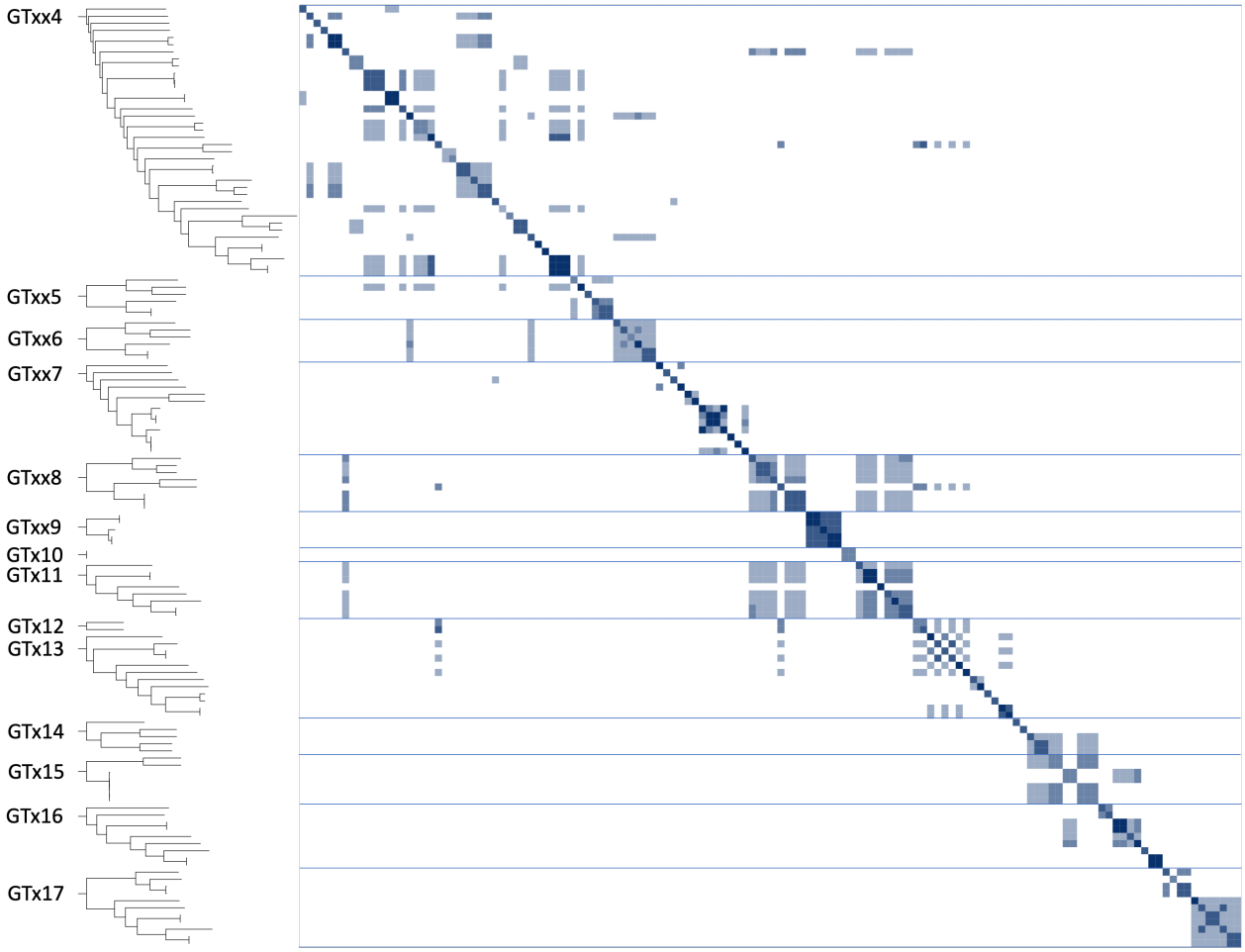


Figure 5: Glycan similarity of sugar repeat units polymerized by BP-Pols. All "seed" BP-Pols where the corresponding transferred oligosaccharide was known were included in the heatmap. A phylogenetic tree is shown for the polymerases in each CAZy family on the left. The glycan similarity scores are shown in a color scale of light blue (identical matches at both –1 and +1 sites) to dark blue (identical matches including both –2, +2 site positions). Blue lines separate the families.

Structural subclass Alexander & Locher	**CAZy clan**	**CAZy families**	**Mechanism**	**Donor**
GT-C_A (7 conserved TM helices)	-	GT53	Inverting	Lipid-P-monosaccharide
	-	GT83	Inverting	Lipid-P-monosaccharide
	-	GT39	Inverting	Lipid-P-monosaccharide
	-	GT57	Inverting	Lipid-P-monosaccharide
	-	GT66	Inverting	Lipid-PP-oligosaccharide
GT-C_B (10 conserved TM helices)	-	GTxx1	Inverting	Lipid-PP-oligosaccharide
	GT-C_{B1}	GTxx3, GTxx7, GTxx8, GTxx9, GTx10, GTx11	Inverting	Lipid-PP-oligosaccharide
	GT-C_{B2}	GTxx2, GTx12, GTx13, GTx14, GTx15, GTx16, GTx17	Retaining	Lipid-PP-oligosaccharide
	GT-C_{B3}	GTxx4, GTxx5, GTxx6	Inverting	Lipid-PP-oligosaccharide
-	-	GT22	Inverting	Lipid-P-monosaccharide
	-	GT50	Inverting	Lipid-P-monosaccharide
	-	GT58	Inverting	Lipid-P-monosaccharide
	-	GT59	Inverting	Lipid-P-monosaccharide

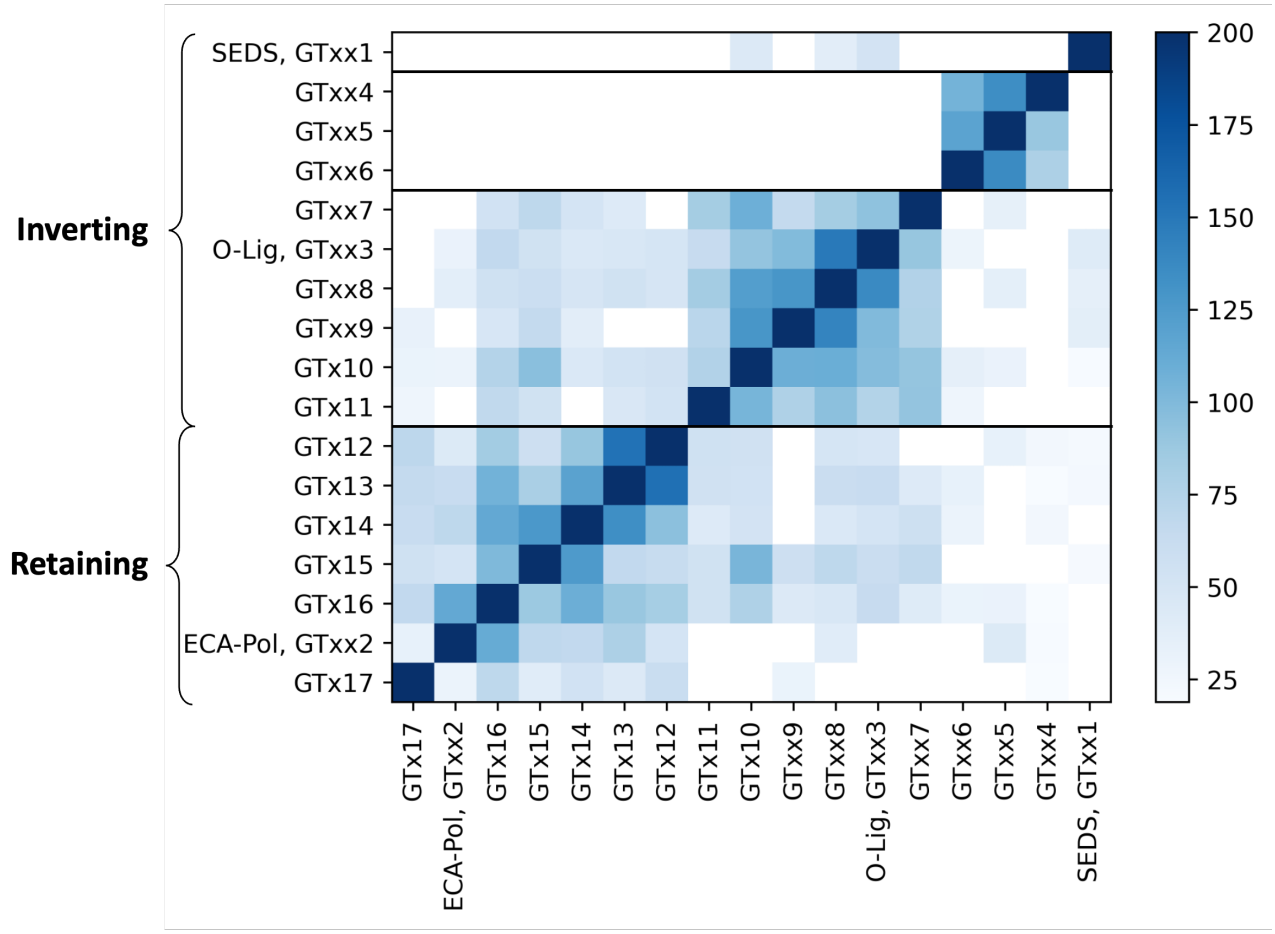


Figure 6: Heatmap of inter-family HHblits bit scores. The HHblits scores are shown on a color scale from white (low similarity score) to dark blue (high similarity score). The HHblits scores depend on the direction of the alignment, and therefore the heatmap is not symmetrical.

have an architecture completely different from that of the two other clans (Fig. 7), with the long loop located close to the N-terminus, and a conservation of one Asp, one His and two Arg residues.

Most of the families in the inverting Clan ~~GT-CB1~~ GT-C_{B1} have two conserved Arg residues and one conserved either Glu/Asp (in the BP-Pols) or His residue (in the O-Lig) (Fig. 7). In the pairwise HHblits alignments and structural superimpositions, the Glu/Asp/His residues align, suggesting that they could play the same role. As an example, the structural superimposition of the published O-Lig structure (7TPG) [40] and an AlphaFold model from one representative of the inverting BP-Pol family GTxx8 is shown in Fig. 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 in O-Lig is placed in the same position as the conserved Glu in the BP-Pol. In O-Lig, the conserved His has been proposed to activate the acceptor, while the two Args are proposed to position the donor by binding to the phosphate groups [40]. We hypothesize that the Glu and Asp residues in the BP-Pols play the same role as the His in O-Lig.

In the retaining clan GT-C_{B2}, the pattern of conservation looks different. Here, most of the families have 2-3 conserved Arg/Lys and one conserved Tyr. As an example of the structural similarity in this clan, the structural superimposition of AlphaFold models from the ECA-Pol family GTxx2 and family GTx16 is shown in Fig 8b. The structures again show low overall similarity (RMSD 5.4Å over 360 residues), but the conserved residues are oriented very similarly. This also shows that ECA-Pols display similarity to the BP-Pols of clan GT-C_{B2}.

Although the peptidoglycan polymerase family, GTxx1 does not cluster in any of the three clans, it does display topographical similarity to clan GT-C_{B1}. In terms of architecture it also contains a long extracellular loop with a conserved Arg and the conserved and essential Asp residue [12]. The Asp residue is in a similar position as the Asp/Glu/His in 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 [40].

3 Discussion

Here we have added 17 glycosyltransferase families (GTxx1 to GTx17) to the CAZy database bringing the total of covered families from 116 to 133. 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. The list of these seed sequences is given in Supplementary Table 1-2 for families GTxx3 to GTx17. No seed sequence was needed for peptidoglycan polymerases (GTxx1) as the family is very tight around two structurally and functionally characterized members.

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 before because of their extreme diversity even within a single species [25], 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 diversity from the NCBI non-redundant database, 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*. Gram-positive bacteria do not produce LPS, but instead capsular polysaccharides (CPS), which are linked to the peptidoglycan layer [42]. Thus a hypothesis could be that the GTxx3 members in *S. pneumoniae* are responsible for the ligation of CPS to the peptidoglycan layer.

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* (GenBank BAQ01516.1) and *A. baumanii* (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. Thus new GT families based on O-antigen polymerases are poised to be formed when additional evidence becomes available.

Moreover, we observe that the sequence diversity within the families we have built is minimal for peptidoglycan polymerases (GTxx1), and then increases gradually from ECA-Pols (GTxx2) to O-Ligs (GTxx3) and is maximal for BP-Pols (GTxx4-GTx17). We hypothesize that sequence diversity reflects the donor and acceptor diversity in each family since the latter increases accordingly.

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

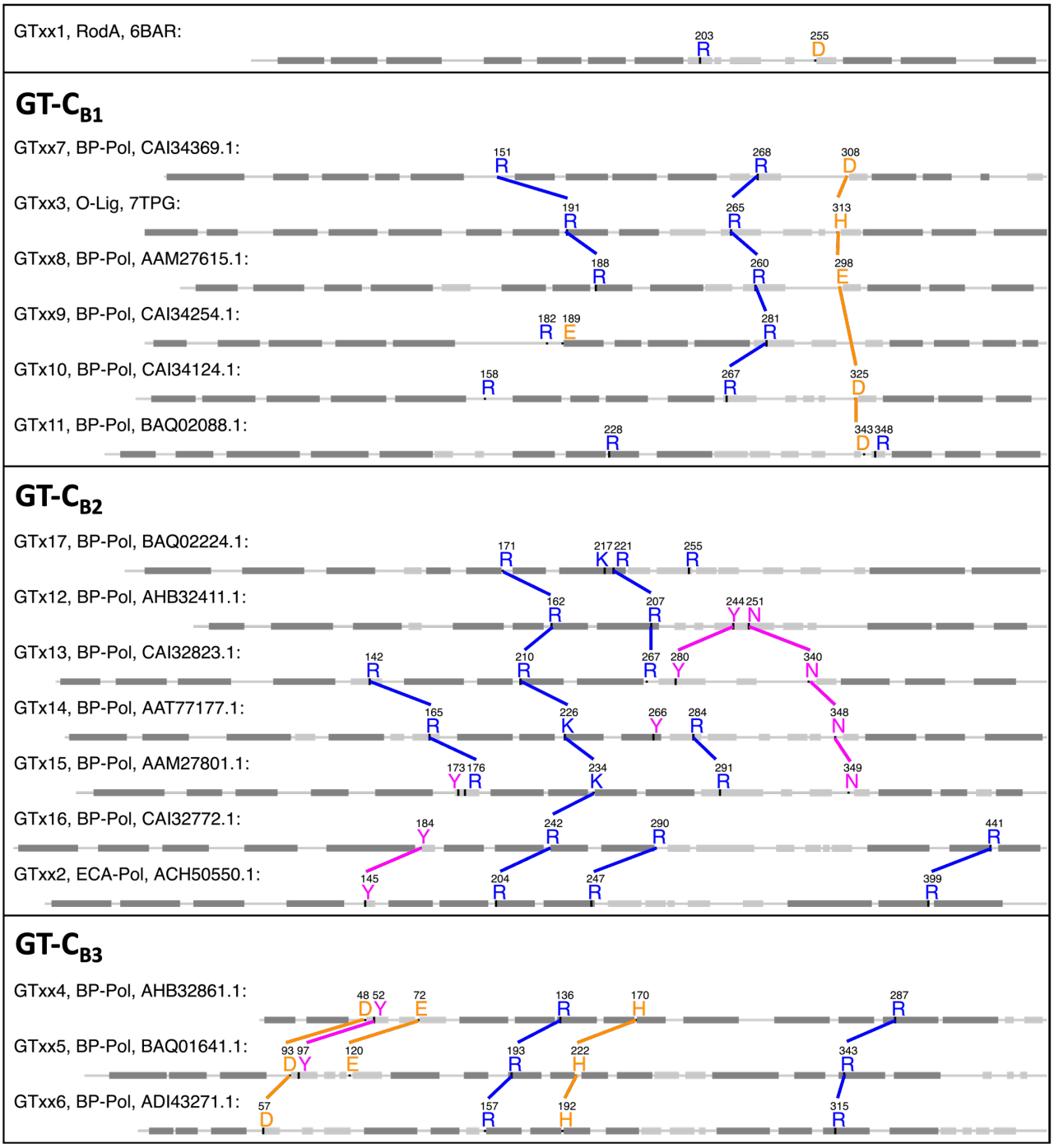
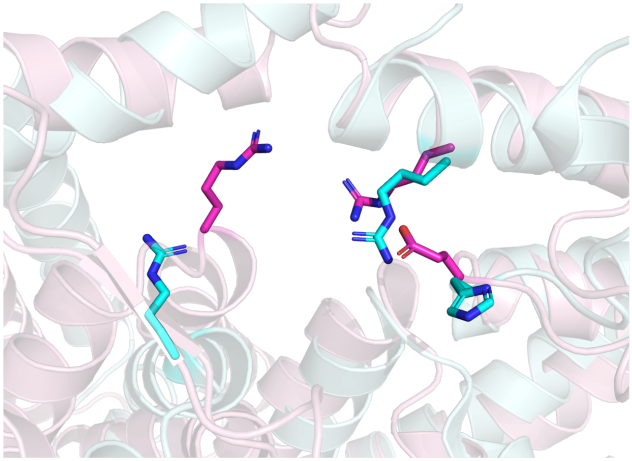
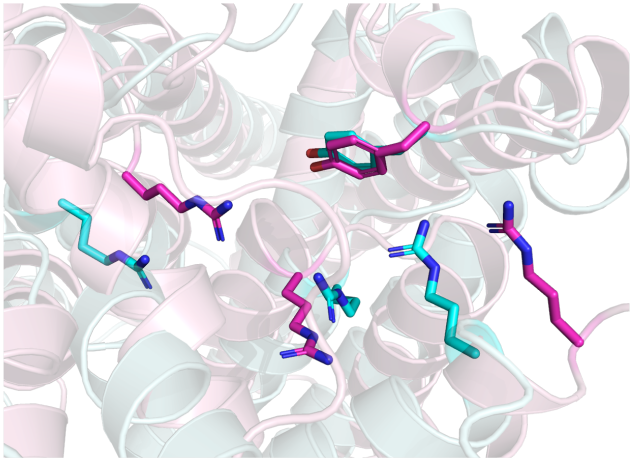


Figure 7: Comparison of conserved residues in the new GT families. The non-aliphatic conserved residues of each of the new CAZy families are shown on representative sequences. Transmembrane helices are shown in dark gray boxes, non-transmembrane helices are shown in light gray boxes. Lines are shown between residues that align in pairwise structural superimpositions. The secondary structure was retrieved from the crystal structures for family GTxx1 and GTxx3 (6BAR and 7TPG respectively) and from AlphaFold models ~~or from experimental structures where available~~for all other families.



Cyan: O-Lig, GTxx3 (7TPG)
Pink: BP-Pol, GTxx8 (AAM27615.1)



Cyan: ECA-Pol, GTxx2 (ACH50550.1)
Pink: BP-Pol, GTx16 (CAI32772.1)

Figure 8: Structural superimposition of different families with conserved residues. a) O-Lig from GTxx3 (PDB 7TPG) and AlphaFold model of BP-Pol from GTxx8 (RMSD 5.3Å over 192 residues). The conserved Glu in GTxx8 is aligning with the conserved His in GTxx3, which is proposed to activate the acceptor [40]. b) AlphaFold models of ECA-Pol from GTxx2 and BP-Pol from GTx16 (RMSD 5.4Å over 360 residues). The conserved residues are all in similar positions.

activation of the acceptor above (SN_2) or below (SN_i) 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 [43]. 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 SN_i 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. Additionally, the SN_i 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 ~~folding~~ fold subclasses [10]. 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 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 were assigned to clan GT-C_{B3}. With 10 transmembrane helices, it is tempting to suggest that this clan may belong to the ~~folding~~ 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 an unexpected 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 of the donor, i.e. the residues closest to the reaction center (Fig. 3). 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 involves a direct translation of glycan IUPAC nomenclature into terms representing backbone configuration, i.e., ignoring chemical modifications and sidechains. Furthermore, a positive similarity score requires identical matches at both donor and acceptor positions (–1 and +1 sites in Fig. 3, respectively). 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 ~~bacterial polysaccharide polymerases~~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 [44]. The multiplicity of polysaccharide biosynthesis loci in some genomes makes it sometimes difficult to assign a particular polysaccharide structure to a particular biosynthesis operon. Although the families described here do not solve all problems, their correlation with the stereochemical outcome of the glycosyl transfer reaction allows to resolve some inconsistencies (~~vide supra~~see above).

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 ~~General methods used for building CAZy families~~

~~The sequence libraries for the different families were built from the seed sequences using "Blastp" from BLAST+ 2.12.0+ [45] against the NCBI non-redundant database version 61. Redundancy reduction was performed using CD-HIT 4.8.1 [46].~~

~~MSAs were generated with MAFFT v7.508 using the L-INS-i strategy (iterative refinement, using weighted sum-of-pairs and consistency scores, of pairwise Needleman-Wunsch local alignments) [47]. HMMs were built using the "hmmbuild" function from HMMER 3.3.2 [48]. The alignments were inspected in Jalview [49]. Finally, the CAZy families were populated by a combination of the "hmmsearch" function from HMMER and Blastp against Genbank.~~

4.1 Alignment-based Clustering (Aclust)

Phylogenetic trees were generated using an in-house tool called Aclust (G.P.Gippert, manuscript in preparation) comprising the following steps. (1) A distance matrix is computed from all-vs-all pairwise local pairwise alignments [50], or from a multiple sequence alignment provided by MAFFT [47]. The distance calculation is based on a variation of Scoredist ([51]), however with distance values normalized by sequence length rather than alignment length. (2) The distance matrix is embedded into orthogonal coordinates using metric matrix distance geometry [52], and a nearest-neighbor joining algorithm is used to create an initial tree. (3) Beginning with the root node of the initial tree, each left and right subtree constitutes disjoint subsets of the original sequence pool, which are embedded and rejoined separately (i.e., step 2 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

The peptidoglycan polymerase family, GTxx1, was built by using "Blastp" from BLAST+ 2.12.0+ [45] 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 (iterative refinement, using weighted sum-of-pairs and consistency scores, of pairwise Needleman-Wunsch local alignments) [47] an HMM model was built using "hmmbuild" from HMMER 3.3.2 [48]. The family was further populated using "hmmsearch" from HMMER 3.2.2 against Genbank with a threshold of XX.

4.3 Building the Enterobacterial common antigen polymerases family (GTxx2)

A sequence library of ECA-Pols was constructed by using ~~"blastp"~~ "Blastp" from BLAST+ 2.12.0+ [45] with the seed sequence (Genbank accession AAC76800.1) against the NCBI non-redundant database ~~as described in the section "General methods used for building CAZy families". All hits~~ version 61 with an E-value ~~smaller than~~ threshold of 1e-60~~were selected. The~~. The hits were redundancy reduced using CD-HIT 4.8.1 [46] 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 ~~the MSA for the family~~an MSA using MAFFT v7.508 with the L-INS-i strategy [47]. An HMM was built based on this MSA using the "hmmbuild" function from HMMER 3.3.2 [48]. The family ~~was built and populated as described in the section "General methods used for building CAZy families"~~GTxx3 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)

37 O-Lig sequences were selected from literature (Supplementary Table 1) and expanded using "blastp" against the NCBI non-redundant database ~~(see section "General methods used for building CAZy families")~~ with an E-value cut-off of 1e-60~~, resulting in 13,431 hits~~. The blast hits were redundancy reduced 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 section ~~4.2~~4.1), 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 ~~, and the~~ with MAFFT v7.508 using the L-INS-i strategy. HMMs were built based on the MSAs using the "hmmbuild" function from HMMER 3.3.2 [48]. The family was populated ~~as described in section "General methods used for building CAZy families"~~using Blastp against Genbank using 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)

~~365~~ 363 BP-Pol sequences were ~~selected from literature (from 2 phyla, 4 orders, 15 species;~~ collected from review papers on biosynthesis of O-antigens and capsular polysaccharides in different species: *Escherichia coli* [25], *Shigella boydii*, *Shigella dysenteriae*, *Shigella flexneri* [26], *Salmonella enterica* [27], *Yersinia pseudotuberculosis*, *Yersinia similis* [28], *Pseudomonas aeruginosa* [17], *Acinetobacter baumanii*, *Acinetobacter nosocomialis* [29] and *Streptococcus pneumoniae* [30] (complete list in Supplementary Table 2). The BP-Pols for A. *baumannii* O7 and O16 were omitted, because of uncertainty of their serotypes [29]. *P. aeruginosa* O15 was omitted, because it has been shown that the BP-Pol reported in [17] is inactivated and that the O-antigen is synthesized via the ABC-dependent pathway rather than the Wzx/Wzy-dependent pathway [53].

The sequence library was expanded using "blastp" against the NCBI non-redundant database ~~(46,644 hits). All hits~~ with an E-value ~~less than~~ threshold of 1e-15~~and a length between 320 and 600 residues were selected (29,372 hits)~~. Redundancy reduction was performed using CD-HIT with a threshold of 95% identity resulting in a pool of 20,850 sequences.

To ~~build a~~ find clusters of BP-Pol sequences that were big 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 of the BP-Pol sequence pool ~~, an all-vs-all pairwise local alignment was performed using~~ were performed using blastp from BLAST+ 2.12.0+. ~~The networks were visualized with Cytoscape [54]. A bit score threshold of 110 was selected and the~~ A series of networks were built using different bit score thresholds. The members of the resulting SSN clusters were identified using NetworkX [55] ~~.~~

~~MSAs and HMMs were~~ and MSAs of the members were built with MAFFT v7.508 using the L-INS-i strategy. The MSAs were inspected using Jalview [49], 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 ~~as described in section 4.1. The HMMs for each cluster~~ using the "hmmbuild" function from HMMER 3.3.2, and the HMMs were compared using HHblits 3.3.0 [56]. ~~The HHblits network was then visualized in Cytoscape [54] with an HHblits score threshold of 160. CAZy families~~ 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 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 biggest superclusters and populated with sequences present in Genbank ~~as described in the section "General methods used for building CAZy families". by~~ a combination of blastp with the seed sequences and hmmsearch. The networks were visualized with Cytoscape [54].

4.6 Analysis of sugar repeat-unit structures

~~A copy of the bacterial records in the CSDB database (http: //csdb. glycoscience.ru) was provided by Philip Toukach [32] and extracted into a listing of~~

In order to analyze the relation between BP-Pol ~~seeds, linking NCBI protein accessions with CSDB entries based on serotype.~~ 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.

Change this The repeat-unit structures were retrieved in the following way:

- For *Acinetobacter baumannii*, the repeat unit structures of the O-antigens have been summarized in [29] . In this study, the linkage made by the polymerase was determined by analysis of the GTs in the gene clusters.
- For *Streptococcus pneumoniae*, the repeat unit structures of the capsular polysaccharides were summarized in [30], and the polymerase linkages were determined based on the initial transferase and the other GTs in the polysaccharide gene cluster. Eight additional repeat-unit structures were ~~cross-checked with the literature. In cases where there were several sugar structures for a serotype in CSDB and in the literature, we chose the candidate that was~~ included which were elucidated after the review paper; from *S. pneumoniae* 16A [34], 33A [35], 33C and 33D [36], 35C and 35F [37], 42 and 47F [57] and 47A [58]. The polymerase linkage has been determined in all these studies, except for those of *S. pneumoniae* 33A and 47A. For 33A, we determined the polymerase linkage based on the gene cluster having the initial trasferase WchA, which transfers a glucose [29]. 47A has WcjG as the initial transferase, which transfers Gal or Galp [29]. Since the repeat-unit contains both Gal and Galp, we could not determine the polymerase linkage unambiguously. However, the sugar unit is very similar to other sugars in the family (most similar to ~~sugar structures for related BP-Pol sequences.~~ *S. pneumoniae* 13, and we proposed the equivalent phase of that one. Finally, a revised structure has been published of 33B [36], and we used that structure instead.

  ~~It is often, but not always, known which bond of the polysaccharide is created by the polymerase . The sugar structures in CSDB are thus shown as the repeat-units acted upon by BP-Pol, i.e., the bond~~
- For *Yersinia pseudotuberculosis*, the repeat unit structures of the O-antigens have been summarized in [28], in which the polymerase linkages were also determined. For *Y. pseudotuberculosis* O3, we used the revised structure [38].
- For *Salmonella enterica*, the repeat unit structures of the O-antigens have been summarized in [27]. In this study, the linkage made by the polymerase ~~is the bond between the rightmost monosaccharide (the −1 site position, see also Fig 3.) and the leftmost monosaccharide (the +1 site position). However, there are cases where~~ was determined by analysis of the GTs in the gene clusters.
- For *Escherichia coli*, the repeat unit structures of the O-antigens have been summarized in [33]. check how they determined the polymerase linkage
- For *Shigella*, the repeat unit ~~is provided in another "phase", ie. ,~~ structures of the O-antigens have been summarized in [26]. In this study, the linkage made by the ~~bond predicted to be catalyzed by~~ polymerase was determined based on the initial sugar unit being GlcpNAc or GalpNAc.
- For *Pseudomonas aeruginosa*, the repeat unit structures were retrieved from the review paper [17]. check how they determined phases *Pseudomonas aeruginosa* O2 and O16 contain two BP-Pol ~~is positioned internally within the linear repeat unit rather than at an end. We cross-checked the "phases" with the literature, and in cases where this was not provided in the literature, we compared them to other sugar structures from related~~ genes; one BP-Pol ~~sequences, and we could predict the polymerase bond and rearrange the sugar structures manually to show the putative correct phase.~~ 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 [59]. The polymerases in our dataset are the ones that polymerize the α bond and we therefore report the sugar structure with the alpha bond.

The CSDB database (http://csdb.glycoscience.ru) [32] was used for finding literature and retrieving linear sugar strings and SNFG image representations of the ~~carbohydrates were generated at the CSDB website~~repeat-unit structures.

Phylogenetic trees for only seed sequences in each of the newly created BP-Pol families were generated using MAFFT v7.508 [47] to supply an initial multiple sequence alignment, followed by Aclust (section 4.2) for distance matrix embedding and clustering. Seed sequences are those where the sugar repeat-unit structure is known. The trees were visualized with the corresponding sugar structures in iTOL [60].

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, adding one to the score for each additional identical match, but terminating at the first non-identity.

To facilitate the scoring, we have chosen to first translate oligosaccharides from IUPAC nomenclature into a set of simplified geometric subunits that represent only monomer dimension and stereochemistry of acceptor and anomeric donor carbon atoms, thus focusing entirely on the glycan backbone (Fig. 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 linear/open, 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), this latter category is assigned in the cases of extensive conformational flexibility such as with alditols or C6 linkages. Chemical modifications, side chains, and the configuration of non-linking carbons are ignored. Further details, limitations and extensions will be presented elsewhere (G.P. Gippert, manuscript in preparation).

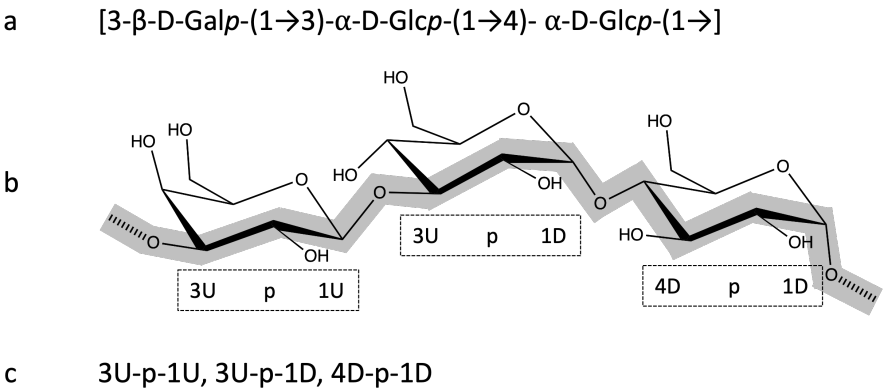


Figure 9: Oligosaccharide translation from IUPAC nomenclature to backbone (geometric) subunits for a trisaccharide consisting of one D-galactopyranose and two D-glucopyranose residues joined by intramolecular β1→3 and α1→4 bonds, respectively, and an intermolecular α1→3 bond formed in the polymerase 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.

4.8 Comparison of the families

Pairwise HHblits analyses [31] were performed for each of the new CAZy families. The HHblits scores were visualized in a heatmap using Python Matplotlib [61].

AlphaFold2 [13] structures were generated of representative proteins from the families using the ColabFold implementation [62] on our internal GPU cluster processed with the recommended settings. The best ranked relaxed model was used[62]. The protein structures were visualized in PyMOL [63] and pairwise ~~structural superimpositions were performed using the CEalign algorithm [64].~~

4.9 ~~AlphaFold structures~~

~~The included AlphaFold2 predicted structures were generated using the ColabFold implementation on our internal GPU cluster processed with the recommended settings using the best ranked relaxed model [62]. The protein structures were visualized in PyMOL [63] and~~ structural superimpositions were performed using the CEalign algorithm [64].

5 Data availability

Accessions to the seed sequences utilized in this work are given in Supplementary Table 1-2 along with the polysaccharide repeat structure; the constantly updated content of families GTxx1 - GTx17 is given in the

online CAZy database at www.cazy.org.

6 Acknowledgements

This work was supported by grant NNF20SA0067193 from the Novo Nordisk Foundation. 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.

7 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.

8 Competing interests

None

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

[6] 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.

[7] 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.

[8] 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 (Pt 3)**, 929–939 (1997).

[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] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Rai, A. K. & Mitchell, A. M. Enterobacterial Common Antigen: Synthesis and Function of an Enigmatic Molecule. *mBio* **11**, e01914–20 (2020). URL https://journals.asm.org/doi/10.1128/mBio.01914-20.

[17] 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.

[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] 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.

[20] Goffin, C. & Ghuysen, J.-M. Multimodular Penicillin-Binding Proteins: An Enigmatic Family of Orthologs and Paralogs. *Microbiology and Molecular Biology Reviews* **62**, 1079–1093 (1998). URL https://journals.asm.org/doi/10.1128/MMBR.62.4.1079-1093.1998.

[21] Taguchi, A. *et al.* FtsW is a peptidoglycan polymerase that is functional only in complex with its cognate penicillin-binding protein. *Nature Microbiology* **4**, 587–594 (2019). URL https://www.nature.com/articles/s41564-018-0345-x.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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).

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] Henrissat, B. & Bairoch, A. Updating the sequence-based classification of glycosyl hydrolases. *The Biochemical Journal* **316 (Pt 2)**, 695–696 (1996).

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] Sonnhammer, E. L. & Hollich, V. [No title found]. *BMC Bioinformatics* **6**, 108 (2005). URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-108.

[52] Crippen, G. & Havel, T. *Distance Geometry and Molecular Conformation*, vol. 15 (Chemometrics Research Studies Press Series, Research Studies Press, 1988).

[53] 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.

[54] 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.

[55] Hagberg, A., Swart, P. & S Chult, D. Exploring network structure, dynamics, and function using networkx (2008). URL https://www.osti.gov/biblio/960616.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

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

[62] 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.

[63] Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8 (2015).

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